portage
SERVICES PARTAGÉS POUR LES DONNÉES DE RECHERCHE
SHARED STEWARDSHIP OF RESEARCH DATA

Beginning With the End in Mind: Building Documentation and Metadata to support Data Deposit and Preservation

Erin Clary (Curation Coordinator, Portage Network)
Krysha Dukacz (Data Manager, Global Water Futures)

September 29, 2020

The “Rime of the Modern Researcher”

Data, data everywhere… but where is the bit I need?

… can I figure out the coding system?

.... is it in a format I can use?

… is it appropriate to my purpose?

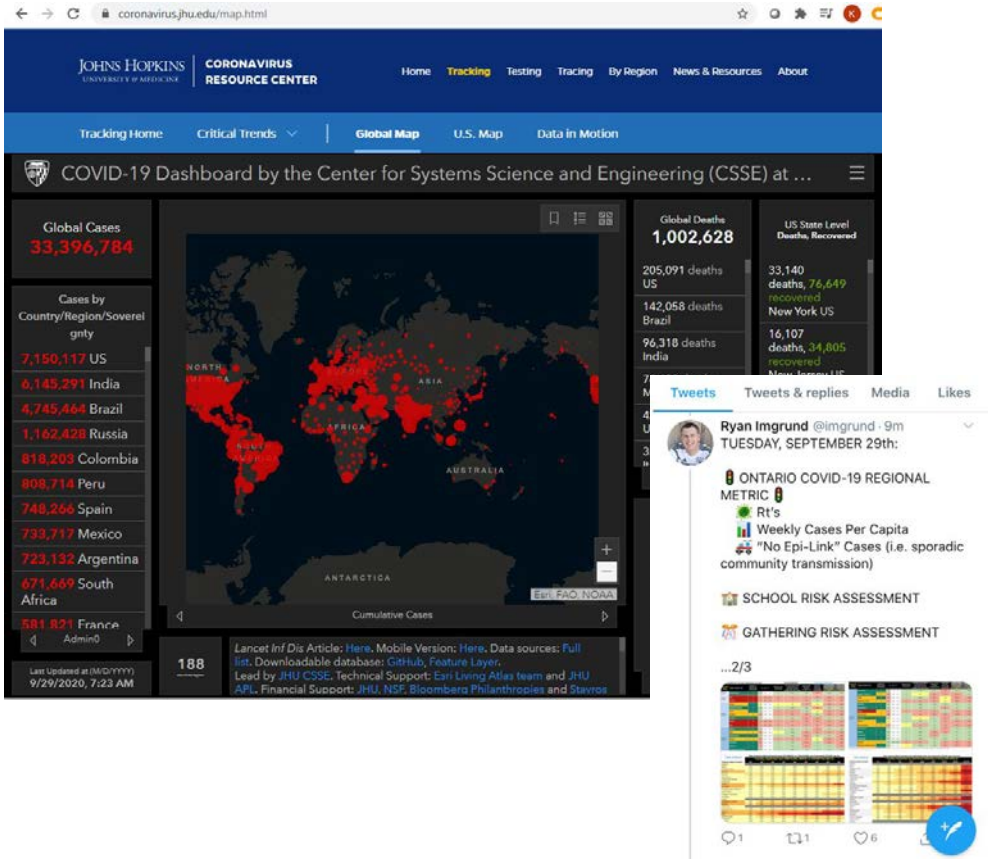
... am I allowed to share it?

The Call for Sharing Publicly Funded Data

Journalists need to know more about COVID-19 transmission, for the sake of all of us

(Sylvia Stead, The Globe and Mail, September 26, 2020)

- No national data standards and needing to source data from provinces, territories and regions.
- Took 10 days to collect data and make it comparable
- Availability of data is spotty, constraints and variables are defined differently
- Variability in data delivery formats: Excel, PDFs, manual copy/paste from mapping application



The Call for Sharing Publicly Funded Data

A new flow for Canadian young hydrologists: Key scientific challenges addressed by research cultural shifts
(Aubry-Wake et al., Hydrological Processes. 2020;34:2001–2006)

- Data collection often focused on long term sites, concentrating data control because of costs
- “Hidden” data: unprocessed, fragmented between publications, proprietary, distributed among various government bodies. Data are difficult to find (word of mouth, who you know)
- Support for data sharing,“including appropriate metadata is essential to effective data-sharing”

North

Arctic understanding limited by patchy field work, scientist says

Areas of Canadian Arctic poorly sampled, could mean faulty assumptions about rate of climate change

Walter Strong · CBC News · Posted: Aug 13, 2018 2:00 AM CT | Last Updated: August 13, 2018

Funder Driven Requirements - Show me the money

Tri-Agency Statement of Principles on Digital Data Management:

Promote excellence in digital data management practices and data stewardship in agency-funded research

Researchers expected to consider:

Data Management Planning	Constraints and Obligations (commercial, legal, ethical)
Adherence to Standards	Collection and Storage
Metadata	Preservation, Retention and Sharing
Timeliness	Acknowledgement and Citation
Efficient and Cost Effective	

http://www.science.gc.ca/eic/site/063.nsf/eng/h_83F7624E.html?OpenDocument

Impending Tri-Agency RDM policy

Encourage a culture of strong data management associated with research excellence

Increase the Canadian researchers recognized/rewarded for data as a valued product of research

Equip Canadian researchers to engage in international research

Increase institutional capacity to support the management of the data produced by researchers

Increase ability for research data to be archived

Reproducibility, Discoverability, Reusability

https://www.ic.gc.ca/eic/site/063.nsf/eng/h_97610.html

The Cost of Data

Cost of Data

Personnel (salaries, benefits)

Infrastructure and Logistics (instrumentation, travel costs, processing, incentives)
Data security, storage, backup
Data cleaning
Data Management

