



**DEEP  
VADOSE ZONE  
PROGRAM**  
@PNNL

# Data Catalog Selection and Implementation

21 August 2024

**Rebecka Bence**

**Aaron Moreno**

**Kenneth Ham**

**Chris Johnson**



PNNL is operated by Battelle for the U.S. Department of Energy

# Agenda

1. Need for data catalogs
2. Function of a data catalog
  - a) Metadata overview
  - b) Required functionality in practice
3. Evaluating data catalog tools
4. Data catalog requirements
  - a) Defining requirements
  - b) Implementation considerations
5. Example of data catalog prototype and features





# Background and Need

- What is a data catalog?
  - Consider a library card catalog
    - Why did you need the catalog?
    - What would have happened if you didn't have the card catalog?
- Definition
  - **From IBM:** An inventory of all data assets in an organization, designed to help professionals quickly find the most appropriate data for any analytical or business purpose



PS3557  
.R5355 Grisham, John  
F57 1991

The firm / John Grisham. 1st. ed.  
New York : Doubleday, c1991.  
42lp. ; 24 cm.

1. Government investigators--Fiction.  
2. Organized crime--Fiction.



# Background and Need

- What need is there for a data catalog?
  - Multiple environmental databases and data assets
    - Data inherently comes in different forms and formats
      - Models
      - Electronic tabulated data
      - Geospatial data
      - Report narratives
      - Tables and figures
      - Images
  - Need a means to find these data resources across multiple systems
  - Multiple agencies/contractors
    - Overlapping "authoritative" data sources

## 3.6.3 Proposed Input Parameters

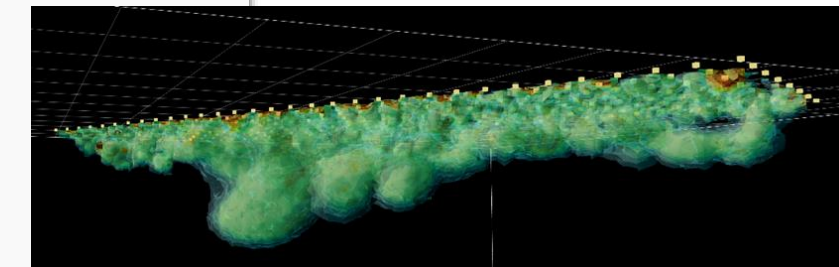
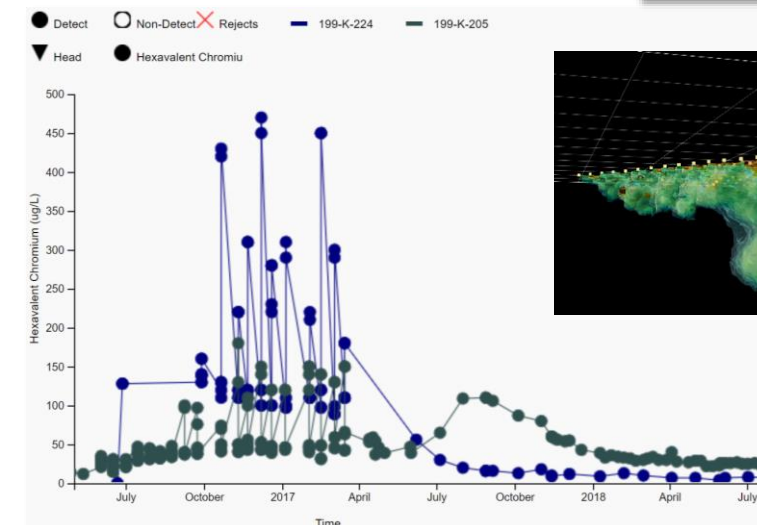
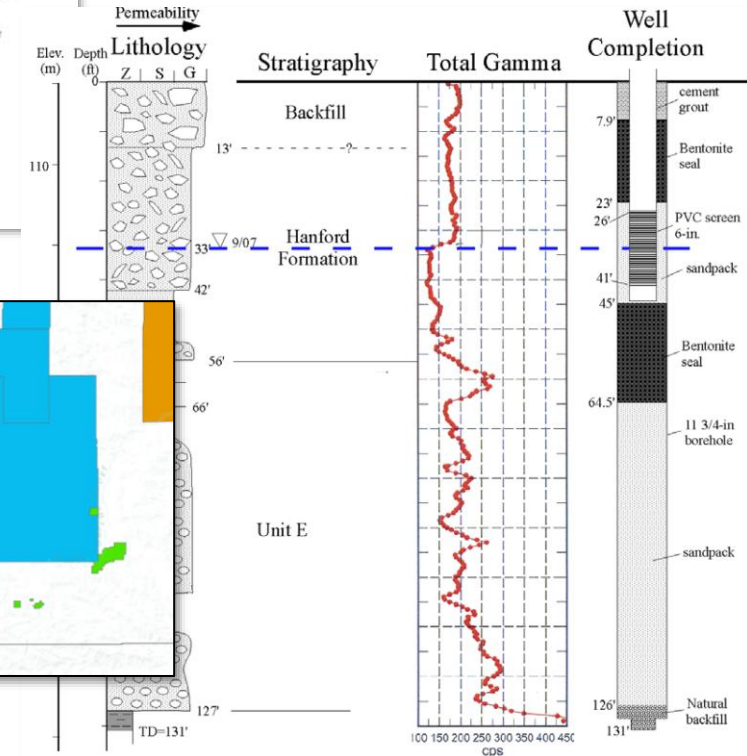
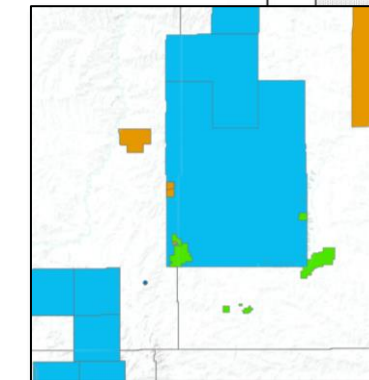
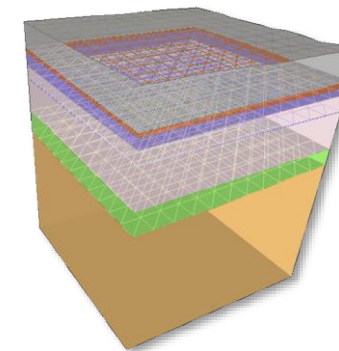
The nominal values in the governing concentration equation for the analytes used in the Composite Analysis are provided in Table 1. The values in Table 1 have been modified to have units of  $C/m^3$ . The modification is performed by multiplying the original data in  $pCvL$  by  $10^{-6}$ .

The stochastic distributions associated with the nonzero coefficients defined in Table 1 are defined using the following rules:

- **Variable  $C_0$**  The triangular distribution will be used for all values of  $C_0$ . The distribution will be symmetric about the midpoint, and the half-range will be 50% of the mid-point. The variable tag will be CB.

