P2P Network Webinar Series

That's So Meta: Metadata Workshop

Part 1: November 17, 2020 and Part 2: November 18, 2020

Program Updates

- Welcome
  - Ashley Ward, Internet of Water
- Take note for today's webinar
  - We are recording!
  - Other administrative notes
- Peer-to-Peer (P2P) Network
- Be sure to register for Part 1 and Part 2

That's so Meta

What are Metadata and why are they important?

Technical Workshop
November 17, 2020

Kyle Onda: kyle.onda@duke.edu

Outline

- What are Metadata?
- Why is metadata important?
- Structuring Metada
- Metadata creation & publishing tools

What are metadata?

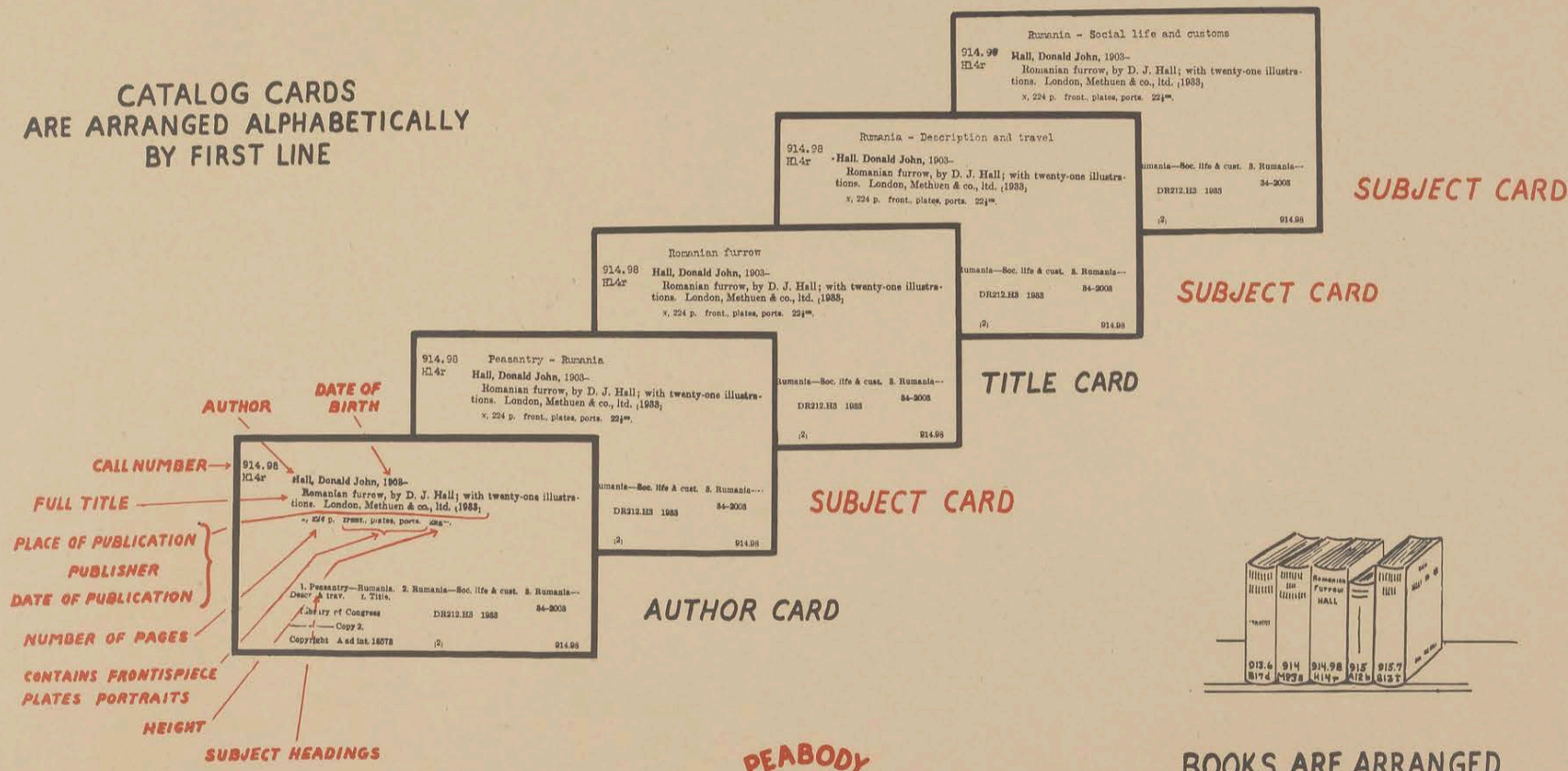
- “Data about data”

What are metadata?

- *Structured* data about data

WHAT THE CATALOG CARD TELLS US

CATALOG CARDS
ARE ARRANGED ALPHABETICALLY
BY FIRST LINE



PEABODY
VISUAL AIDS
PUBLISHED BY
FOLLETT BOOK COMPANY · CHICAGO

BOOKS ARE ARRANGED
FROM LEFT TO RIGHT
NUMERICALLY BY CALL NUMBER

Prepared under the direction of Miss Ruby Ethel Cundiff for the Peabody Library School course in Teaching the Use of the Library.
Committee: Juanita McMillan, chairman; Esther J. Sheets; Maralea Arnette; Lorainne Binkley; William Eury

Why is metadata important?

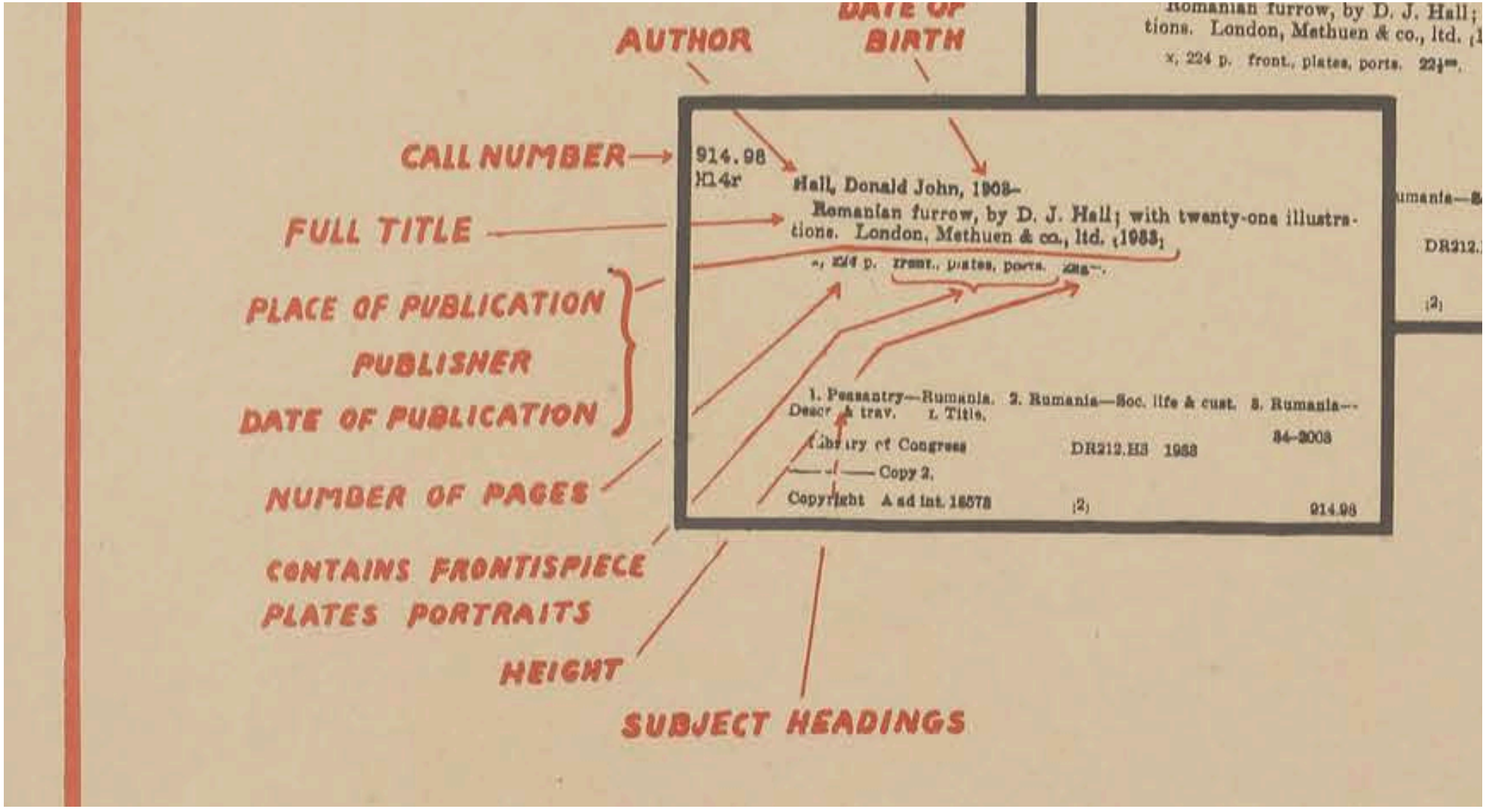
- *Discovery*
- *Exchange*
- *Administration*
- *Use*