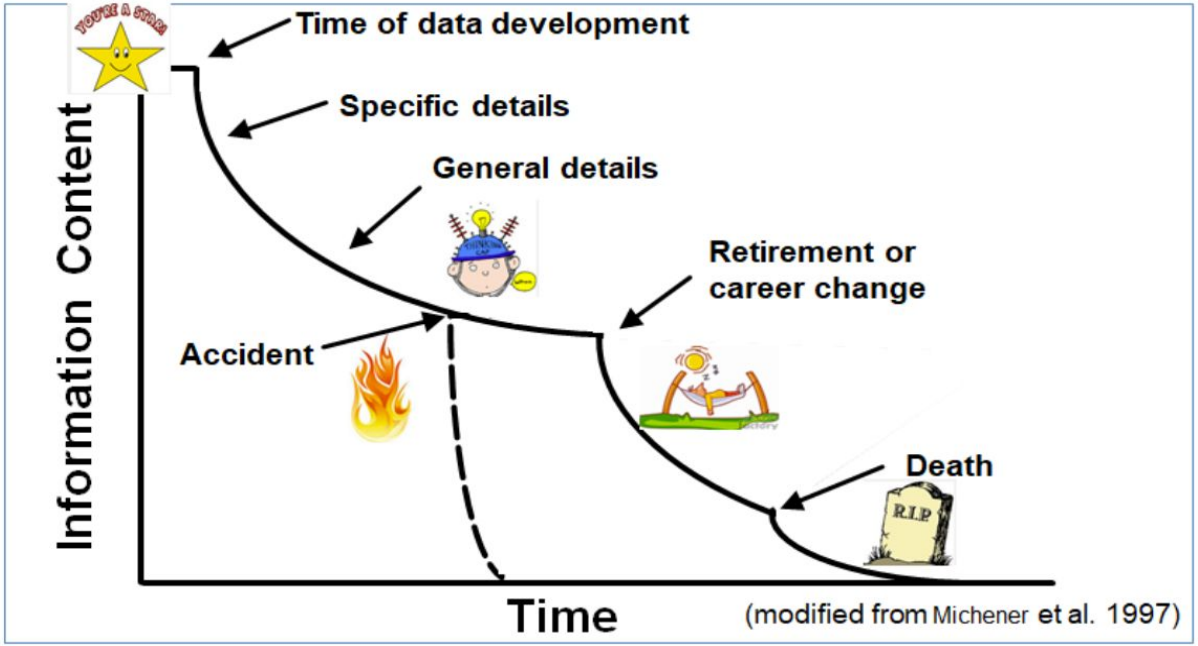
Blood, sweat and tears

Sometimes irreplaceable (one-time opportunity to collect)

The Value of Data

Value of Data

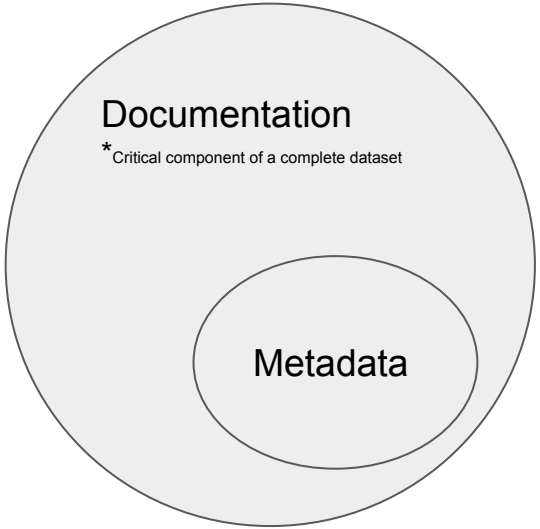
- The value of data is in the information that can be derived from its use.
- The information contained in data requires documentation to provide context
- Value is augmented by Reuse



DataONE Education Module: Metadata. DataONE. Retrieved Sept24, 2020. From http://www.dataone.org/sites/all/documents/L07_Metadata.pptx

What is Metadata?

"**Metadata** are a subset of core data **documentation**, which provides standardised structured information explaining the purpose, origin, time references, geographic location, creator, access conditions and terms of use of a data collection" (UK Data Archive).



INFORMATION	ABOUT DATA
WHO?	-Collected the data? -Processed the data? -To contact for more information? -Owns the data?
WHAT?	-Are the data about? -Parameters were measured? -Format are the data in? -Is the data quality? -Are the use constraints? -Are the appropriate uses?
WHEN?	-Were the data collected? - Were the data processed?
WHERE?	-Were the data collected ? -Are the data held?
WHY?	-Were the data collected?
HOW?	-Were the data collected / processed? -Do I access the data? -Were the data quality assessed?

ftp.ncddc.noaa.gov › pub › Metadata › DISL_June30

Need to Know for each dataset:

- Where it is
- How to access it
- What it can be used for
- Known issues/Quality
- Collection Methods
- Ethical /Privacy issues
- Licensing
- How to cite

Beginning with the End in Mind...

How can the value of data be protected…

- Figure out what is needed in the end and plan out the steps to get there.
- Make a plan - Formal Data Management Plan (DMP) or informal process
- Reuse already contained in your research plan.

ecohydrology research group | water INSTITUTE | UNIVERSITY OF WATERLOO | portage

Data Management Plan for Ecohydrology Research Group

Abstract

This exemplar DMP was created by the Ecohydrology Research Group at the University of Waterloo, with the purpose of providing a standardized DMP model for its many multi- and cross-disciplinary research projects. Here, fundamental processes and procedures core to the ERG have been incorporated as a starting point for its faculty, staff, and HQPs (Highly Qualified Personnel) when developing their research projects. This exemplar demonstrates the utility of a model DMP being used by a research group or collective to maintain best practices in data management.

Principal Investigators:

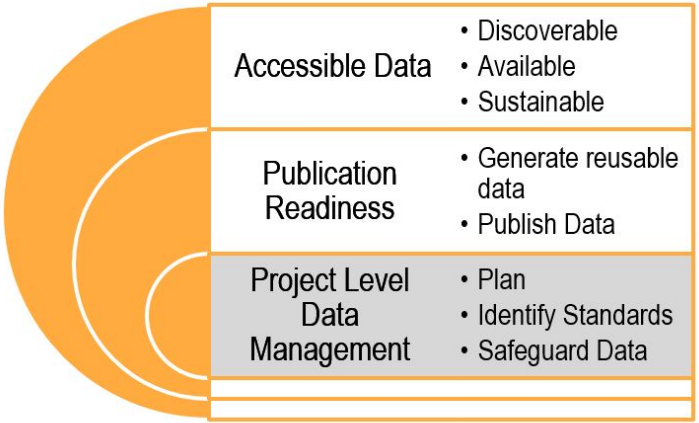
Philippe van Cappellen, Fereidoun Reza Nezhad

DMP exemplars will o be available on the Portage Network site soon. Excerpt courtesy of Bhaleka Persaud and the University of Waterloo Ecohydrology Research Group

DMP considers:

- Ethics
- Data collection
- Licensing
- Metadata
- Documentation
- Sharing
- Storage

Portage DMP Assistant



Benefits of Well-documented Data

Researcher	Publisher	Funders
Efficiency	Rigorous peer review	Transparency
Ease of collaborative work	Validation and reproducibility	Accountability
Credit and impact	Defensible authority	Return on Investment

This is the End…

Sharing with others (just metadata piece or whole datasets)

Mackenzie DataStream

FRDR DFDR
FEDERATED RESEARCH DATA REPOSITORY
DÉPÔT FÉDÉRÉ DE DONNÉES DE RECHERCHE

AMERIFLUX

PANGAEA.