Table 1. Nominal Coefficient Values for Background Concentrations in the Columbia River

Analyte ID	$C_0$	$M_0$	$\lambda_0$	Fallout?
$^{14}C$	$5.3 \times 10^{-10}$	0	0	No
$^{14}C$	0	0	0	No
$^{137}Cs$	0	$5.49 \times 10^{-12}$	0.223	Yes
$^{137}Eu$	0	0	0	Very small
$^{137}I$	$1.5 \times 10^{-8}$	$3.04 \times 10^{-8}$	0.0562	Yes
$^{137}I$	0	$1.1 \times 10^{-14}$	$4.41 \times 10^{-6}$	Very small
$^{235}Pu$	0	0	0	Very small
$^{239}Pu$	$8.7 \times 10^{-12}$	0	0	No
$^{239}Pu$	Not modeled	Not modeled	Not modeled	No



**DEEP  
VADOSE ZONE  
PROGRAM**  
@PNNL



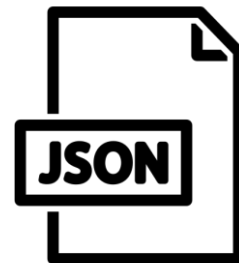
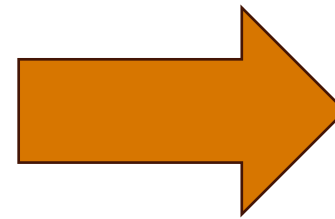
# Basic Functionality

- Catalogs are a key aspect to making data FAIR (findable, accessible, interoperable, and reusable)
- To make data FAIR, catalogs store **metadata**

```
PS3557
.F5355      Grisham, John
F57 1991

The firm / John Grisham. 1st. ed.
New York : Doubleday, c1991.
42lp. ; 24 cm.

1. Government investigators--Fiction.
2. Organized crime--Fiction.
```



```
<gmd:MD_Metadata xmlns:gmd="http://www.isotc211.org/2005/gmd" xmlns:gco="http://www.isotc211.org/2003/gco" xmlns:gml="http://www.opengis.net/gml" xmlns:xlink="http://www.w3.org/1999/xlink">
  <gmd:fileIdentifier>
    <gco:CharacterString>1686241220912r3853225710481458</gco:CharacterString>
  </gmd:fileIdentifier>
  <gmd:language>
    <gco:CharacterString>en</gco:CharacterString>
  </gmd:language>
  <gmd:hierarchyLevel>
    <gmd:MD_ScopeCode codeSpace="ISOTC211/19115" codeList="http://www.isotc211.org/2003/isotc211_19115" codeListValue="dataset">dataset</gmd:MD_ScopeCode>
  </gmd:hierarchyLevel>
  <gmd:hierarchyLevelName>
    <gco:CharacterString>Data</gco:CharacterString>
  </gmd:hierarchyLevelName>
  <gmd:contact>
    <gmd:CI_ResponsibleParty>
      <gmd:organisationName>
        <gco:CharacterString>RJ Lee Group Inc.</gco:CharacterString>
      </gmd:organisationName>
      <gmd:role>
        <gmd:CI_RoleCode codeSpace="ISOTC211/19115" codeList="http://www.isotc211.org/2003/isotc211_19115" codeListValue="author">author</gmd:CI_RoleCode>
      </gmd:role>
    </gmd:CI_ResponsibleParty>
  </gmd:contact>
  <gmd:dateStamp>
    <gco:Date>2023-06-08</gco:Date>
  </gmd:dateStamp>
  <gmd:metadataStandardName>
    <gco:CharacterString>ISO 19139/19115 Metadata for Datasets</gco:CharacterString>
  </gmd:metadataStandardName>
  <gmd:metadataStandardVersion>
    <gco:CharacterString>2003</gco:CharacterString>
  </gmd:metadataStandardVersion>
  <gmd:identificationInfo>
```





# Metadata Standards

- How information is organized in a metadata file matters for readability and interoperability

RJ LEE GROUP

Heather Medley  
Central Plateau Cleanup  
Company, LLC (Env)  
PO Box 1464  
Richland, WA 99352  
Client Project: Anions

## Laboratory Report

RJ Lee Group No.: W208169  
SDG No: RJLG22C0529  
Samples Received: 08/18/22 09:10  
Analysis/Prep Date: 08/19/22 04:23  
Report Date: 08/31/22 07:34

Sample Name: B45NF7		Batch No: BH20086		Date Sampled: 08/17/22 12:06		
RJ Lee Grn. ID: W208169-01		Matrix: Water		Date Analyzed: 08/19/22 04:23		
Analyte	Method	Result µg/mL	PQL µg/mL	MDL µg/mL	Dilution Factors	Qualifiers
Bromide	EPA 300.0	0.29	0.20	0.10	2	D
Chloride	EPA 300.0	19.7	1.50	0.75	50	D
Fluoride	EPA 300.0	0.28	0.04	0.02	2	D
Nitrate as NO3-N	EPA 300.0	41.9	1.15	0.60	50	D
Nitrite as NO2-N	EPA 300.0	< 0.03	0.06	0.03	2	U, D
Sulfate	EPA 300.0	46.5	7.50	3.75	50	D

```

<gmd:MD_Metadata xmlns:gmd="http://www.isotc211.org/2005/gmd" xmlns:gco="http://www.opengis.net/gml" xmlns:xlink="http://www.w3.org/1999/xlink"
  <gmd:fileIdentifier>
    <gco:CharacterString>1686241220912r3853225710481458</gco:CharacterString>
  </gmd:fileIdentifier>
  <gmd:language>
    <gco:CharacterString>en</gco:CharacterString>
  </gmd:language>
  <gmd:hierarchyLevel>
    <gmd:MD_ScopeCode codeSpace="ISOTC211/19115" codeList="http://www.isotc211.org/19115/code/MD_ScopeCode" codeListValue="dataset">dataset</gmd:MD_ScopeCode>
  </gmd:hierarchyLevel>
  <gmd:hierarchyLevelName>
    <gco:CharacterString>Data</gco:CharacterString>
  </gmd:hierarchyLevelName>
  <gmd:contact>
    <gmd:CI_ResponsibleParty>
      <gmd:organisationName>
        <gco:CharacterString>RJ Lee Group Inc.</gco:CharacterString>
      </gmd:organisationName>
      <gmd:role>
        <gmd:CI_RoleCode codeSpace="ISOTC211/19115" codeList="http://www.isotc211.org/19115/code/CI_RoleCode" codeListValue="author">author</gmd:CI_RoleCode>
      </gmd:role>
    </gmd:CI_ResponsibleParty>
  </gmd:contact>
  <gmd:dateStamp>
    <gco:Date>2023-06-08</gco:Date>
  </gmd:dateStamp>
  <gmd:metadataStandardName>
    <gco:CharacterString>ISO 19139/19115 Metadata for Datasets</gco:CharacterString>
  </gmd:metadataStandardName>