Why is metadata important? - Discovery

- Need to be able to search for data by elements such as

- - Identifier
- - Type
- - Title
- - Subject/Topic
- - Publisher
- - Author
- - Temporal coverage
- - Spatial coverage

Why is metadata important? - Discovery



Why is metadata important? - Exchange

- Need to be able to know how to send and receive data.
- - File formats
- - Data transfer interfaces
- - Download links

Why is metadata important? - Administration

- *Who is allowed to use this data, and for what purposes?*
- *How should this data be cited?*
- *What version of this data is it, and where can previous versions be found?*
- *Where should questions about this data be directed?*

Why is metadata important? - Use

- *Why was the data created?*
- *How was the data collected?*
- *What is the data about?*
- *What is the data quality?*
- *How are the column headers defined?*
- *What are the data values in each column restricted to?*

Structuring Metadata

- *Schema*
- *Metadata Standards*
  - *Content*
  - *Value*
  - *Structure*
- *Data Dictionary/ Feature Catalogs/ Entity & Attribute Information*

Structuring Metadata

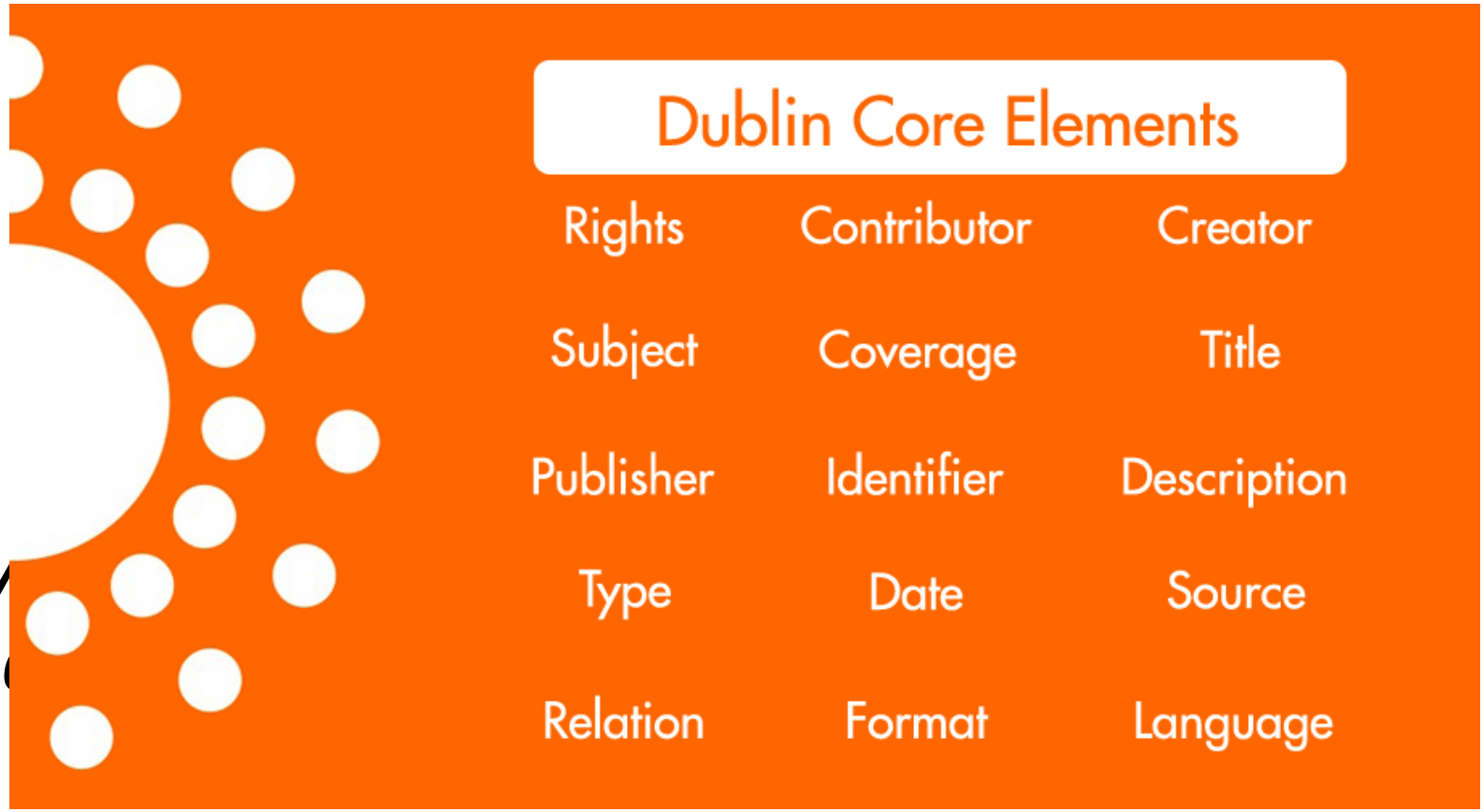
- *Schema*
- *Standards*
  - *Content*
  - *Value*
  - *Structure*
- *Data Dictionary/ Feature Catalogs/ Entity & Attribute Information*

List of metadata elements.

Structuring Metadata

Example: *https://www.dublincore.org/specifications/dublin-core/dces/*

- *Schema*
- *Standards*
  - *Content*
  - *Value*
  - *Structure*
- *Data Dictionary/ Attribute Inform*



Structuring Metadata

- *Schema*
- *Standards*
  - *Content*
  - *Value*
  - *Structure*
- *Data Dictionary/ Feature Catalogs/ Entity & Attribute Information*

Definitions of metadata elements and how they should be generated

Structuring Metadata

Example: https://www.dublincore.org/specifications/dublin-core/dcmi-terms/#http://purl.org/dc/elements/1.1/type

Term Name: type	More details
URI	http://purl.org/dc/elements/1.1/type
Label	Type
Definition	The nature or genre of the resource.
Comment	Recommended practice is to use a controlled vocabulary such as the DCMI Type Vocabulary [DCMI-TYPE]. To describe the file format, physical medium, or dimensions of the resource, use the Format element.
Type of Term	Property
Note	A second property with the same name as this property has been declared in the dcterms: namespace. See the Introduction to the document DCMI Metadata Terms for an explanation.