CIOOS
CANADIAN INTEGRATED OCEAN OBSERVING SYSTEM

Scholars Portal Dataverse

zenodo

Principles to consider

Personal accomplishment and reliability, journal publication

- DOIs for Publication
- Funding Compliance
- Retraction Avoidance

*Tweet used with permission

Tools for Building Metadata

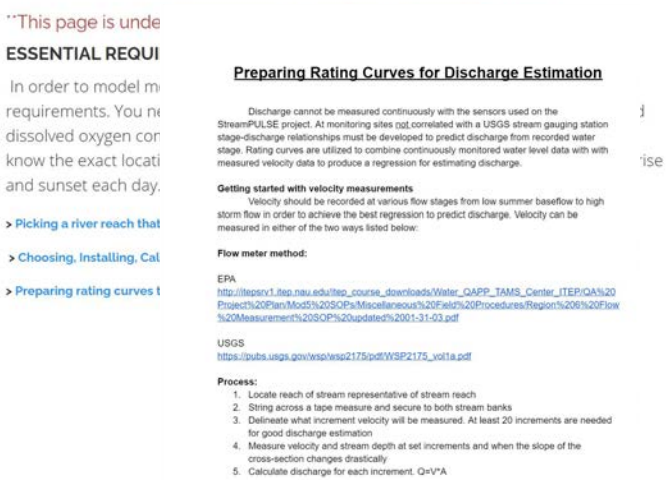
- Standard Operating Procedures
- Log files or Field Notes
- File Organization
- File Naming
- Analysis Scripts
- README
- Metadata standards (General and Disciplinary)
- Collaborative platforms

Standard Operating Procedures

Standard Operating Procedures (SOPs) provide step by step instructions on how to successfully complete a task. This enables consistency over time and among different people. Common examples include:

- Field Visit SOP may include what to bring, tasks that need to be completed at each site (download data, check sensors, take sample, potential issues/solutions)
- Instrument Setup - location, connectivity, calibration, expected output/QA/QC, troubleshooting
- Data Management - Folder Structures; Protocols for uploading field data; Storage locations (3-2-1 backup); File Naming, Transition expectations; Data Acquisition steps; Data upload

Standard Operating Procedures - Examples and Resources



"This page is unde

ESSENTIAL REQUI

In order to model m
requirements. You n
dissolved oxygen cor
know the exact locati
and sunset each day.

> Picking a river reach that

> Choosing, Installing, Cal

> Preparing rating curves t

Preparing Rating Curves for Discharge Estimation

Discharge cannot be measured continuously with the sensors used on the StreamPULSE project. At monitoring sites not correlated with a USGS stream gauging station stage-discharge relationships must be developed to predict discharge from recorded water stage. Rating curves are utilized to combine continuously monitored water level data with with measured velocity data to produce a regression for estimating discharge.

Getting started with velocity measurements

Velocity should be recorded at various flow stages from low summer baseflow to high storm flow in order to achieve the best regression to predict discharge. Velocity can be measured in either of the two ways listed below:

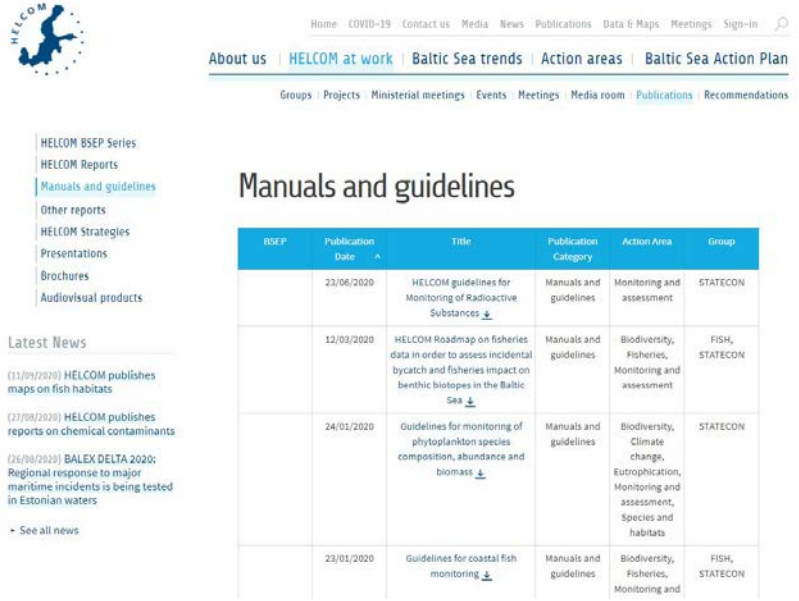
Flow meter method:

EPA
http://itepsrv1.itep.nau.edu/itep_course_downloads/Water_QAPP_TAMS_Center_ITEP/QA%20Project%20Plan/Mod5%20SOPs/Miscellaneous%20Field%20Procedures/Region%206%20Flow%20Measurement%20SOP%20updated%2001-31-03.pdf

USGS
https://pubs.usgs.gov/wsp/wsp2175/pdf/WSP2175_vol1a.pdf

Process:

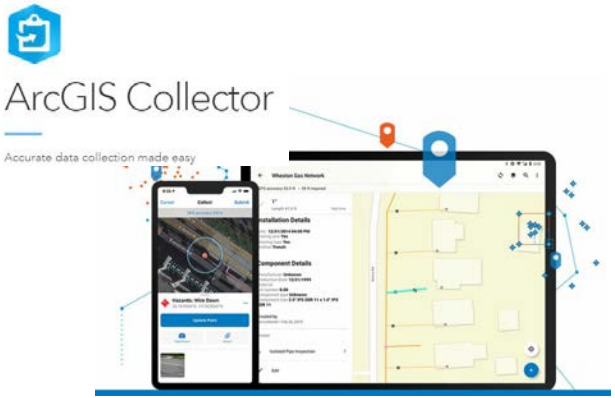
1. Locate reach of stream representative of stream reach
2. String across a tape measure and secure to both stream banks
3. Delineate what increment velocity will be measured. At least 20 increments are needed for good discharge estimation
4. Measure velocity and stream depth at set increments and when the slope of the cross-section changes drastically
5. Calculate discharge for each increment. Q=V*A

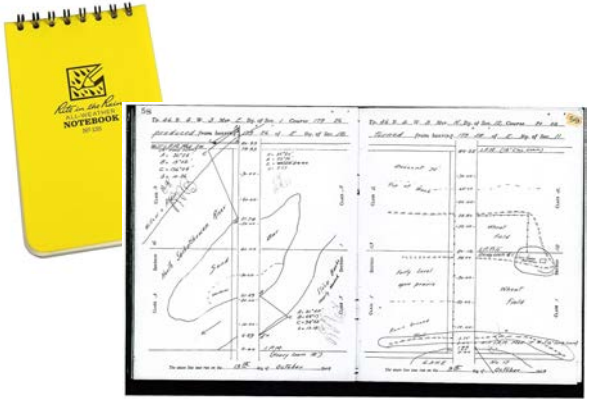


https://helcom.fi/helcom-at-work/publications/manuals-and-guidelines/

Field Notes and Log files

- Various methods such as collector apps, field books, field notes, site reports
- Can be manual or electronic