```



DEEP  
VADOSE ZONE  
PROGRAM  
@PNNL



# Metadata Standards and Schema

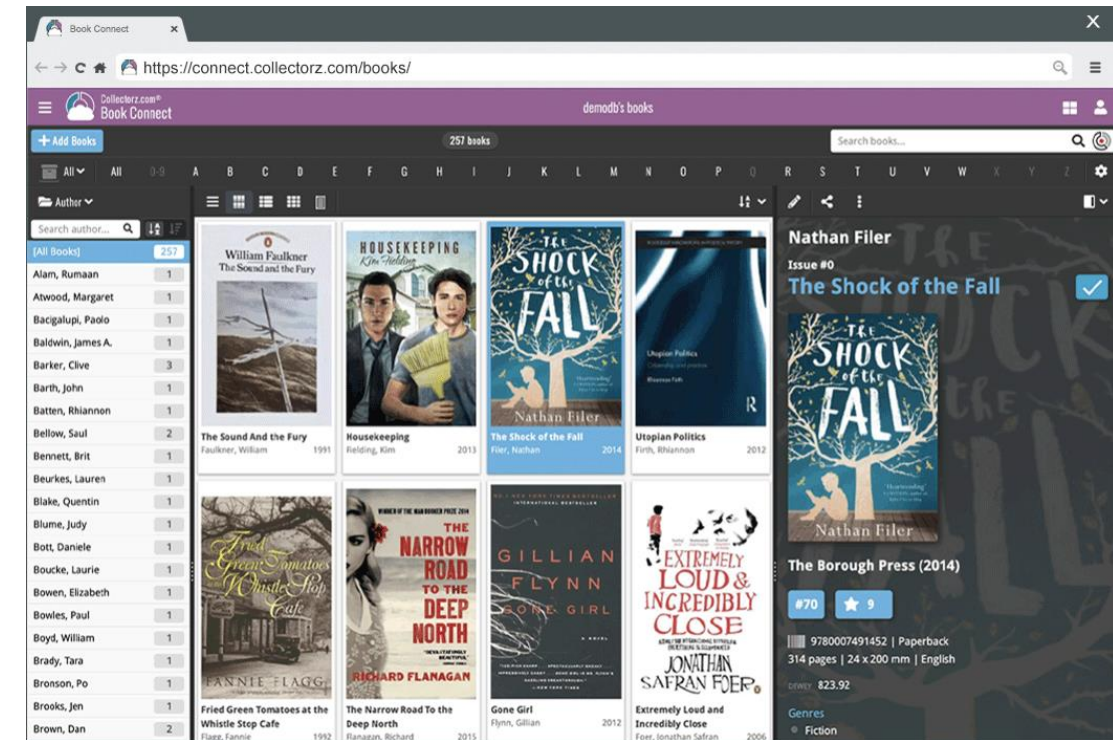
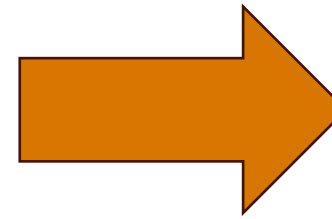
- There are many types of metadata standards – generic to domain-specific
  - Dublin Core
  - International Organization for Standardization (ISO) 19XXX series
  - Federal Geographic Data Committee (FGDC)
  - Darwin Core
  - NASA's Standards
- *Schema* is the part of the standard that outlines the overall structure of the metadata and addresses how to handle common components like dates, names, and places
- Adopting vs. adapting metadata schema
- Encoding schemes to further standardization







# Catalog Functions



- To provide access to datasets, a data catalog must:
  - Manage records
  - Enable users to find datasets of interest
  - Provide a way to retrieve the dataset of a selected catalog record
  - Be able to manage/store large numbers of catalog records







# Catalog Functions: Cataloging

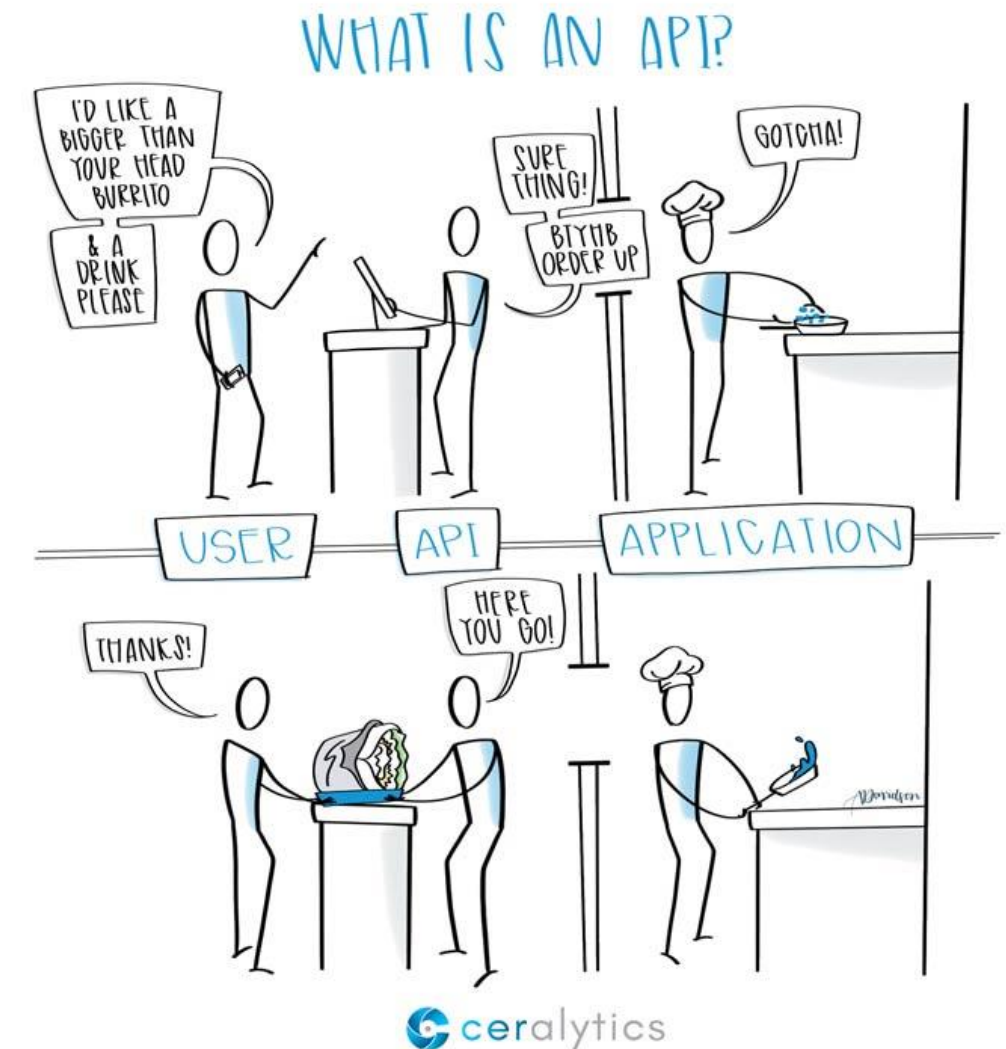
- A record is created by compiling descriptive and well-organized metadata
- A record usually includes a link to the dataset/asset
  - Assisting features:
    - ✓ A well documented metadata standard/schema
    - ✓ Ability to validate incoming metadata
    - ✓ A metadata creation wizard or tool

Call Number	Author: Last, First
Summary of Book	
Genre: Fiction, Non-fiction, Poetry	
Date: YYYYMMDD	



# Catalog Functions: Cataloging

- The number of records to catalog will continue to grow
- The ability to easily import and export data is a key function of the catalog
  - Minimize manual data entry and transformation
  - Minimize the need to transform metadata (i.e., import and export metadata in common standard schemas)
  - Possible Solutions
    - Application Programming Interfaces (APIs)
    - REpresentational State Transfer (REST) APIs