- *Data Dictionary/ Feature Catalogs/ Entity & Attribute Information*

Structuring Metadata

- *Schema*
- *Standards*
  - *Content*
  - *Value*
  - *Structure*
- *Data Dictionary/ Feature Catalogs/ Entity & Attribute Information*

Controlled vocabularies (allowable term lists)

Structuring Metadata

Example: ISO 19115 **Topics**

- *Schema*
- *Standards*
  - *Content*
  - *Value*
  - *Structure*
- *Data Dictionary/ Feature Catalogs/ Entity & Attribute Information*

Farming
Biota
Boundaries
Climatology
Economy
Elevation
…

Structuring Metadata

- *Schema*
- *Standards*
  - *Content*
  - *Value*
  - *Structure*
- *Data Dictionary/ Feature Catalogs/ Entity & Attribute Information*

Definitions of formats to encode metadata so that it is machine readable

Structuring Metadata

- *Schema*
- *Standards*
  - *Content*
  - *Value*
  - *Structure*
- *Data Dictionary/ Feature Catalogs/ Entity & Attribute Information*

Examples: XML, JSON, JSON-LD

Structuring Metadata

- *Schema*
- *Standards*
  - *Content*
  - *Value*
  - *Structure*
- *Data Dictionary/ Feature Catalogs/ Entity & Attribute Information*

Descriptions of data attributes and allowed values. Embed directly in metadata or include link to document (website, pdf, csv, etc)

Structuring Metadata

Example: USGS Streamflow Statistics

attr

- **attrlabl:** site_no
- **attrdef:** A unique number for each site where hydrological data has been collected and streamflow statistics have been calculated.
- **attrdefs:** U.S Geological Survey

attrdomv

- **udom:** Unique 8-digit code identifying site where hydrologic data has been collected.

attr

- **attrlabl:** huc12
- **attrdef:** Hydrologic unit code 12-digit designation.
- **attrdefs:** U.S. Geological Survey

attrdomv

codesetd

- **codesetn:** Hydrologic Unit Code (HUC)
- **codesets:** https://nhd.usgs.gov/data.html

attr

- **attrlabl:** decade
- **attrdef:** A 4-digit number indicating the beginning calendar year of a decade. For example, a row with '1960' listed for decade contains streamflow statistics based on observed daily-mean streamflow values for the period January 1, 1960 through December 31, 1969 for a given site (site_no).
- **attrdefs:** Producer defined

attrdomv

- **udom:** A 4-digit number that represents the first year of the decade in which the streamflow data was collected. On site may have multiple decades of streamflow data.

https://usgs.ornl.gov/metadata/catalog/html/Lower_Mississippi-Gulf_Water_Science_Center/5b218db8e4b092d9652874c7.html

Metadata creation tools - Standards

- *Dublin Core (https://www.dublincore.org/specifications/dublin-core/dcmi-terms/)*
- *ISO 191** series*
- *FGDC Standards (https://www.fgdc.gov/standards/list)*
- *https://schema.org/datasets*

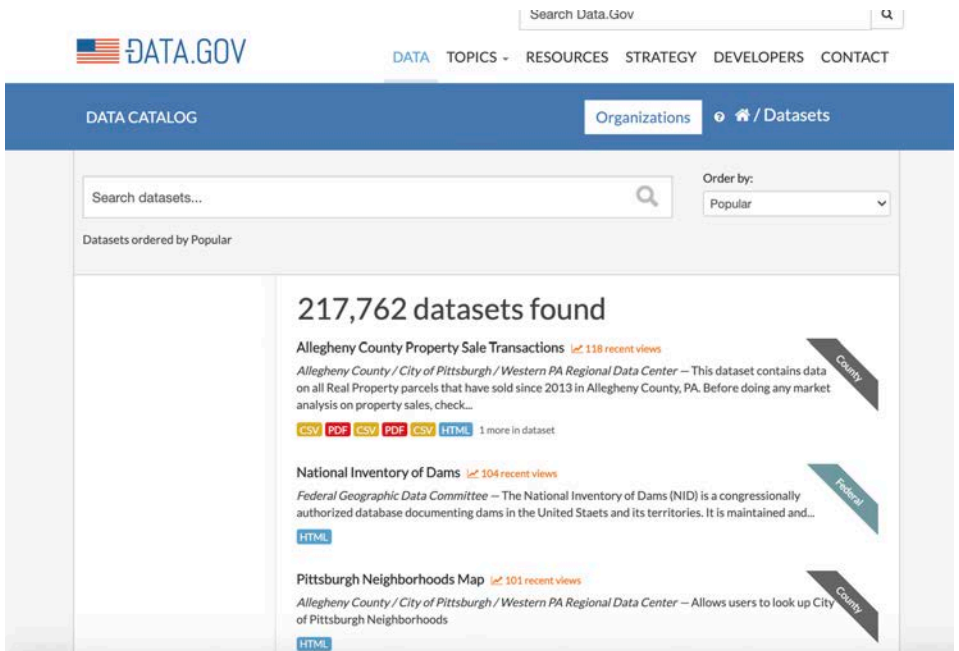
Metadata creation tools – Controlled vocabularies and data value lists

- *ODM2 Controlled vocabularies for hydrologic observations (http://vocabulary.odm2.org)*
- *NEMI catalog for water quality analytes and methods (https://nemi.gov)*
- *QUDT vocabularies for quantities, units, dimensions, and data types (http://www.qudt.org)*
- *Make and publish your own as needed*