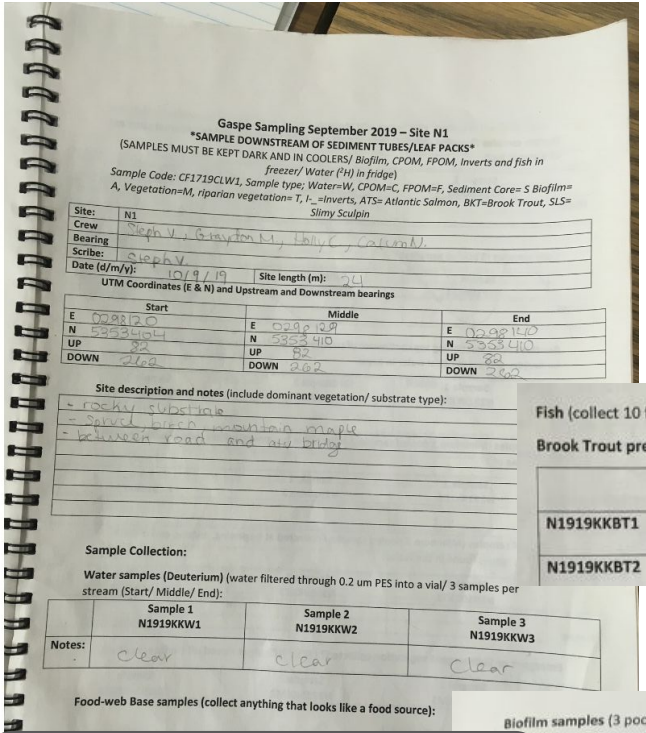




https://www.isc.ca/About/History/LandSurveys/FieldBooks/Pages/Elements.aspx

See also:
Great comparison of Electronic Lab Notebooks from Harvard Biomedical Data Management.
https://datamanagement.hms.harvard.edu/electronic-lab-notebooks
Pain (2019) How to Keep a Lab Notebook. Science. https://doi.org/10.1126/science.caredit.aaz3678

Field Notes and Log files - Examples and Resources



Gaspe Sampling September 2019 – Site N1

SAMPLE DOWNSTREAM OF SEDIMENT TUBES/LEAF PACKS

(SAMPLES MUST BE KEPT DARK AND IN COOLERS/ Biofilm, CPOM, FPOM, Inverts and fish in freezer/ Water (4°) in fridge)

Sample Code: CF1719CLW1, Sample type; Water=W, CPOM=C, FPOM=F, Sediment Core= S Biofilm= A, Vegetation=M, riparian vegetation= T, I-_=inverts, ATS= Atlantic Salmon, BKT=Brook Trout, SLS= Slimy Sculpin

Site:	N1
Crew	Steph V., Grayton M., Holly C., Catlin N.
Bearing	
Scribe:	Steph V.
Date (d/m/y):	10/9/19
Site length (m):	24

UTM Coordinates (E & N) and Upstream and Downstream bearings

	Start	Middle	End
E	0298120	0298129	0298120
N	5353404	5353410	5353410
UP	82	82	82
DOWN	262	262	262

Site description and notes (include dominant vegetation/ substrate type):

- rocky substrate
- spruce, birch, mountain maple
- between road and an bridge

Sample Collection:

Water samples (Deuterium) (water filtered through 0.2 um PES into a vial/ 3 samples per stream (Start/ Middle/ End):

	Sample 1 N1919KKW1	Sample 2 N1919KKW2	Sample 3 N1919KKW3
Notes:	clear	clear	clear

Food-web Base samples (collect anything that looks like a food source):

Fish (collect 10 fin clips):

Brook Trout present? (YES / NO)

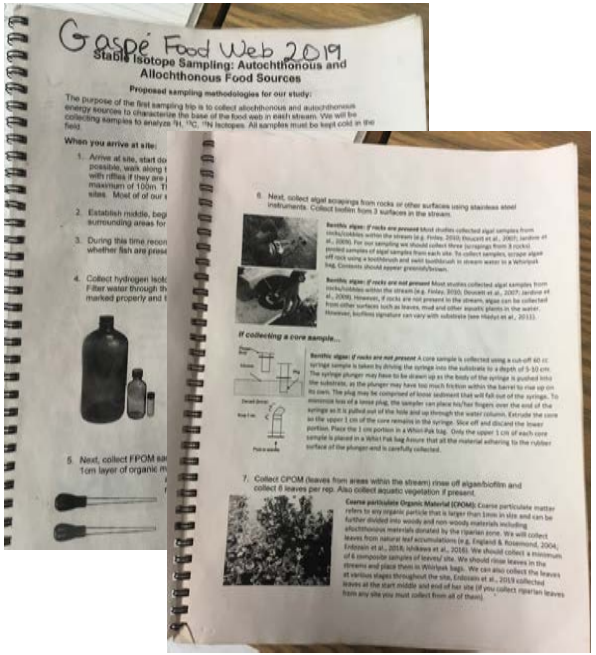
	Length (cm)	Weight (g)
N1919KKBT1	16.4	42.69
N1919KKBT2	6.8	2.24

Biofilm samples (3 pooled samples) Note surface you collect from (try and get minimal samples ...ter in bag):

Sample 1 N1919KKA1	Sample 2 N1919KKA2	Sample 3 N1919KKA3

Field Notes capture the who, what, where, why, and how of data collection

Standard Operating Procedures can be included for easy reference.



Gaspé Food Web 2019

Stable Isotope Sampling: Autochthonous and Allochthonous Food Sources

Proposed sampling methodologies for our study:

The purpose of the first sampling trip is to collect allochthonous and autochthonous energy sources to characterize the base of the food web in each stream. We will be collecting samples to analyze ^{2}H, ^{13}C, ^{15}N isotopes. All samples must be kept cold in the field.

When you arrive at site:

1. Arrive at site, start do... possible, walk along t... with riffles if they are ... maximum of 100m. T... sites. Most of of our ...
2. Establish middle, beg... surrounding areas for ...
3. During this time recon... whether fish are pres...
4. Collect hydrogen isot... Filter water through th... marked properly and t...
5. Next, collect FPOM sa... 1cm layer of organic m...
6. Next, collect algal scrapings from rocks or other surfaces using stainless steel instruments. Collect biofilm from 3 surfaces in the stream.

Benthic algae: if rocks are present Most studies collected algal samples from rocks/cobbles within the stream (e.g. Finlay, 2010; Doucett et al., 2007; Jardine et al., 2006). For our sampling we should collect three (scrapings from 3 rocks) pooled samples of algal samples from each site. To collect samples, scrape algae off rock using a toothbrush and swirl toothbrush in stream water in a Whirlpak bag. Contents should appear greenish/brown.

Benthic algae: if rocks are not present Most studies collected algal samples from rocks/cobbles within the stream (e.g. Finlay, 2010; Doucett et al., 2007; Jardine et al., 2006). However, if rocks are not present in the stream, algae can be collected from other surfaces such as leaves, mud and other aquatic plants in the water. However, biofilms signature can vary with substrate (see Hladyz et al., 2011).

If collecting a core sample...

Benthic algae: if rocks are not present A core sample is collected using a cut-off 60 cc syringe sample is taken by driving the syringe into the substrate to a depth of 5-10 cm. The syringe plunger may have to be drawn up as the body of the syringe is pushed into the substrate, as the plunger may have too much friction within the barrel to rise up on its own. The plug may be comprised of loose sediment that will fall out of the syringe. To minimize loss of a loose plug, the sampler can place his/her fingers over the end of the syringe as it is pulled out of the hole and up through the water column. Extrude the core so the upper 1 cm of the core remains in the syringe. Slice off and discard the lower portion. Place the 1 cm portion in a Whirl-Pak bag. Only the upper 1 cm of each core sample is placed in a Whirl Pak bag Assure that all the material adhering to the rubber surface of the plunger end is carefully collected.