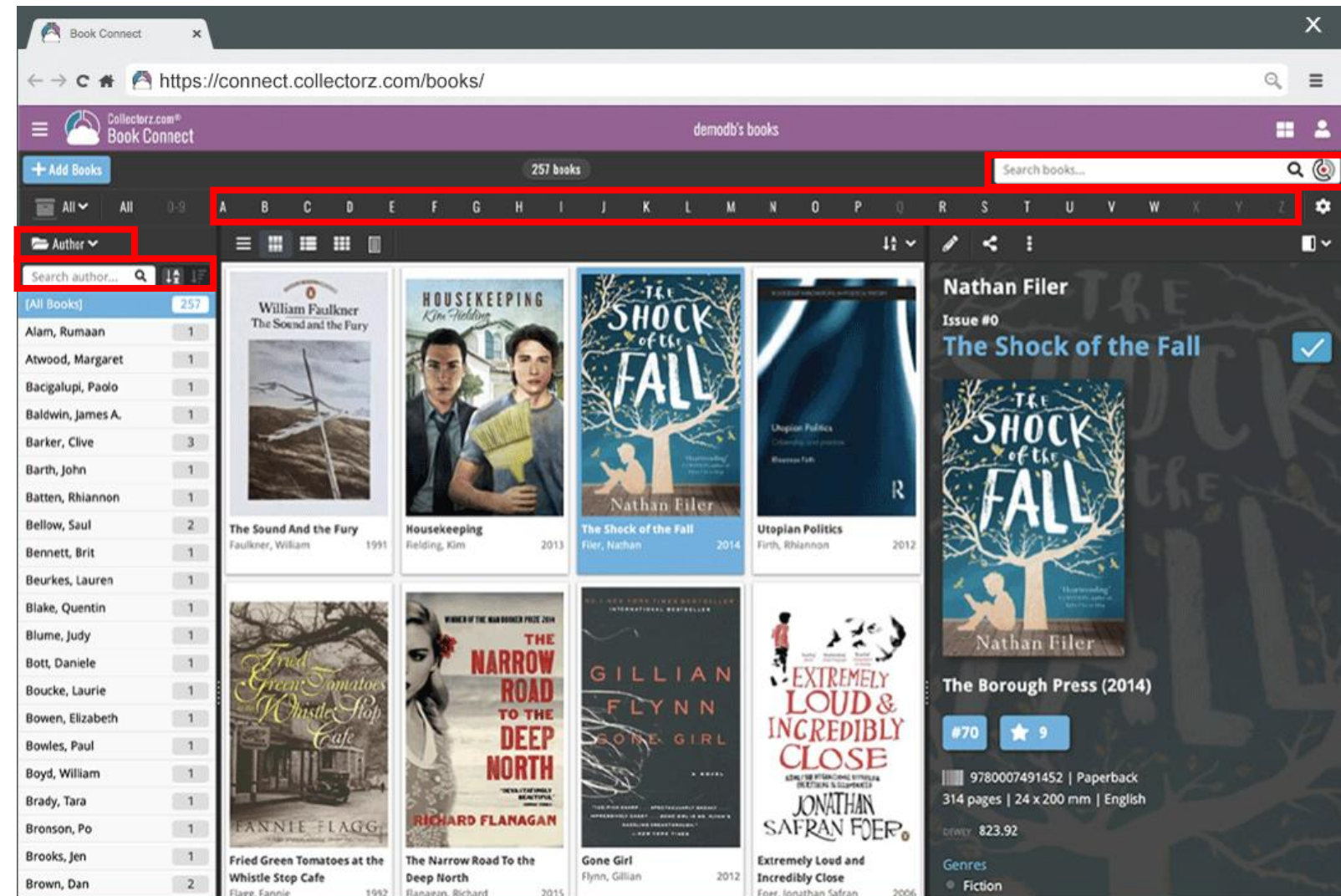






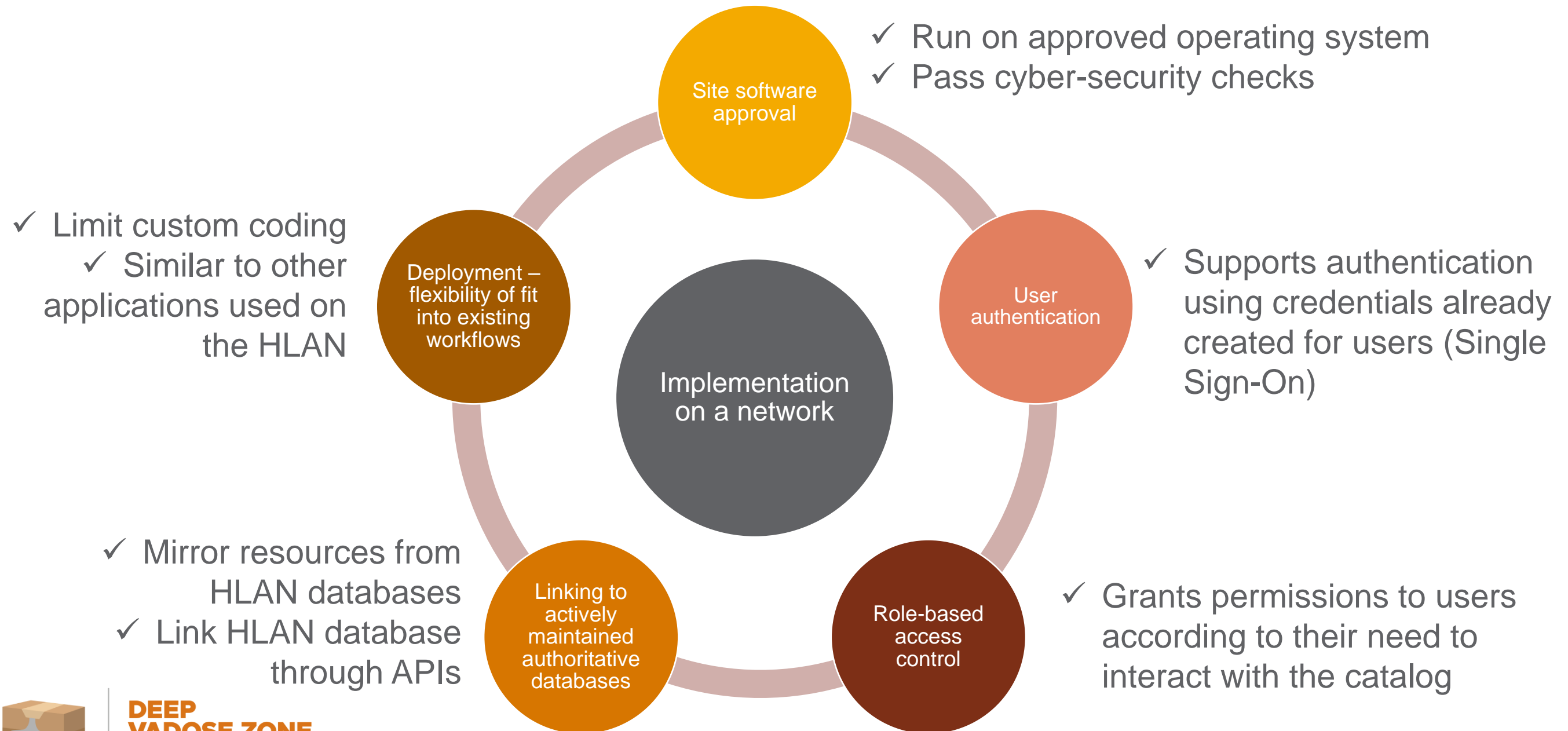
# Catalog Functions: Find and Retrieve

- Finding a record relies on imposing selection criteria on the metadata fields
  - Filter – words or spatial data
- Retrieve the record directly
  - Inspect the resource
  - Access the resource (if role-based restrictions allow it)
    - ✓ Links to the resource allows for cataloging large datasets
    - ✗ This can limit the ability to map, plot, or filter data within the catalog interface



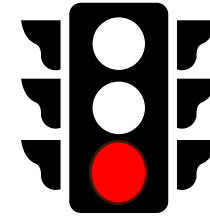


# Catalog Functions: Practical Implementation Considerations at Hanford





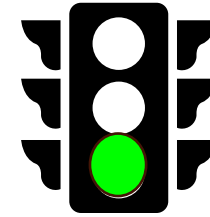
# Review Questions



1. What does FAIR stand for?
2. Name two common minimum requirements for metadata.
3. True or False – I should upload Word documents, PDF files, and Excel files to a catalog



# Review Questions



1. What does FAIR stand for?  
**Findable, Accessible, Interoperable, Reusable**
2. Name two common minimum requirements for metadata
  - a. Title
  - b. Description
  - c. Date
  - d. File format
  - e. Metadata Standard
  - f. Unique identifier
  - g. Contact information
3. True or False – I should upload Word documents, PDF files, and Excel workbooks to a catalog  
**False**







# Catalog Software Evaluation

- Compared leading commercial and open-source data catalog platforms
  - Criteria to assess functionality
  - Capabilities supporting data discoverability, retrieval, and archiving
  - Metadata standard compatibility
  - Integration into chosen network: Hanford network (HLAN)

## Proprietary Platforms

- ArcGIS Enterprise Sites
- Junar
- OpenDataSoft
- Socrata

## Non-proprietary Platforms

- Energy Data eXchange (EDX)
- Comprehensive Knowledge Archive Network (CKAN)
- DKAN (Drupal-based open data portal based on CKAN)





# Example Deployments

- Existing deployments were helpful to understand capabilities
- Clockwise from top left:
  - OpenDataSoft – City of Vancouver Open Data Portal
  - DKAN – USDA Ag Data Commons
  - Junar – City of Palo Alto Open Data Portal
  - Socrata – City of Austin, Texas, Open Data Portal

**954 records**  
No active filters  
Filters

Search records...