Metadata Publishing Tools

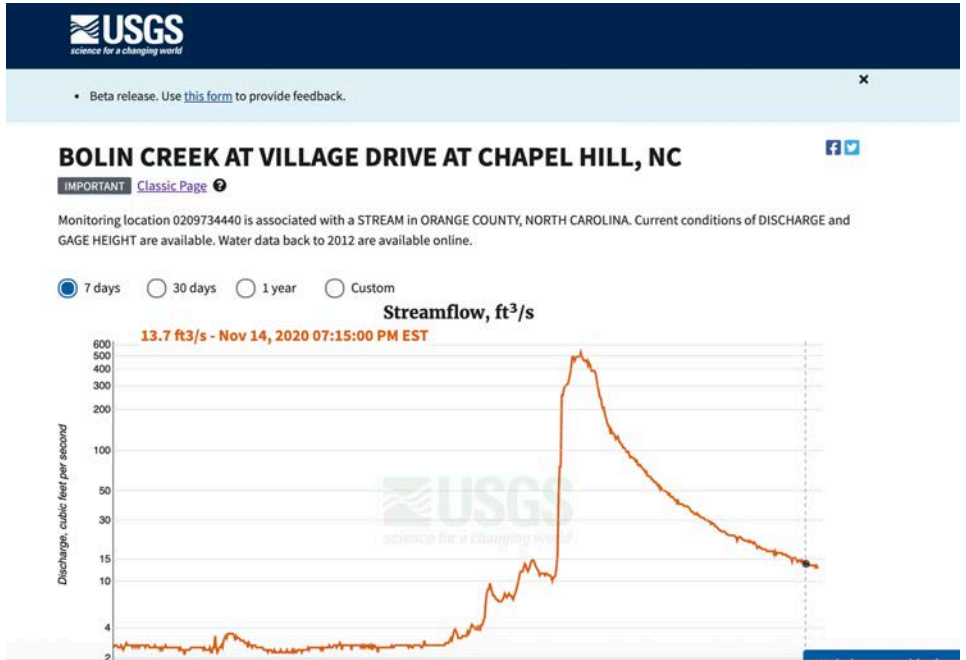
Data Catalogs

- *Publish and manage metadata from many datasets in a searchable catalog*



Embedded Microdata in Dataset Websites

- *Publish datasets or individual data points indexable by search engines*



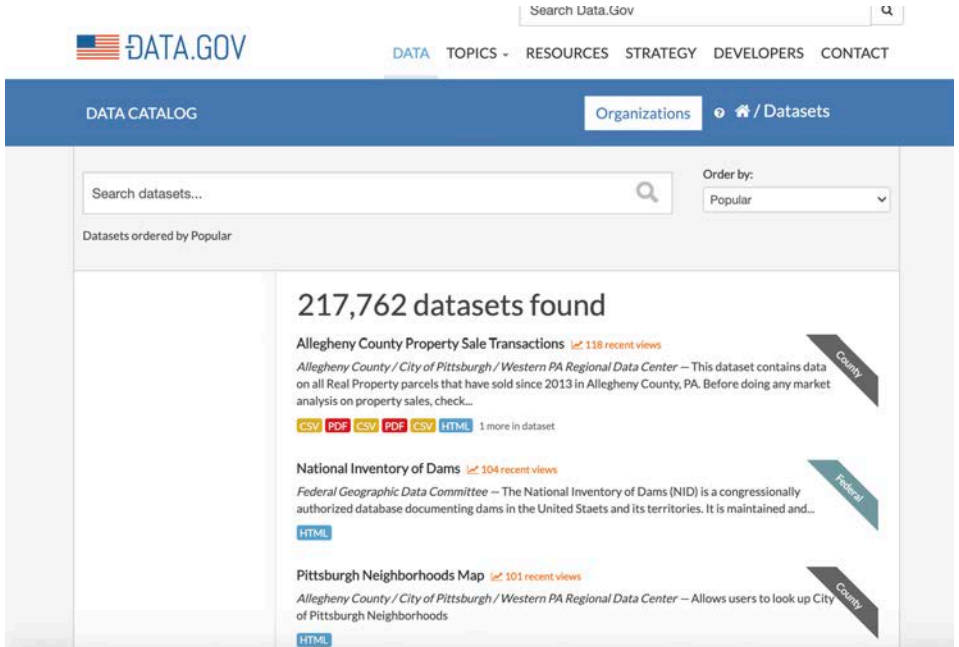
Metadata Publishing Tools

Data Catalogs

- *Publish and manage metadata from many datasets in a searchable catalog*

Embedded Microdata in Dataset Websites

- *Publish individual datasets indexable by search engines*



Metadata Publishing Tools – Data Catalogs

- *Full-service metadata catalogs and repositories*
  - *Dataverses (e.g. https://dataverse.harvard.edu)*
  - *Hydroshare (https://hydroshare.org)*

- *Self – deployable*
  - *CKAN (open source). (e.g. https://catalog.newmexicowaterdata.org)*
  - *CKAN (cloud) (e.g. https://opengov.com, https://dataopian.com)*
  - *Geonode (open source) (e.g. https://geonode.wfp.org)*
  - *Socrata (paid) (e.g. https://data.colorado.gov)*
  - *ESRI ArcGIS Hub (e.g. https://data-ncdenr.opendata.arcgis.com)*

Metadata Publishing Tools – Embedded Microdata

- *Microdata guidance*
  - *jsonld.com (https://jsonld.com/json-ld-generator/)*
  - *Google Datasets (https://developers.google.com/search/docs/data-types/dataset)*

- *Feature-level metadata publishing/Landing page generators*
  - *OGC API-Features https://github.com/opengeospatial/ogcapi-features/blob/master/implementations.md*

CUAHSI-Supported Community Data Sharing and Management Tools

Dr. Jerad Bales

Executive Director, CUAHSI

IoW – CUAHSI – WaDE That's So Meta Technical Workshop.

CUAHSI (*Ku-WAH-see*)

- *Strengthening interdisciplinary collaboration in the water resources community,*
- *Empowering the community by providing critical infrastructure, and*
- *Promoting education in the water sciences at all levels*

Community Services: Resources and training to build capacity and extend capabilities.

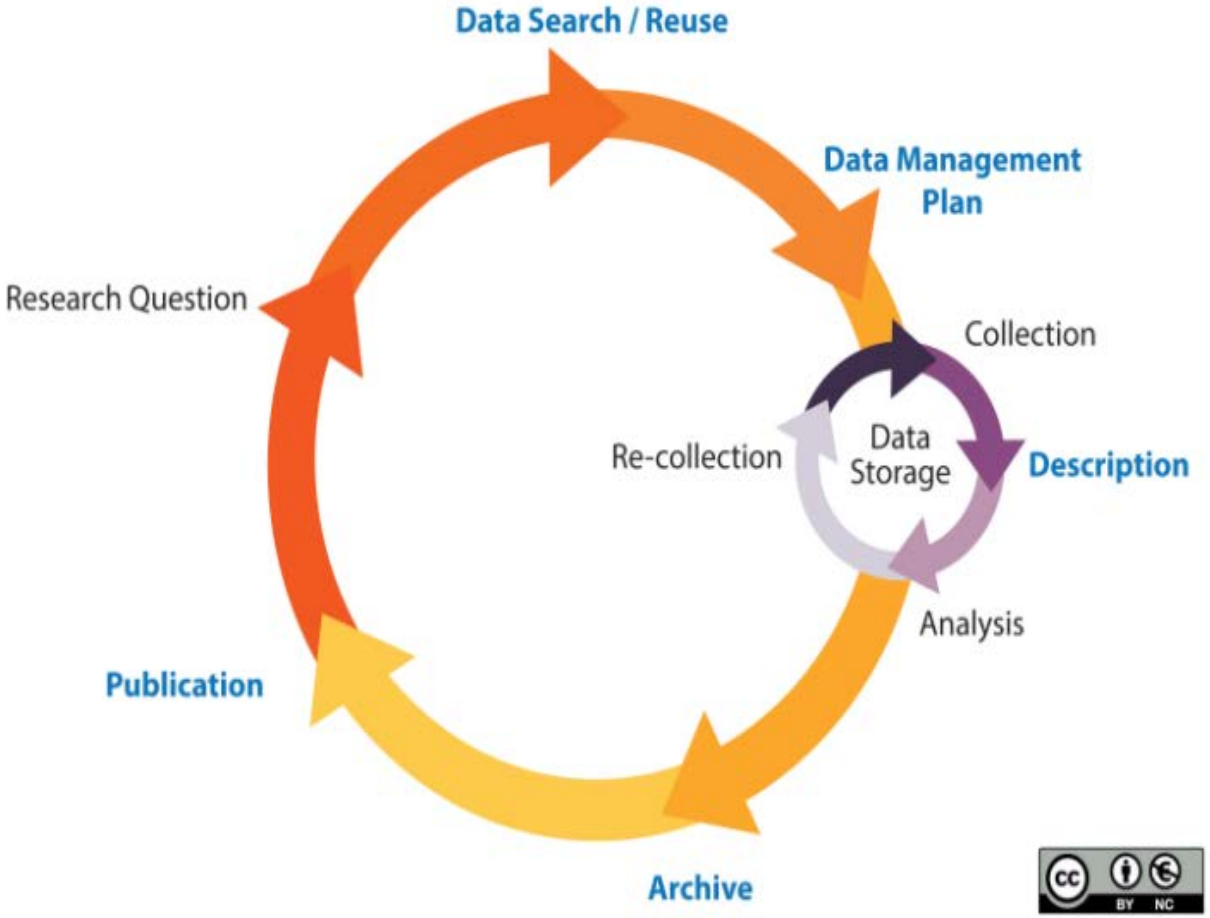
Data Services: Access to data and models to facilitate research, water-resources management, and education.

HYDROSHARE

jupyterhub

The Data Management Lifecycle

How are data handled from the beginning to the end of your project?




SCIENTIFIC DATA

OPEN

SUBJECT CATEGORIES
» Research data
» Publication characteristics

Comment: The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson *et al.*

UCSC University Library, Research Data Management

The Data Management Lifecycle

5 Essential Components of the DMP:

1. Anticipated Products
2. Data Formats & Standards
3. Dissemination, Access, and Sharing of Data
4. Re-Use, Re-Distribution, and Production of Derivatives
5. Archiving of Data