7. Collect CPOM (leaves from areas within the stream) rinse off algae/biofilm and collect 6 leaves per rep. Also collect aquatic vegetation if present.

Course particulate Organic Material (CPOM): Coarse particulate matter refers to any organic particle that is larger than 1mm in size and can be further divided into woody and non-woody materials including allochthonous materials donated by the riparian zone. We will collect leaves from natural leaf accumulations (e.g. England & Rosemond, 2004; Evtimov et al., 2018; Ishikawa et al., 2016). We should collect a minimum of 6 composite samples of leaves/ site. We should rinse leaves in the streams and place them in Whirlpak bags. We can also collect the leaves at various stages throughout the site, Evtimov et al., 2019 collected leaves at the start middle and end of her site (if you collect riparian leaves from any site you must collect from all of them).

Field book printed on Rite in the Rain paper and bound for a cohesive record.

Shared with permission of The Kidd Lab (Dr. Karen Kidd)

File Organization

https://www.therussianstore.com/media/wysiwyg/Traditional_Russian_Matryoshka.jpg

- Increases findability and efficiency
- Prompts for critical data management elements (RawData, Licensing etc.)

Good ideas:

- Use a **ToFile** folder to collect items that don't have a place yet or if you are short on time
- Be aware of **maximum path length**
- **Avoid spaces and special characters** which programs and code may not interpret correctly
- For teams and labs, create an **empty file structure that can be shared** to enhance consistency or use a script to generate the structure

```
\ToFile
\ActiveProjects
    \Site1Project
     Site1Project_ReadMe
        \Code
        \Data
            \Tower1_2017
             Tower1_2017_ReadMe
                \Raw
                \Processed
                \Published
            \Tower1_2018
             Tower1_2018_ReadMe
                \Raw
                \Processed
                \Published
            \Third_Party_Data
                \LiDAR_2017
                 LiDAR_2017_ReadMe
                 LiDAR_2017_Data
        \Figures
        \Methods
        \Papers
        \Plans
        \Publications
        \ReferenceMaterials
        \Shared
            \Name_Date
            Name_Date_Request
            Name_Date_SharingAgreement
            \Name_Date_Data
\Finance
    \FormTemplates
    \Non_Project
    \Site1Project
\RefMaterials
\Templates
    \DocTemplates
        \Code
        \Metadata
        \Plans
        \SOP
    \EmptyPrjFolderStructure
```

https://osf.io/k83qx/wiki/A%20Good%20Start%20on%20Folder%20Structure/

File Organization - Examples and Resources

```
├── Code
│   ├── clean_raw_data.r
│   ├── analysis_1.r
│   └── analysis_2.r
├── Data
│   ├── Raw_data
│   │   ├── file_a.raw
│   │   └── file_b.raw
│   └── Processed_data
│       ├── file_a.csv
│       └── file_b.csv
├── Outputs
│   ├── Figures
│   └── Models
└── README.txt
```

```
├── Site_Name
│   ├── Observations
│   │   ├── 20190520_SiteCode_streamflow.csv
│   │   ├── 20190720_SiteCode_streamflow.csv
│   │   └── 20190920_SiteCode_streamflow.csv
│   └── Processed_data
│       └── SiteCode_contentDescription.csv
├── Site_Name
│   ├── Observations
│   │   ├── 20190520_SiteCode_streamflow.csv
│   │   ├── 20190720_SiteCode_streamflow.csv
│   │   └── 20190920_SiteCode_streamflow.csv
│   └── Processed_data
│       └── SiteCode_contentDescription.csv
├── Documentation
│   ├── Instrument_calibration.csv
│   ├── Data_collection_methods.txt
│   └── Quality_assurance.txt
└── README.txt
```

In terms of a filing system, we suggest starting with the following file structure somewhat like the following:

- Clubs (e.g., NREM GSO, SASA, Grebe)
- Courses (e.g., NREM507, SUSTAG509)
- FieldNotes
- Miscellaneous
- Personal
- Project
    - Analyses
        - AnalysisX
        - AnalysisY…
    - GrantDocumentation-Reports
    - Images
    - MeetingNotes
    - Presentations
    - Proposal
    - RawData
    - Readings
    - Thesis/Dissertation
        - Chapter 1
        - Chapter2…s

https://www.nrem.iastate.edu/landscape/lab-data-and-file-management-sops

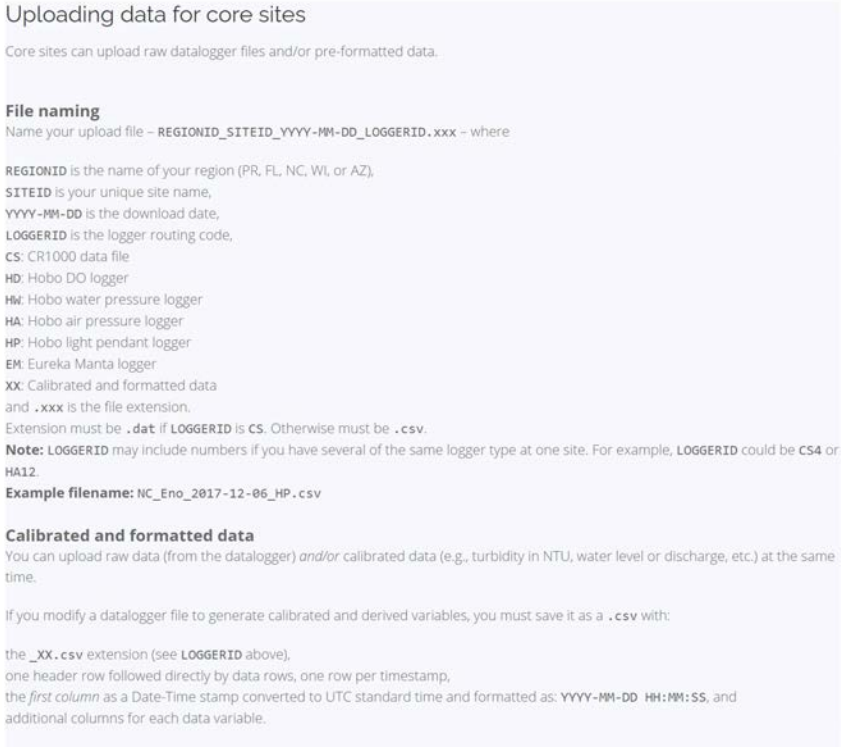
For further guidance: NIST Electronic File Organization Tips:
https://www.nist.gov/system/files/documents/pml/wmd/labmetrology/ElectronicFileOrganizationTips-2016-03.pdf

File Naming Guidelines

- Balance between concise and descriptive. May contain:
  - Project name, site name, type of data or analysis
  - Date (YYYYMMDD format)
  - Version number (v_01, v_02, etc.)
- Avoid using:
  - Spaces (use - or _ instead)
  - Special characters (~ ! @ # $ % ^ & * ( ) ` ; < > ? , [ ] { } ' " and |)
  - Potentially sensitive or restricted information
- Lead numbers with a zero (e.g, 001, 002 … 010, 011 … 100, 101)
- Try to make filenames unique (directory structure may change over time)
- Document your naming conventions to help with consistency

Resources: UBC File Naming Guidelines, Stanford Best Practices for File Naming

File Naming - Example



Uploading data for core sites

Core sites can upload raw datalogger files and/or pre-formatted data.