**TYPE**

Pedestrian Actuated Signal	397
Fixed Time	259
Semi Actuated	192
Fully Actuated	49
RRFB	34
Special Crosswalk	11
> More	

**Geo Local Area**

Downtown	164
Kitsilano	70
Fairview	65
Renfrew-Collingwood	64
Kensington-Cedar Cottage	60
Mount Pleasant	55
> More	

**Traffic signals**

New records are currently being added to this dataset, this process may take a while. Please note that visualizations might be incomplete in the meantime. Don't hesitate to refresh your page regularly!

This dataset contains the locations of the City's traffic signals.

**Data currency**  
This data is updated frequently in the normal course of business, however priorities and resources determine how fast a change in reality is reflected in the database. The extract on this website is updated weekly.

**Data accuracy**  
Traffic signal locations are in the approximate centre of the intersection.

**Dataset Identifier** traffic-signals  
**Downloads** 439  
**Data Owner** City of Vancouver  
**Data Team** Engineering Services  
**Search Terms** light  
**Themes** Streets and transportation  
**Keywords** traffic  
**License** Open Government Licence - Vancouver  
**Modified** July 19, 2021 3:08 AM  
**Publisher** City of Vancouver  
**Follow**

**Proposed Corridor Construction Program**  
City Infrastructure

The following 34 "investment packages" were derived from recommendations in Corridor Mobility Plans for the nine corridors eligible for 2016 Mobility Bond construction funding in accordance with the ballot language approved by voters in November 2016. Those corridors are: North Lamar Boulevard, Burnet Road, Airport Boulevard, East MLK Jr.

Updated April 14, 2021  
Data Provided by City of Austin Corridor Program Office

**About this Dataset**

Updated April 14, 2021  
Data Last Updated February 23, 2018  
Metadata Last Updated April 14, 2021  
Date Created February 22, 2018

Views 549  
Downloads 1,016

Data Provided by City of Austin Corridor Program Office  
Dataset Owner saradanae  
Contact Dataset Owner

**City of Austin**  
Department Capital Planning Office  
Additional Information  
Frequency As Needed  
Digital Object Identifier (DOI)  
DOI Number https://doi.org/10.26000  
Topics  
Category City Infrastructure  
2016 mobility bond, 201 road projects, sidewalk  
Tags  
Show More

USDA Ag Data Commons  
U.S. DEPARTMENT OF AGRICULTURE  
Providing Central Access to USDA's Open Research Data

Log in

Datasets Software & Tools About Us News Contact Us

Home / BAR- The Bio-Analytic Resource for Plant Biology

**Filter By:**  
License  
U.S. Public Domain

**Other Access**  
The information on this page (the dataset metadata) is also available in these formats:  
JSON RDF  
via the DKAN API

**Social**  
Twitter LinkedIn Reddit Google+ Facebook

**BAR- The Bio-Analytic Resource for Plant Biology**

BAR is a collection of web-based, user-friendly tools for exploring, visualizing, and analyzing large datasets from plants. Supported are expression data, Next-Gen sequence data, protein-protein interactions, polymorphisms / conservation, and protein 3-D structures.

The BAR is funded in part by Centre for the Analysis of Genome Evolution and Function, grants from the Canada Foundation for Innovation to Dr. Provart, and from Genome Canada to the Arabidopsis Research Group at the Department of Cell and Systems Biology, University of Toronto. The BAR may be used to explore large-scale data sets from Arabidopsis and other species, and for hypothesis generation.

**BAR - The Bio-Analytic Resource for Plant Biology**  
Explore Data

Field	Value
Modified	2019-08-05
Release Date	2018-01-23
Identifier	f6aa9999-516c-44d8-97f3-60b52bbe2629
Publisher	University of Toronto
License	U.S. Public Domain
Contact Name	Provart, Nicholas
Contact Email	nicholas.provart@utoronto.ca
Public Access Level	Public

**Extended Metadata:**

Field	Value
Authors	University of Toronto
Peer Reviewed	No
Product Type	Software tool

Field	Value
ISO Topic(s)	biota
National Agricultural Library Thesaurus Term	evolution, grants, Canada, Arabidopsis, genes, data collection, protein-protein interactions, membrane proteins, genomics, gene expression, fluorescence, genetic markers, transcription factors, binding sites, databases, DNA, nucleotide sequences, gene expression regulation, bioinformatics, seed coat, Magnoliopsida, Medicago, soybeans, potatoes, Arachis, grapes, Liliopsida, corn, barley,

Search for Open Data



**DEEP  
VADOSE ZONE  
PROGRAM**  
@PNNL





# Assessment of Data Catalog Software Tools

- F = fails to meet requirements
- M = meets requirements
- E = exceeds requirements (i.e., meets requirements and delivers additional desired features)
- Lowest rating for any capability area was assigned as the overall rating

Capability	ArcGIS	CKAN	DKAN	EDX	Junar	OpenDataSoft	Socrata
Catalog	E	E	E	M	F	E	E
Find	E	E	E	M	E	M	E
Retrieve	E	E	E	E	M	E	E
Large dataset storage	E	E	E	M	F	F	E
Hanford Site software approval	E	E	E	E	M	M	E
Single sign-on	E	E	E	F	F	E	E
Role-based access	E	E	E	M	F	M	E
Linking/federation	E	E	E	E	E	E	E
Deployment	E	M	M	M	F	F	M
Overall rating	E	M	M	F	F	F	M





# Key Assessment Findings

- Propriety software-as-a-service (SaaS) model of delivering a data catalog
  - E.g., Junar, OpenDataSoft
  - Favors consistency across customers at expense of customization/configurable roles
  - Limits size of dataset hosted on the shared commercial platform
  - Socrata, however, provided custom roles and gateways allowing local datasets to be incorporated into an online catalog
- Inability to incorporate authentication from site network (e.g., HLAN) was an issue
  - EDX and Junar couldn't meet that capability
- Initial evaluation found strong candidates ("M" or "E" ratings)
  - ArcGIS Enterprise Sites, Socrata, CKAN, and DKAN
  - Provide fully self-hosted options, allowing for greater control and flexibility
- Explored ArcGIS Hub (non-enterprise version of ArcGIS Enterprise Sites)
  - Straightforward to use for cataloging datasets and managing metadata
  - Difficult to implement some requirements (e.g., custom keyword lists and other limitations)





# Catalog Requirements

- Cataloging data so it can be discovered, retrieved, interpreted, and reused usually requires context-specific requirements to be met to be functional
- Requirements can be implemented in different aspects of the catalog:

Requirement Implementation Area	Description
Metadata Schema*	Defines the structure of catalog entries and what is available for searching and organizing the catalog
Metadata Editor	Controls how catalog entries are created, requiring certain fields, limiting field contents, etc.
Data Catalog Interface	Defines how users interact with the catalog, how searches are built, and what fields are available to be searched
Network Setting	Defines and controls user authorization and roles, and facilitates interaction among data systems