Data Management Plan

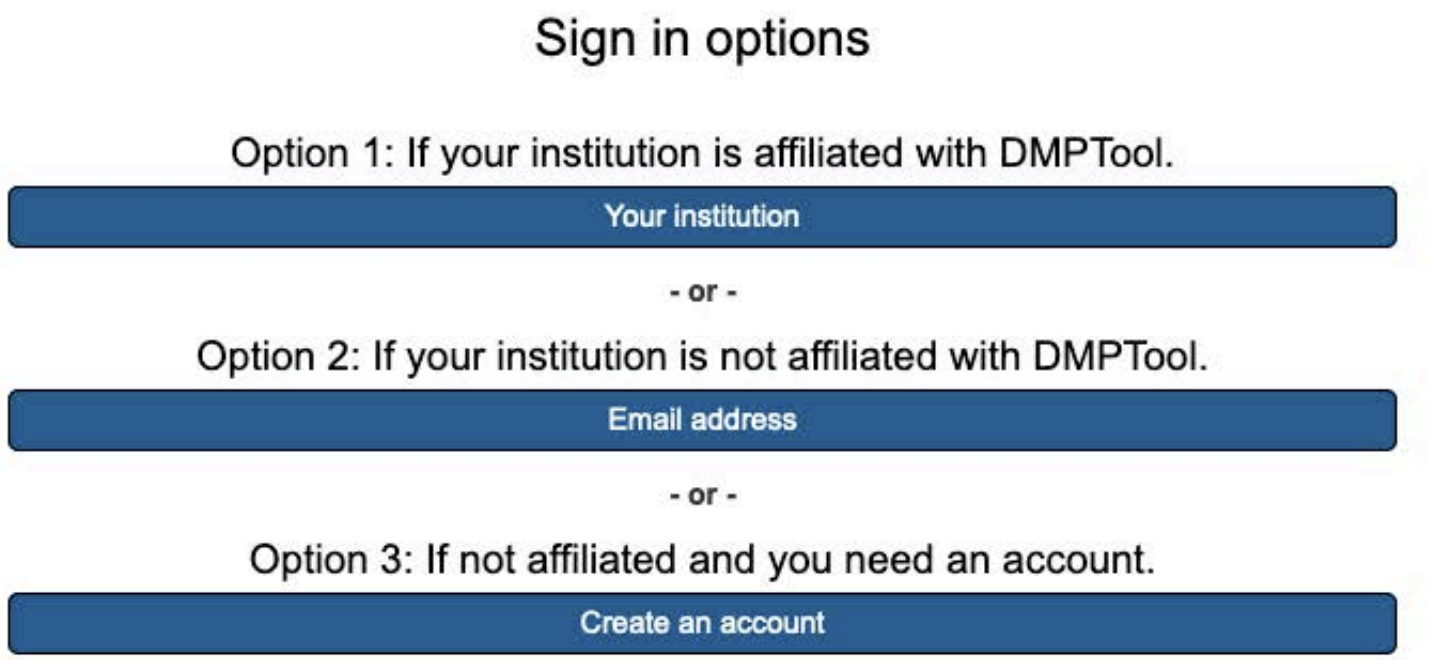
- **Start BEFORE work begins**
- **Be well documented and updated throughout process**
- **Covers lifecycle of data like collection, storage, analysis, sharing, publishing**

Hurricane Maria DMP

CUAHSI Website: *https://www.cuahsi.org/data-models/data-management-plans/*

Templates for Creation of DMP

DMPTool
Build your Data Management Plan



https://dmptool.org/get_started

https://www.cuahsi.org/data-models/data-management-plans/

A Little about MetaData

- ***Schema***: A list of ***elements*** are used to capture information about a resource.
- ***Standards*** tell us how to populate each element of the schema
  - **Content Standards**
  - **Data Value Standards, including controlled vocabulary.**
  - **Data Structure Standards: insures readability by both humans and machines.**

MetaData Make Data *Interoperable*

https://guides.library.ucsc.edu/c.php?g=618773&p=4306381

A Little more about MetaData

- ***Schema***: A list of ***elements*** are used to capture information about a resource.

Examples of metadata schema:

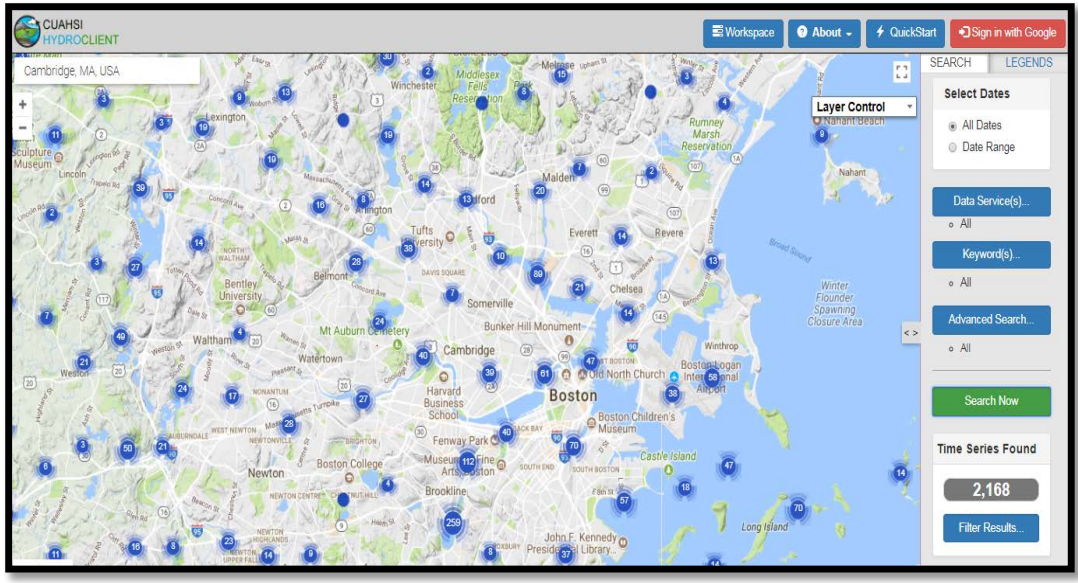
- ODM2 (HIS): ***https://github.com/ODM2/ODM2/wiki/Doc***
- Dublin Core Metadata (HydroShare):

https://www.dublincore.org/schemas/

https://help.hydroshare.org/creating-and-managing-resources/best-practices-for-describing-your-resource-with-metadata/

Data Search/Reuse: HydroClient--Discover Data

- **Discover, view, download time-series data all in the same format**
- **Obtain data from:**
  - **U.S. Federal Agencies (e.g. USGS, NOAA, NASA, EPA)**
  - **Academic projects**
  - **Non-profit watershed groups**
  - **International governments and projects**
- **Publish your time series data on HydroClient**
- **ODM2 Metadata Model (see paper in folder)**




HIS DEMO
Follow Along!

Data Workflow

- **Easily create a digital instance of a dataset or model**
- **Quickly share it with colleagues (perhaps privately at first)**
- **Add value through collaboration, annotation, and iteration**
- **Describe with metadata**
- **Eventually…share publicly or formally Publish (with doi) for reuse.**

Basic HydroShare Workflow

How it works

1

Create data

Collect your data using the same methods you use now. HydroShare supports a broad set of hydrologic data types.

2

Upload to HydroShare

Upload your data files to HydroShare through the web user interface. HydroShare will automatically extract as much metadata as it can from the files you upload.

3

Describe with metadata

Use HydroShare's simple metadata entry forms to finish describing your data so that your colleagues can find, access, and interpret it.

4

Share with colleagues

Choose who has access to the data and models you have uploaded to HydroShare. You can share with individual users or make your resources public for everyone to access.

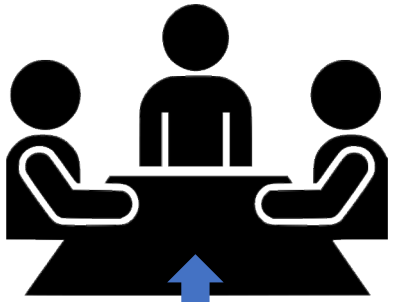
5

Permanently Publish

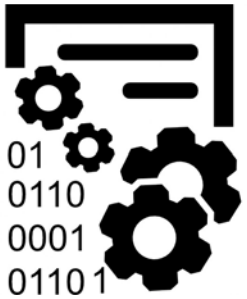
Obtain a Digital Object Identifier (DOI) so your work can be easily cited. Reference related journal publications in your metadata.

https://www.hydroshare.org/

Jeff, Amber, and Dave met and conceptualized a new study



Collected data over a period of several months



Amber compiled the data and uploaded to HydroShare

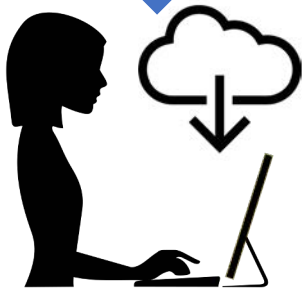
Dave downloaded the data and wrote Matlab code to implement the visualizations we needed – uploaded code to HydroShare

Finished writing the report and submitted for publication

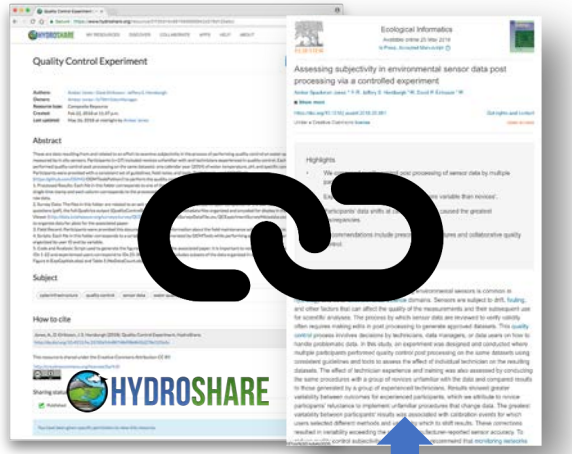
Reviews came back and we needed to revise the paper, the code, and the data

Life gets complicated

Amber downloaded the Matlab code and data to make the necessary changes. Report resubmitted.