File naming

Name your upload file – `REGIONID_SITEID_YYYY-MM-DD_LOGGERID.xxx` – where

`REGIONID` is the name of your region (PR, FL, NC, WI, or AZ),
`SITEID` is your unique site name,
`YYYY-MM-DD` is the download date,
`LOGGERID` is the logger routing code,
`CS`: CR1000 data file
`HD`: Hobo DO logger
`HW`: Hobo water pressure logger
`HA`: Hobo air pressure logger
`HP`: Hobo light pendant logger
`EM`: Eureka Manta logger
`XX`: Calibrated and formatted data
and `.xxx` is the file extension.
Extension must be `.dat` if `LOGGERID` is `CS`. Otherwise must be `.csv`.
Note: `LOGGERID` may include numbers if you have several of the same logger type at one site. For example, `LOGGERID` could be `CS4` or `HA12`.
Example filename: `NC_Eno_2017-12-06_HP.csv`

Calibrated and formatted data

You can upload raw data (from the datalogger) *and/or* calibrated data (e.g., turbidity in NTU, water level or discharge, etc.) at the same time.

If you modify a datalogger file to generate calibrated and derived variables, you must save it as a `.csv` with:

the `_XX.csv` extension (see `LOGGERID` above),
one header row followed directly by data rows, one row per timestamp,
the *first column* as a Date-Time stamp converted to UTC standard time and formatted as: `YYYY-MM-DD HH:MM:SS`, and
additional columns for each data variable.

E.g., http://pulseofstreams.weebly.com/uploading_data.html

Scripts and syntax files

- If possible use a scripted language to process and analyze your data
- Promote transparency and reproducibility
- Don’t alter your raw data directly
  - Input raw data -> generate new file as output
- Include comments
  - These make each line of code human readable and the function is clear
- Keep track of any external dependencies
  - List name and version of packages or libraries you use

Reference (and more great advice): Borer ET, Seabloom EW, Jones MB, and Schildhauer M (2009) Some Simple Guidelines for Effective Data Management. The Bulletin of the Ecological Society of America. https://doi.org/10.1890/0012-9623-90.2.205.

Analysis Scripts - Example

Digital Elevation Map (DEM) and derived products, slope and aspect. Runtime around 5-10 minutes

After running 'src_borders.R' we now have a set of polygons ('snrc_std.shp', loaded as `snrc.sf`, below) with which to split the data into blocks (*ie* mapsheets). The path of this shapefile is stored in the `cfg.borders` metadata list, which can be loaded from the file 'borders.rds' located in the `data.dir` directory.

```
# download again to rewrite existing files?
force.download = FALSE

# load the helper functions
library(here)
source(here('utility_functions.R'))

# load the borders info and shapefiles for mapsheets
cfg.borders = readRDS(here('data', 'borders.rds'))
snrc.sf = sf::st_read(cfg.borders$out$fname$shp['snrc'])
snrc.codes = cfg.borders$out$code
```

As with the 'borders' collection, we will fill in the metadata as we go, and save a copy to disk at the end

```
# create the source data directory, metadata list and its path on disk
collection = 'dem'
cfg = MPB_metadata(collection)
```

Data in FRDR at Koch DC, Lewis MA (2020) Raster datasets relevant to mountain pine beetle outbreak ecology in the province of British Columbia. Federated Research Data Repository. https://doi.org/10.20383/101.0283 with code used to derive the files in GitHub at https://github.com/deankoch/rasterbc_src and code to import a subset of data into R for modelling at https://github.com/deankoch/rasterbc.

ReadMe files

- General information
  - Dataset name, project summary, contact information, date range, funding sources,
- Data and file overview
  - File name, description of content, format, notes
- Sharing and access information
  - Limits of use, how to access data, citation
- Methodological information
  - Instrumentation, links to protocols and manuals, QA/QC, processing details
- Data-specific information
  - Variable names, units, missing data codes
- Details of anomalies necessary for interpreting the data and its suitability for use

Further guidance is available in UBC's Quick Guide: Creating a README for your dataset and Cornell University's Guide to writing "readme" style metadata (template available for download).

ReadMe files - Examples and Resources

README.md

bpwtpR: Buffalo Pound Water Treatment Plant (BPWTP) data integration & reporting

Summary

This project was completed as part of the MITACs accelerate program (IT17066) and the CFREF Global Water Futures Initiative FORMBLOOM (awarded to Venkiteswaran and Baulch).

Highly Qualified Personnel

- Megan L. Larsen, Wilfrid Laurier University, mlarsen@wlu.ca, ORCID 0000-0002-4155-1293
- Anthony A.P. Baron, University of Saskatchewan

Principle Investigators

- Jason Venkiteswaran, Wilfrid Laurier University, jvenkiteswaran@wlu.ca, ORCID 0000-0002-6574-7071

Data sources

- Provide links to any data used from external providers and/or submodules and their associated management source (e.g. IISD-ELA, OMOECC).

Funding sources

- MITACs Accelerate Program (IT-17066)
- FORMBLOOM

Keywords

List keywords separated by commas

Additional information and support

- Sensitive Data Flag - Human Participants: NO
- Sensitive Data Flag - Indigenous Partnerships: NO
- Sensitive Data Flag - Government Partnerships: NO
- Sensitive Data Flag - Industry Partnerships: YES
- Access Restrictions

Variables

Potential variables include:

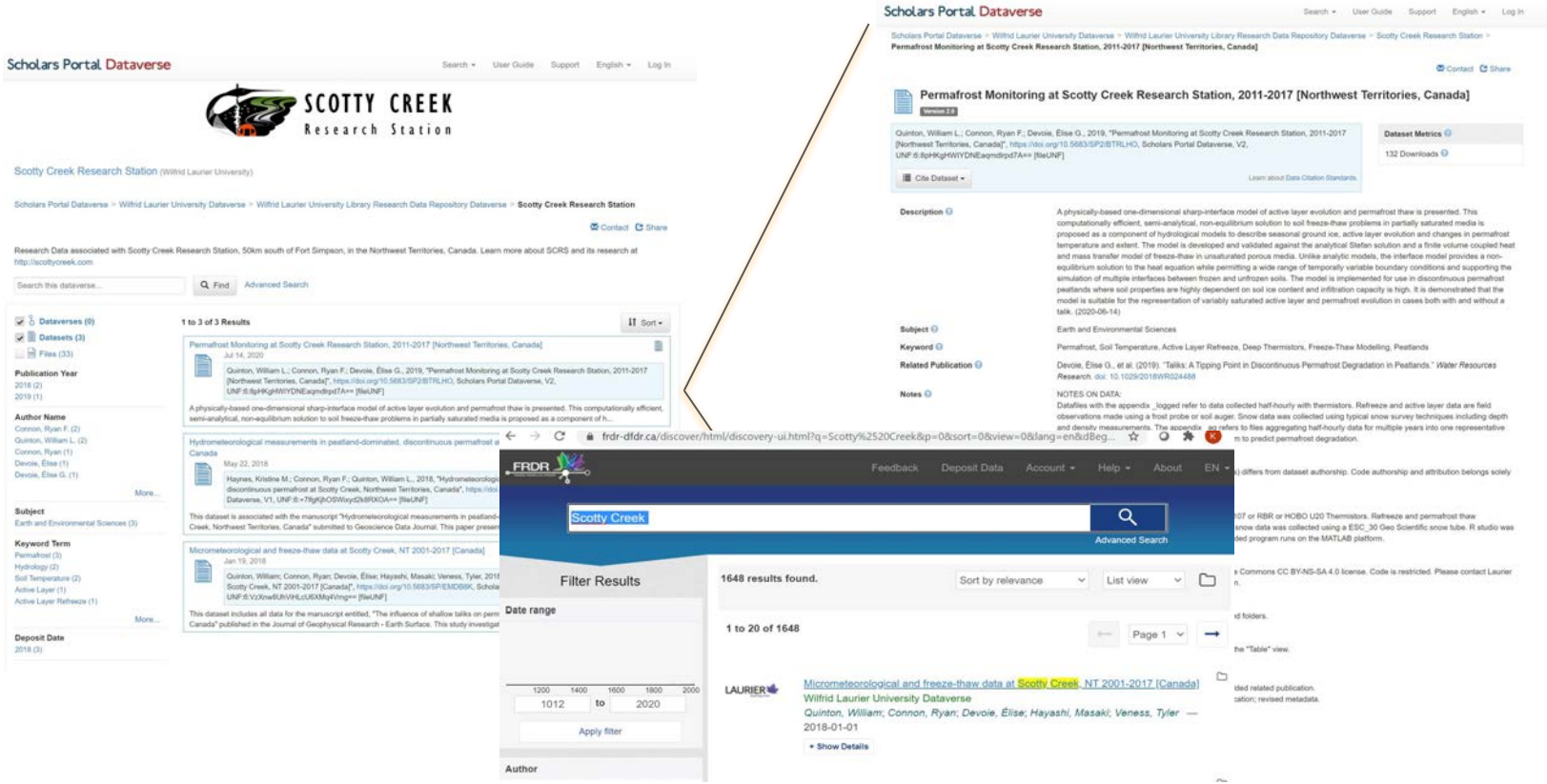
Date-Time (UTC)
DO (mg/L)
Saturation DO (mg/L)
Water Temperature (°C)
Water Pressure (kPa)
Air Temperature (°C)
Air Pressure (kPa)
Depth (m)
Discharge (m3/s)
Velocity (m/s)
Light, PAR (µmol/m2/s)
Light, lux
Specific Conductivity (mS/cm or µS/cm)
pH
fDOM (frac)
fDOM (mV from sensor)
Turbidity (NTU)
Turbidity (mV from sensor)
Nitrate (mg/L)
CO2 (ppm)

*GitHub example courtesy of Dr. Jason Venkiteswaran (Wilfrid Laurier University)
https://jivenky.github.io/rules/ ; https://github.com/biogeochem/musical-funicular

Metadata Standards - General Purpose

Element	Notes	E.g., (https://doi.org/10.20383/101.0193)
Title	Be concise and descriptive. Choose something appropriate for use in a data citation.	Unmanned aerial vehicle structure from motion and lidar data for sub-canopy snow depth mapping
Author	Dataset creators. Be prepared to provide author affiliations and ORCIDs.	Harder, Phillip. University of Saskatchewan. https://orcid.org/0000-0003-2144-2767
Description	It may help to think of the dataset description as you would an article abstract. This is a high-level summary that addresses the nature and scope of the dataset.	Unmanned Aerial Vehicles (UAV) have had recent widespread application to capture high resolution information on snow processes and the data herein was collected to address the sub-canopy snow depth challenge...
Subject	Keywords that describe the dataset.	Unmanned aerial vehicle; lidar; snow depth, point cloud, digital surface model
Spatial coverage	Place where data was gathered. Place name, geographic coordinates, and/or bounding box.	Fortress Mountain Snow Laboratory (Fortress), Kananaskis, AB; 50.833 -115.220
Temporal coverage	Data collection period(s).	Start: 2018-09-07 / End: 2019-04-25
Related identifier	Link to associated publications, code, protocols, methods, and other information that gives context to the data.	https://doi.org/10.5194/tc-14-1919-2020
Rights	Terms of use. Note: the terms you can assign might be limited if your data are derived from external sources.	Creative Commons Attribution 4.0 International (CC BY 4.0)
Funder	Granting agency or funder(s) that supported your research.	Natural Sciences and Engineering Research Council of Canada (NSERC)