# Esri Geoportal Server

- Free, open-source tool with web interface and broad functionality
  - Flexible, customizable, and compatible with ArcGIS tools
  - Supports enterprise authentication
- Catalog
  - Inventories the metadata
  - Find & view resources
  - OGC catalog service compliant
  - Multiple interfaces to resources
    - ✓ E.g., RESTful endpoint for data transfer
- Metadata editor
  - Metadata creation/editing
  - Supports ISO 19000-series standards
    - ✓ ISO 19115 North American Profile (NAP)
- Harvester
  - To incorporate existing data catalog entries



**DEEP  
VADOSE ZONE  
PROGRAM**  
@PNNL

OGC = Open Geospatial Consortium

The screenshot displays the Esri Geoportal Server web interface. At the top, there's a navigation bar with 'Geoportal', 'Catalog', 'Map', and 'About'. Below this is a 'Search Catalog' section with a search bar containing 'environment' and a 'Search' button. The results show 417 items, sorted by 'By Relevance'. The first three results are:

- Environment**: Created by zguo on 2020-09-28. Description: Prior to the use of the Environmental Watercourse Mapping and RAR SPEA map layers, the user should review the applicable user notes and disclaimer. <http://www.coquitlam.ca/planning-and-> [HTML](#) [XML](#) [JSON](#) [Links](#) [Add to Map](#) [Preview](#)
- Sedimentary Environment**: Created by zguo on 2020-09-28. [HTML](#) [XML](#) [JSON](#) [Links](#) [Add to Map](#) [Preview](#)
- Structuurvisie - IJsselmeer**: Created by admin on 2020-11-02. Description: Kaartlaag Structuurvisie Noord-Holland 2040. Op de kaart zijn de contouren van het IJsselmeer te zien. Zie ook paragraaf 4.1.1 in de Structuurvisie Noord-Holland 2040. Op 28 september 2015 hebben PS bij [HTML](#) [XML](#) [JSON](#) [Links](#) [Add to Map](#) [Preview](#) [Thumbnail](#) [Download \(HTTP\)](#)

Below the map view, there are filters for 'Time Period', 'Date', 'Owner', 'Topic Category', and 'Metadata Type'.



# Catalog Requirements: Theming and Authoritative Naming

- Keywords categorize data, usually by theme, place, stratum, or temporal aspect
- Topic keywords should be able to be translated or adapted to site-specific theme keywords

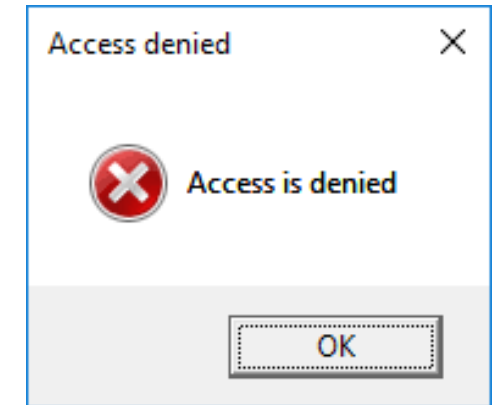
ISO Topic Keywords	Hanford Environmental Data Theme
Farming	Atmospheric
Biota	Biota
Boundaries	External Radiation
Climatology	Groundwater
Meteorology/ atmosphere	Miscellaneous Material
Economy	<u>Pore</u> Water
Elevation	Sediment
Environment	Soil Gas
Geoscientific information	Soil
Health	Surface Water
Imagery/ basemap/ earth cover	Waste Solid
Intelligence	Waste Water
Military	
Inland waters	
Location	
Oceans	
Planning/ cadaster	
Society	
Structure	
Transportation	
Utilities/ communication	

Requirement Implementation Area	Requirement
Metadata Schema	Support custom lists for keywords, including multiple types of place keywords to reflect differing “uses” of a site
Metadata Editor	Enforce custom lists for keywords
Data Catalog Interface	Make custom keywords accessible in search criteria
Network Setting	Consistent use of theming nomenclature across environmental data information systems



# Catalog Requirements: Resource Access Limitations

- There may be a need to limit access to catalog record for security purposes
- Access limitation can be both to the catalog record and to actual resource
- Can be difficult to enforce – federation with existing network settings/authentication is usually key



Requirement Implementation Area	Requirement
Metadata Schema	Support differentiating classes of records by access limitation
Metadata Editor	Require access limitation entry into appropriate metadata field
Data Catalog Interface	Enforce limitation on access to catalog record based on network identity or role
Network Setting	Maintain lists of identities with similar roles and access







# Catalog Requirements: Resource Use Limitations

- Related to but different from access restriction
- Use limitations drive what a user should or should not do with the resource
- Can be difficult to enforce

## How to credit OpenStreetMap

Where you use OpenStreetMap data, you are required to do the following two things:

- Provide credit to OpenStreetMap by displaying our copyright notice.
- Make clear that the data is available under the Open Database License.



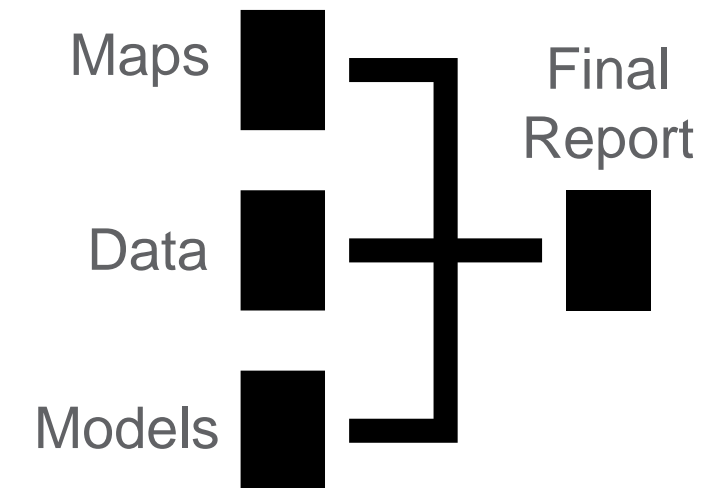
Requirement Implementation Area	Requirement
Metadata Schema	Support specifying constraints on use of the resource
Metadata Editor	Support or require use limitation entry in appropriate metadata field
Data Catalog Interface	Prominently display constraints on use of the resource
Network Setting	Not Applicable. (Limiting how data is used outside of the catalog involves voluntary compliance with terms of use or external controls)





# Catalog Requirements: Link Data to Deliverable

- Accessing the source data for a final report or output is a key part of data transparency and reproducibility
- Can produce more robust analyses by contextualizing new data with historical data
- Spectrum of complexity – site specific need drives the requirement within constraint of implementation areas



Requirement Implementation Area	Requirement	Alternative Requirement
Metadata Schema	Support aggregate datasets	Support a deliverable as a resource
Metadata Editor	Link aggregate datasets to a deliverable	Use lineage fields to identify data sources
Data Catalog Interface	Make aggregate datasets findable based on a deliverable	Handle deliverables as a resource in a catalog entry
Network Setting	Enable linking to deliverables to deliverable database	Enable linking to deliverables in IDMS





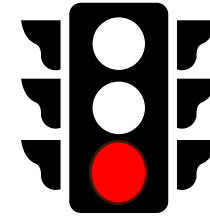
# Catalog Requirements: Testing Results for Esri Geoportal

Requirement	Schema	Editor	Catalog Interface	Network	Notes
Theming	✓	✓	✓	✗	Consistent use of theming nomenclature across environmental data systems not in use
Access Limitations	✓	✓	✗	✓	Continuing to test if role-based permission can be applied to a metadata field
Use Limitations	✓	✓	✓	N/A	
Link to Deliverable	✓	✓	✓	✗	Alternative requirements only. Linking to document deliverable system unable to be tested due to security constraints.