Report and HydroShare resource (code/data) published with links!



Slide courtesy of Jeff Horsburgh

Resource Metadata

LCZO-Stream Water Chemistry, Streamflow / Discharge, Hysteretic response of solutes and turbidity at the event scale across forested tropical montane watersheds - Luquillo Experimental Forest (2016-2017)

Open with...

Authors: Adam Wymore | Miguel C Leon | James B Shanley | William H McDowell
Owners: Miguel C Leon
Resource type: Composite Resource
Storage: The size of this resource is 1.1 MB
Created: Mar 18, 2019 at 5:28 p.m.
Last updated: May 22, 2019 at 6:17 p.m. by Miguel C Leon
DOI: 10.4211/hs.f8420c1447fe440eb93e656b2db0b64d
Citation: See how to cite this resource

Sharing Status: Published
Views: 164
Downloads: 2
+1 Votes: 1 other +1 this +1
Comments: No comments (yet)

Subject Keywords

hurricanes and tropical storms | LCZO | hysteresis | github | specific conductivity | sensors | turbidity | Luquillo Critical Zone Observatory

Resource Level Coverage

Spatial

Coordinate System/Geographic Projection:
WGS 84 EPSG:4326

Coordinate Units:
Decimal degrees

Place/Area Name:
Luquillo experimental forest

North Latitude	East Longitude
18.4705°	-65.5554°
South Latitude	**West Longitude**
18.1458°	-65.9660°

Temporal

Start Date: 08/06/2016
End Date: 09/21/2017

References

Sources

Derived From: https://doi.org/10.3389/feart.2019.00126

Credits

Funding Agencies

This resource was created using funding from the following sources:

Agency Name	Award Title	Award Number
NSF-EAR		0722476
NSF-EAR		1331841
NSF-DEB		0620919
NSF-DEB		1239764
NSF-DEB		1546686
NSF-EF		1442444
University of New Hampshire	New Hampshire Agricultural Experiment Station	1006760

How to Cite

Wymore, A., M. C. Leon, J. B. Shanley, W. H. McDowell (2019). LCZO-Stream Water Chemistry, Streamflow / Discharge, Hysteretic response of solutes and turbidity at the event scale across forested tropical montane watersheds - Luquillo Experimental Forest (2016-2017), HydroShare, https://doi.org/10.4211/hs.f8420c1447fe440eb93e656b2db0b64d

Copy

Author Identifiers

ORCID
https://orcid.org/0000-0002-1341-5681

GoogleScholarID
https://scholar.google.com/citations?user=ScWTFoQAAAAJ&hl=en

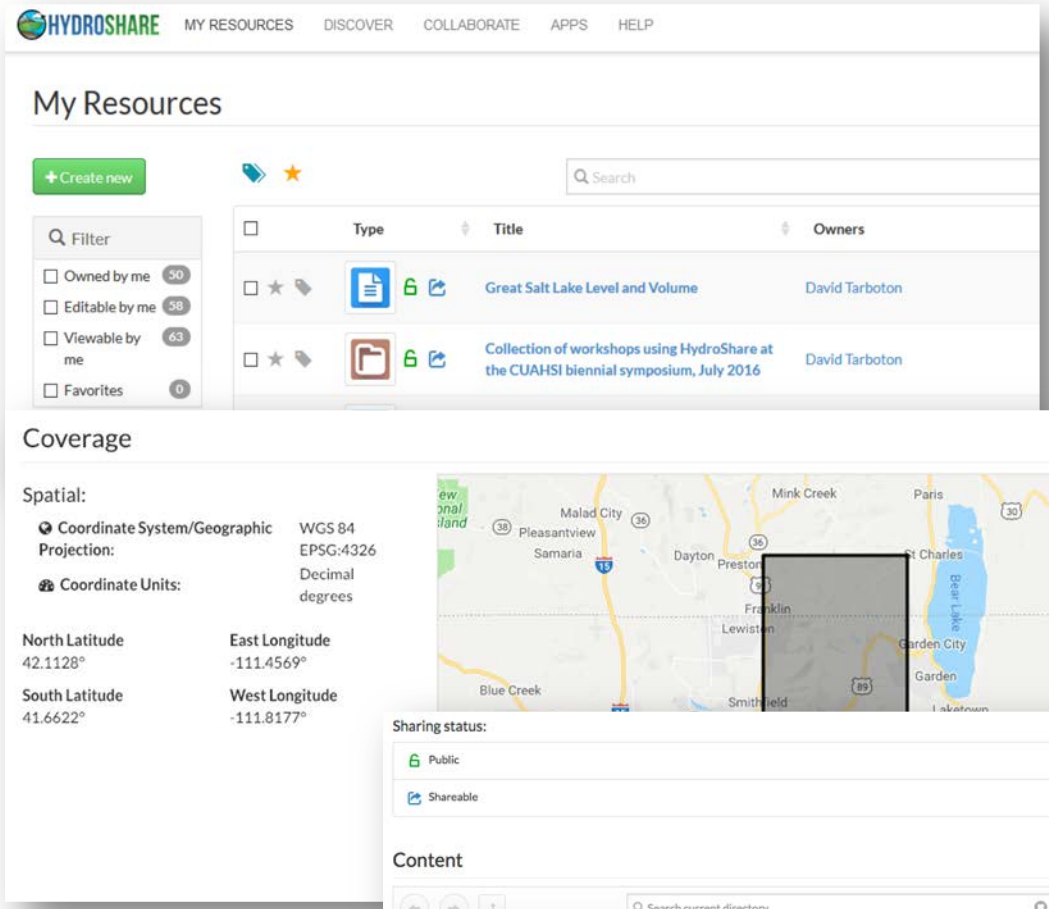
ResearchGateID
https://www.researchgate.net/profile/Anthony_Castronova

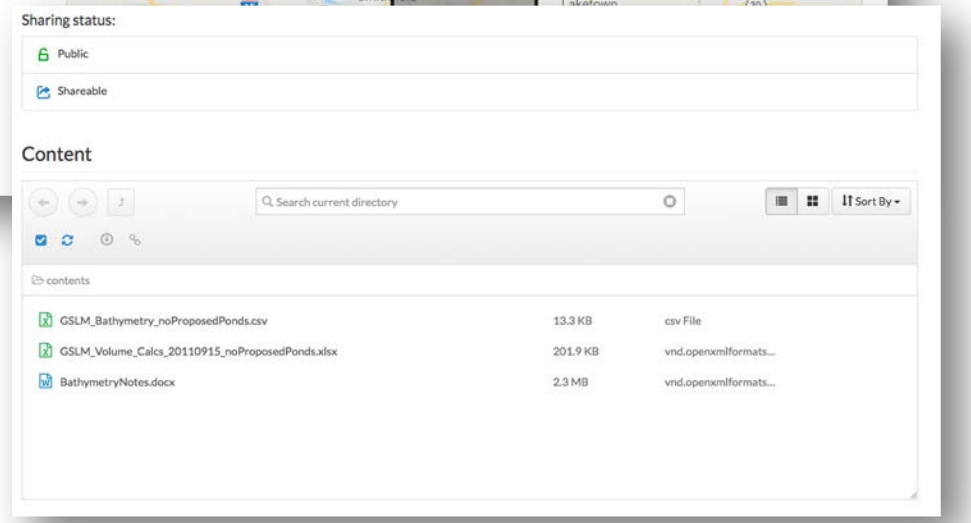
Resources, Groups, and Communities

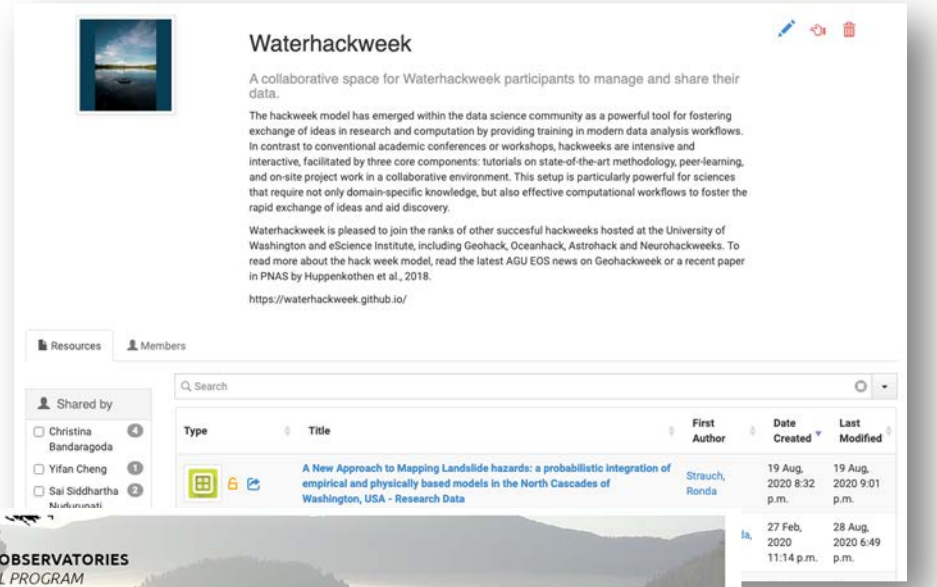
Group

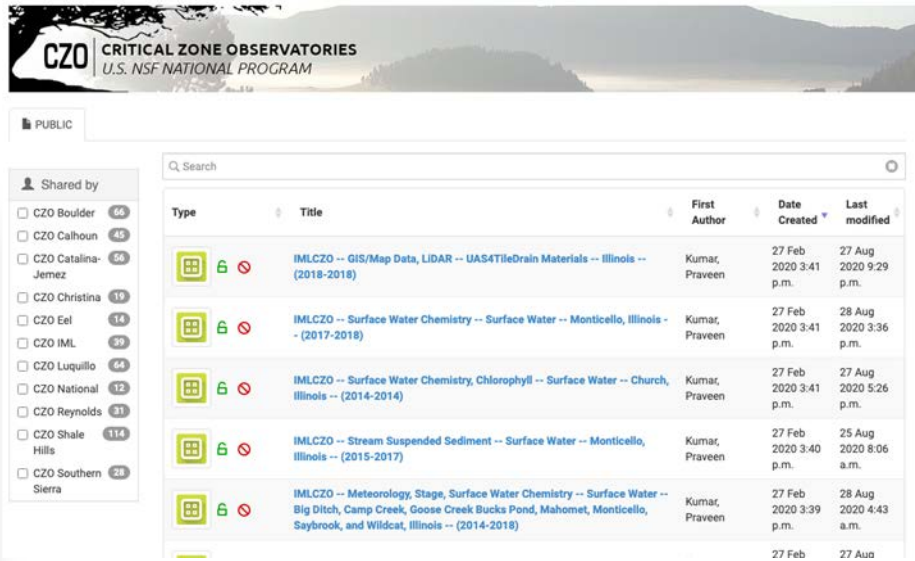
Community

Resource









Interfacing with HydroShare Data

(Application Programming Interface)

HYDROSHARE

Data

HydroShare GIS

OPeNDAP

MATLAB Online

jupyter

CUAHSI JupyterHub

binder

National Water Model …

Data Rods Explorer App

CUAHSI HIS

Sharing hydrologic data

CUAHSI

UNIVERSITIES ALLIED FOR WATER RESEARCH

Share workflows and models

CUAHSI Domain Subsetter Workflow - NWM

Open with...

Authors:	Danielle Tijerina | Anthony Michael Castronova	**Sharing Status:**	Private
Owners:	Anthony Michael Castronova | Danielle Tijerina	**Views:**	30
Resource type:	Composite Resource	**Downloads:**	11
Storage:	The size of this resource is 85.5 KB	**+1 Votes:**	Be the first one to +1 this.
Created:	Jun 18, 2019 at 7:53 p.m.	**Comments:**	No comments (yet)
Last updated:	Jun 20, 2019 at 6:06 p.m. by Anthony Michael Castronova		
Citation:	See how to cite this resource		

Abstract

The purpose of this resource is to provide a workflow of how to use the CUAHSI Domain Subsetter - NWM edi
(subset.cuahsi.org) introduces a collaborative effort for preparing, publishing, and sharing subsets of the Nati
watershed scales. Our hope is that these efforts will lower the barrier of entry for using and applying these mo
scientists from a diverse spectrum of expertise. With a combination of modern cyberinfrastructure technique
tools, researchers will have access to subsets of National Water Model information that would otherwise requ
resources. This work provides the foundation onto which similar efforts can be applied to other large-scale m

Collect NLDAS

This notebook demonstrate
onto an existing WRF-Hydr

1. Objective

This notebook illustrates th
domain. This include:

1. Subsetting NLDAS dat
2. Downloading NLDAS d
3. Regridding NLDAS usi

Running a WRF-Hydro Simulation

This notebook demonstrates how to executed the WRF-Hydro model using the `DOMAIN` and `FORCING` data collected in the two previous notebooks. We'll be using a containerization technology called Docker to simplify this process and eliminate the need for compiling any source code.

> Docker is a tool designed to make it easier to create, deploy, and run applications by using containers. Containers allow a developer to package up an application with all of the parts it needs, such as libraries and other dependencies, and ship it all out as one package. By doing so, thanks to the container, the developer can rest assured that the application will run on any other Linux machine regardless of any customized settings that machine might have that could differ from the machine used for writing and testing the code.

Please see the following link for further description of Docker:
https://opensource.com/resources/what-docker

Before we begin, we need to make sure that Docker is installed on your computer. If you're unsure if Docker is already installed on your computer, open a terminal and type `docker -v`. If it's installed, you should see output similar to: `Docker version 18.09.2, build 6247962`.

Install Docker by following the instructions at this link.

```
In [ ]: # verify the docker is installed and running
        !docker -v
```

Subsetting the National Water Model Domain

Authors: Tony Castronova acastronova@cuahsi.org
Danielle Tijerina dtijerina@cuahsi.org

jupyterhub Physical_Hydrology_Homework_complete_version (autosaved)

Logout | Control Panel

File Edit View Insert Cell Kernel Help | Not Trusted | No Kernel

Run | Markdown | **Memory:** 79 / 4096 MB

Problem 2. Uncertainty

Do your coding and calculation below.