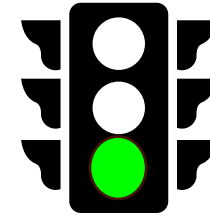
# Review Question



1. What are the four requirement areas you could consider when determining and testing requirements for a catalog?



## Review Question



1. What are the four requirement areas you could consider when determining and testing requirements for a catalog?
  - a. Metadata schema – does the selected schema fulfill your site-specific requirements?
  - b. Metadata editor – can site-specific metadata requirements be easily applied in the catalog's metadata editor 'wizard' or in an automated way?
  - c. Catalog interface – does the chosen software meet the usability needs of your target users? Will it find, retrieve, and display catalog entries in a way that fulfills site-specific requirements?
  - d. Network – what network requirements does the catalog software need to meet to be functional when deployed?





# Hanford Case Study

- Defined a use case
  - ✓ Hanford-specific requirements and HLAN enterprise data system implementation
- Tested available software
  - ✓ Selected Esri Geoportal Server
- Selected metadata standard
  - ✓ ISO 19115
- Defined catalog requirements
  - ✓ Geophysical data as a use case
- Began prototype testing
  - ✓ Prototype catalog built using local environment, hosting metadata from the Hanford Administrative Record and other open sources
  - ✓ Customization and Hanford-specific requirements (e.g., Hanford-specific theming, catalog entry restriction) were demonstrated







# Collaboration and User Stories...

- Meet as a “Hanford Working Group” every two weeks with site contractor and sponsor representatives
  - Refines requirements
  - Functional testing

- User stories developed



A scientist needs data on all waste sites including location.



A geophysicist need seismic data, magnetic and gravity for the suprabasalt sediments in the 600 area.



A hydrologist needs to locate hourly historic barometric pressure to normalize aquifer tests in the 1960's in 200E.

How do I...?

Environmental  
Dashboard  
Application

Hanford Maps

Hanford  
Waste Information  
Data System

SOCRATES

Contractor  
Datasets

Hanford GIS

Hanford Well  
Information  
System

PHOENIX

Future Datasets

Hanford Environmental Information System



DEEP  
VADOSE  
PROGRAM  
@PNNL



# Prototype in Esri Geoportal Server

Search using map feature

Sort and filter (customizable)

Hanford themed keywords

Metadata 'tiles' to view, edit, add to map, etc.

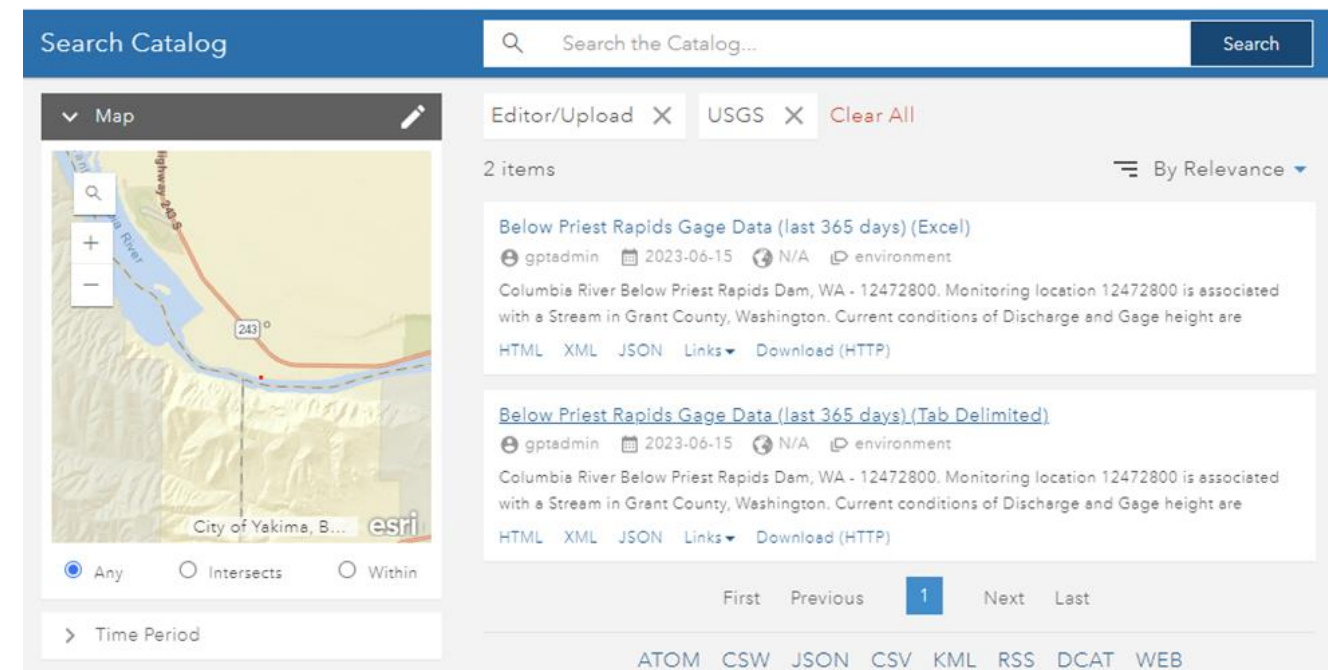
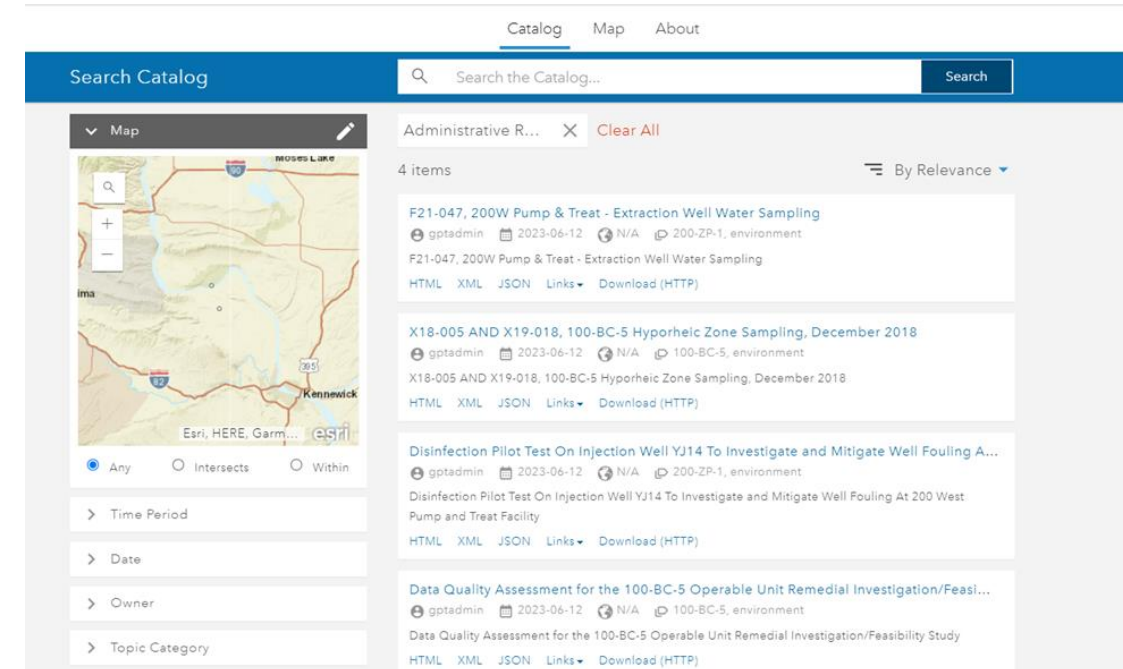
The screenshot displays the Esri Geoportal Server interface. At the top, there's a "Search Catalog" header with a search bar containing "200-ZP-1" and a "Search" button. Below the header, the left sidebar shows a "Map" view with a map of the Hanford area. The right pane shows search results for "200-ZP-1", listing two items: "F21-047, 200W Pump &amp; Treat - Extraction Well Water Sampling" and "Disinfection Pilot Test On Injection Well YJ14 To Investigate and Mitigate Well Fouling A...". The interface includes filters for "Time Period", "Date", "Owner", "Topic Category", "Metadata Type", "Organizations", and "Keywords". The "Keywords" section is expanded, showing "200-ZP-1 (2)". At the bottom, there are links for "ATOM", "CSW", "JSON", "CSV", "KML", "RSS", "DCAT", and "WEB".