Let's create a dataframe of the given inputs using R. For this, we will use **data.frame** function of R. The c function in R is used to create a vector with values you provide explicitly.

```
In [ ]: data = data.frame("Watershed" = c("Connecticut", "Yukon", "Euphrates", "Mekong"),
                       "Area.km2" = c(20370, 932400, 261100, 663000),
                       "Precipitation.mmperyr" = c(1100, 570, 300, 1460),
                       "Streamflow.cms" = c(386, 5100, 911, 13200),
                       "Precipitation_RelErr"= c(10, 20, 10, 15),
                       "Streamflow_RelErr" = c(5, 10, 10, 5))
        data
```

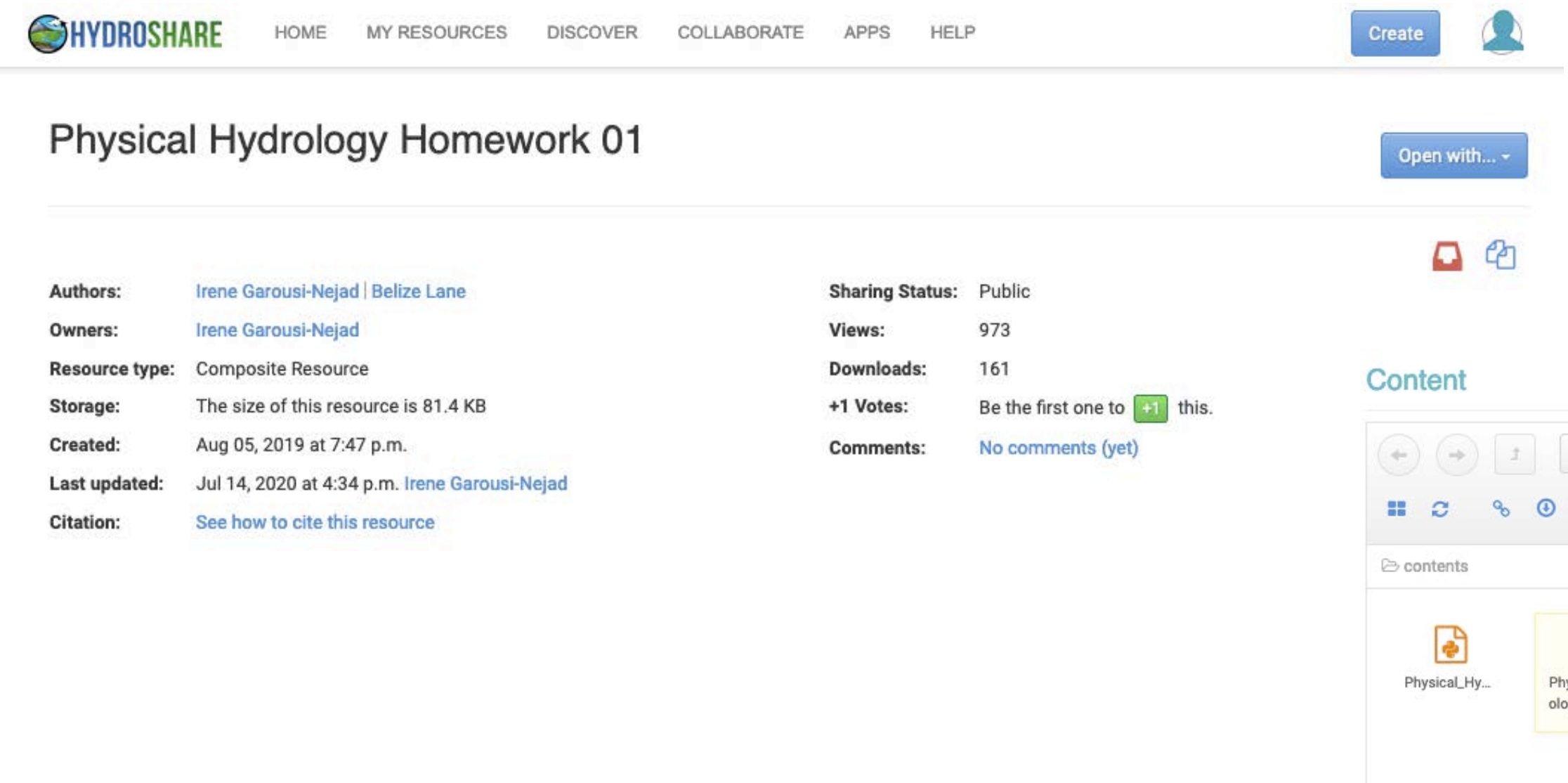
To calculate the long-term average evapotranspiration, we need the long-term average precipitation and the long-term average streamflow values. Note that these two components should have the same units when using in a water balance equation. The given precipitation data has the unit of $\frac{mm}{yr}$. While the streamflow data has the unit of $\frac{m^3}{s}$. In the cell below, we read the column of steamflow from the dataframe and convert its unit and then save the results as a new column that is added to the dataframe.

To read a column from a dataframe, use $ sign after the name of the dataframe followed by the name of the column. Look at the cell below.

```
In [ ]: data$Streamflow.cms
```

To add a new column to an existing dataframe, use the name of the dataframe and the $ sign followed by the name of the new column. Look at the cell below. It adds a new column called Area.m2 to the dataframe. The values for this column are the drainage areas with the unit of $\mathbf{m^2}$.

Educational Applications



Problem 2. Uncertainty

Do your coding and calculation below.

Let's create a dataframe of the given inputs using R. For this, we will use **data.frame** function of R. The c function in R is used to create a vector with values you provide explicitly.

In [ ]:
```
data = data.frame("Watershed" = c("Connecticut", "Yukon", "Euphrates", "Mekong"),
                  "Area.km2" = c(20370, 932400, 261100, 663000),
                  "Precipitation.mmperyr" = c(1100, 570, 300, 1460),
                  "Streamflow.cms" = c(386, 5100, 911, 13200),
                  "Precipitation_RelErr"= c(10, 20, 10, 15),
                  "Streamflow_RelErr" = c(5, 10, 10, 5))
data
```

To calculate the long-term average evapotranspiration, we need the long-term average precipitation and the long-term average streamflow values. Note that these two components should have the same units when using in a water balance equation. The given precipitation data has the unit of $\frac{mm}{yr}$. While the streamflow data has the unit of $\frac{m^3}{s}$. In the cell below, we read the column of steamflow from the dataframe and convert its unit and then save the results as a new column that is added to the dataframe.

To read a column from a dataframe, use $ sign after the name of the dataframe followed by the name of the column. Look at the cell below.

In [ ]:
```
data$Streamflow.cms
```

To add a new column to an existing dataframe, use the name of the dataframe and the $ sign followed by the name of the new column. Look at the cell below. It adds a new column called Area.m2 to the dataframe. The values for this column are the drainage areas with the unit of $\mathbf{m^2}$.

HYDROSHARE

MY RESOURCES DISCOVER COLLABORATE APPS HELP ABOUT

About HydroShare

What?

HydroShare is a system operated by The Consortium of U
enables users to share and publish data and models in a
shareable and discoverable manner. HydroShare include
content in HydroShare providing users with a gateway to

With HydroShare you can: share data and models with co
manipulate a broad set of hydrologic data types and mode
aggregate resources into collections; discover and access
programming interface (API) to programmatically access
models with data in HydroShare.

Who?

HydroShare is freely accessible to everyone, but primarily
others interested in water-resources data and tools.

Why?

The goal of HydroShare is to advance hydrologic science
products resulting from their research and/or data collecti
use data beyond the purpose for which it was originally co
investments; manage, archive, and publish data in a disco

How?

HydroShare has been, and continues to be, developed th
primary grants supporting HydroShare development were
currently through an NSF award to CUAHSI (EAR-133860
1664061, OAC-1664018, and OAC-1664119 (2017-2021)

HydroShare Support

Get quick answers to the most common questions about how to get started with and use HydroShare.

Introduction to HydroShare
Learn more about what you can do with HydroShare

HydroShare Resource Types
Learn more about data and model resources in HydroShare

Uploading and Publish Data
Create new datasets an models and describe wit metadata

Sharing and Privacy
Share resources with your colleagues, control who has access, and learn how to formally publish

Discovering Resources
Discover datasets and models shared by others

Participation and Collaboration
Create user groups and collaborate with your colleagues

Apps

Modeling

About HydroShare

Email: help@cuahsi.org

jdbales@cuahsi.org


HydroShare DEMO
Follow Along!

Q & A

Please submit your *content-related questions* in the webinar's **Q&A box**. The moderator will read your question aloud.

Administrative questions can be placed in the **"Chat" box**.

Final Notes

IoW Contact Information:
internetofwater@duke.edu
https://internetofwater.org/

SAP2P Network:
Website: https://internetofwater.org/state-agency-p2p-network/
Webinars: https://internetofwater.org/webinars/

Join us for continued conversation at the IoW P2P Forum:
https://p2pforum.internetofwater.org/

Poll: Please participate in a quick poll!

Follow-Up Information

- Links to the webinar recording and slides will be distributed once posted

Thank you for your participation!