# Geoportal Server – Example Data Records

- Hanford Administrative Record entries
  - Via web scraping
  - Geospatial context (operable unit) was added
- Records for Columbia River stage gage data
  - Station 12472800, below Priest Rapids Dam
  - Retrieved from USGS Daily Values Web Service

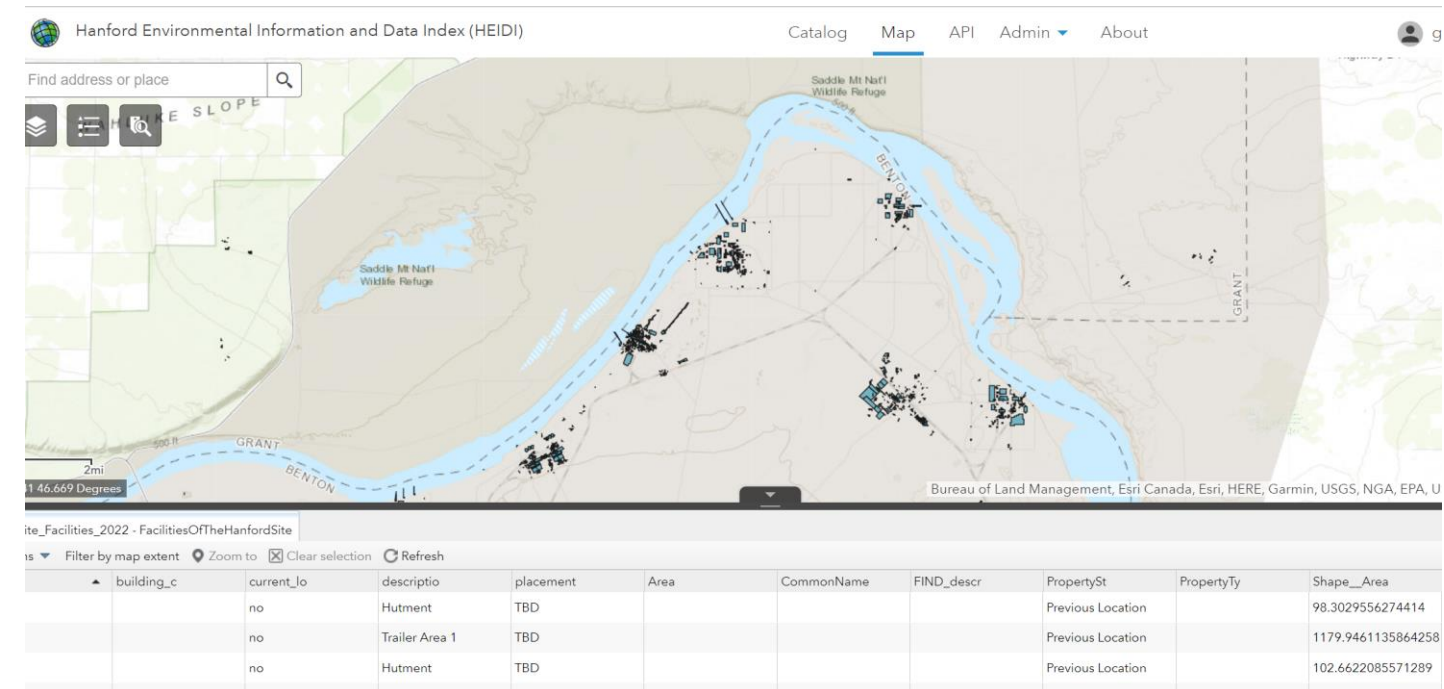
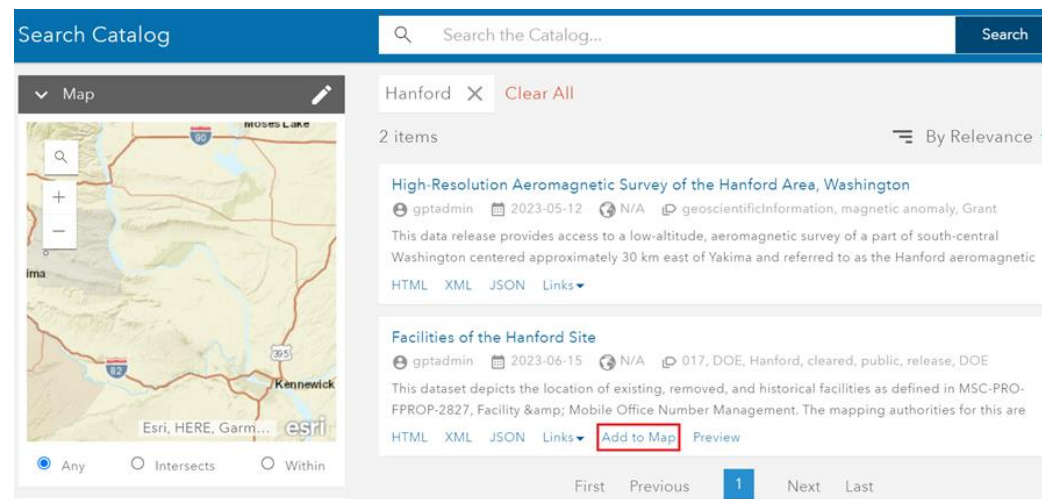






# Geoportal Server – Add to Map Feature

- Example: “Facilities of the Hanford Site” added to map



- Custom filter fields, linking databases to resources, and helpful navigation tips



**DEEP  
VADOSE ZONE  
PROGRAM**  
@PNNL

## Source of Origin

Environmental Dashboard Application (3)  
PHOENIX (2)  
[SOCRATES \(2\)](#)

Filter Guide



# Benefits of a Prototype and Full-Scale HEIDI

## Short Term

- A catalog implementation plan at the ready
- Metadata files for cataloging and template use
- Cyber security approval confirmation

## Long Term

- Refinement of geological framework models
- Capitalize on existing data to inform sampling plans and other environmental studies
- Effective, data-driven, strategic decisions to meet long term remediation goals



# Summary

- Findable, accessible, interoperable, and reusable (FAIR) data resources are critical
  - Cost effective use / re-use of data
- Metadata based on standards is the backbone of a data catalog
  - May need some customization for site-specific requirements
  - Describes data lineage/quality, resource location, access restrictions
  - Used to facilitate needs such as theming and consistent place/entity naming
- Catalog requirements should be developed for each implementation area and should be tested before full-scale deployment
  - Catalog requirements are generally site/case specific





## Summary

- Evaluated data catalog tools and found potential candidates
  - ArcGIS Enterprise Sites, Socrata, CKAN, and DKAN
  - Further testing led to Esri Geoportal Server
- Have built a prototype HEIDI with Hanford and external data resources
- Next: Complete the prototype with additional data and federated user authentication



Hanford Environmental Information and Data Index (HEIDI)



# Acknowledgements

- Additional PNNL contributors: Yusuf Afzal (software developer) and Julianne Schneider (research librarian)
- Funding for this work was provided by the U.S. Department of Energy Richland Operations Office under the Deep Vadose Zone (DVZ) Project







**DEEP  
VADOSE ZONE  
PROGRAM**  
@PNNL

# Thank You

Rebecka Bence  
[rebecka.bence@pnnl.gov](mailto:rebecka.bence@pnnl.gov)