General Purpose Elements Aid Discovery



Metadata Standards - Disciplinary

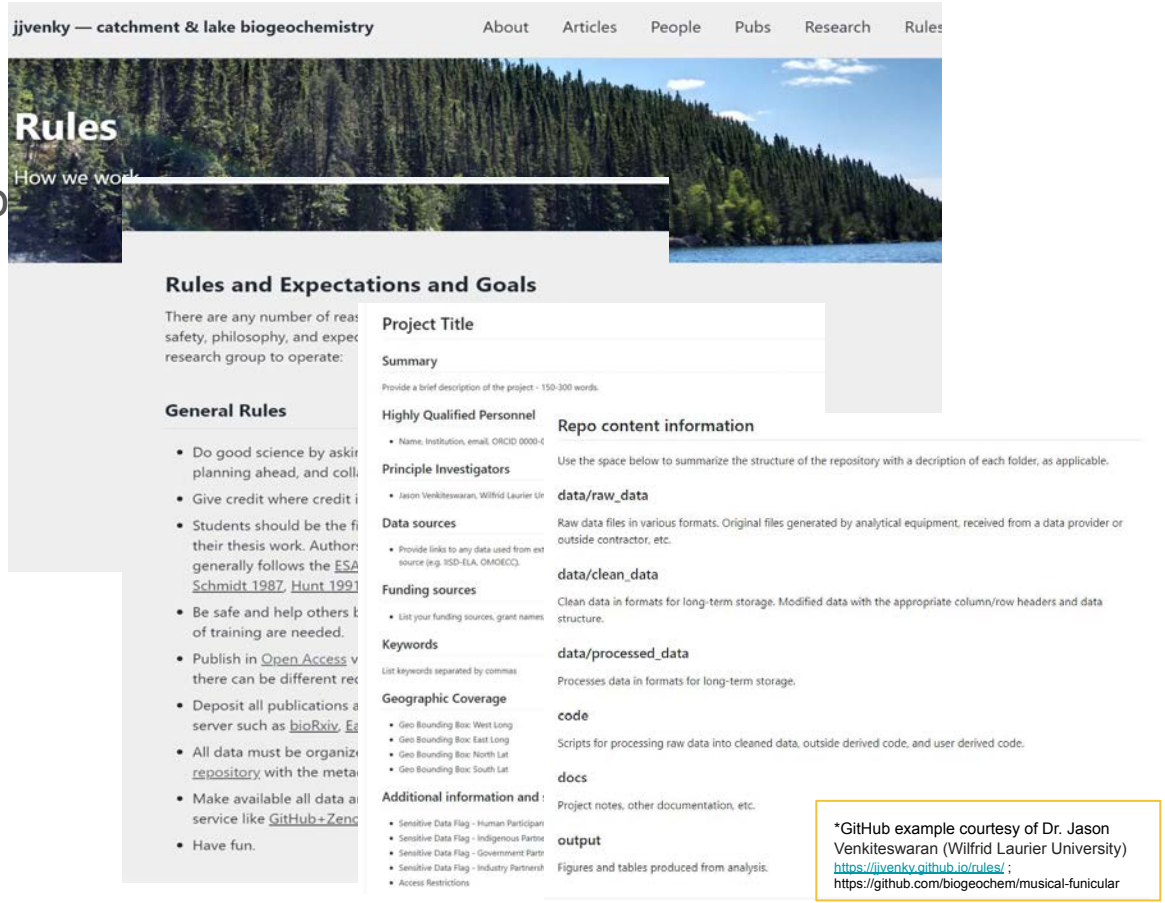
Standard	Use for	Tools
ISO 19115	Geographic information in many disciplines (e.g., Glaciology, Hydrogeology, Oceanography, Climatology)	mdEditor https://www.mdeditor.org/ (also supports FGDC)
Ecological Metadata Language (EML)	Ecology and environmental sciences	Morpho https://knb.ecoinformatics.org/tools/morpho
Data Documentation Initiative (DDI)	Social, behavioral, and economic sciences	Many options available at https://ddialliance.org/resources/tools
Digital Imaging and Communications in Medicine (DICOM)	Medical imaging information and related data	Attribute confidentialtiy profiles http://dicom.nema.org/medical/dicom/current/output/html/part15.html#chapter_E
Investigation/Study/Assay tab-delimited (ISA-TAB)	Omics-based experiments	Many options available at https://isa-tools.org/software-suite.html
Crystallographic Information Framework (CIF)	Crystallographic and related structured data	Many options available at https://www.iucr.org/resources/cif/software
Flexible Image Transport System (FITS)	Astronomical data (images and multi-dimensional data)	Many options available at https://fits.gsfc.nasa.gov/fits_utility.html

See Digital Curation Centre <https://www.dcc.ac.uk/guidance/standards/metadata> or RDA Metadata Standards Working Group <https://rd-alliance.github.io/metadata-directory/> for more standards and tools. See also https://fairsharing.org/ for standards and policies.

Collaborative platforms

Collaborative tools provide a platform for sharing procedures, data, code and other project related information. The helps to create a common understanding and identify authoritative versions.

- Google (forms and documents)
- Open Science Framework
- GitHub
- CEDAR Workbench



*GitHub example courtesy of Dr. Jason Venkiteswaran (Wilfrid Laurier University) https://jjvenky.github.io/rules/ ; https://github.com/biogeochem/musical-funicular

Collaborative platforms

Product	Notes	Storage limit (free)	Getting Started
Google Drive	• Set up shared Drive for your research project • Develop shared documentation, and adjust settings to allow offline editing of documents • Use forms to standardize metadata acquisition	15 GB	https://www.google.ca/forms/about/
OSF	• Option to set global storage location to Montréal so your documents are hosted on Canadian servers • Keep your full project private to your team, or choose to share individual components of the project more widely • Use storage add-ons to connect your project to Drive, GitHub and other external services	Unlimited	OSF guidance at https://help.osf.io/hc/en-us OSF instructional videos https://osf.io/3yw4y/wiki/home/ UBC Research Commons upcoming OSF workshop, Oct 5 (open to public).
GitHub	• Can be used for more than software - also use for creating and maintaining documentation • Built for collaboration and version control • Free basic accounts for teams (unlimited public/private repositories, unlimited collaborators) https://github.com/pricing	500 Mb	Jason Brodeur's upcoming webinar for the GWF/Portage webinar series, Oct 6. 1:00 PM EDT. Git cheat sheets https://training.github.com/ (Eng & Fra) and GitHub's beginner guide Hello World Library Carpentry Introduction to Git https://librarycarpentry.org/lc-git/
CEDAR Workbench	• Create rich web-based metadata-acquisition forms • Pre-configure forms, and pre-set your team's defaults • Templates can be versioned • APIs allow you to import existing metadata or export metadata • Developed for biomedical sciences, but you can create your own templates	N/A	CEDAR user guide https://metadatacenter.github.io/cedar-manual/

In Summary:

- Have a data management plan that addresses documentation and metadata
- Budget time
  - Add a buffer so metadata isn't edged out by other competing priorities!
- Budget resources
  - Metadata creation takes people and time. If you are writing a new grant proposal, consider budgeting staff time for metadata and other data management tasks
  - OpenAIRE RDM Costing Tool https://www.openaire.eu/how-to-comply-to-h2020-mandates-rdm-costs
- Integrate into workflow
- Collaborative effort

Checklist for success ...

- ❏ Use SOPs to help you manage things consistently.
- ❏ Follow best practice guidance to name and structure files.
- ❏ Create a README, codebook, and other documentation necessary to understand your data and interpret it correctly.
  - ❏ Document data collection instruments, methods, quality control measures, and specialized software needed to view or manipulate data.
  - ❏ Describe the contents of your directories and/or files.
  - ❏ Provide attribution to any external data sources.
- ❏ For tabular data, define all variables and allowable values including null values. Include units of measure where appropriate.
- ❏ Scripts and code should be well commented and dependencies accounted for.
- ❏ Provide links to associated publications, code in external repositories, and any other information that provides more context to your data.
- ❏ Select a license that respects the constraints of any data you may have reused (for help choosing a Creative Commons license, see https://chooser-beta.creativecommons.org/)

Resources list

- Data Management Planning
  - DMP Assistant, bilingual tool for preparing data management plans (DMPs) https://assistant.portagenetwork.ca/
  - OpenAire. How to identify and assess Research Data Management (RDM) costs. https://www.openaire.eu/how-to-comply-to-h2020-mandates-rdm-costs
  - Borer ET, Seabloom EW, Jones MB, and Schildhauer M (2009) Some Simple Guidelines for Effective Data Management. The Bulletin of the Ecological Society of America. https://doi.org/10.1890/0012-9623-90.2.205
  - FAIR Principles: https://www.go-fair.org/fair-principles/
  - CARE Principles: https://www.gida-global.org/care
  - DataONE Data Management Skill Building Hub https://dataoneorg.github.io/Education/

- Standard Operating Procedures - Documenting expectations and best practices
  - DataONE, Best Practices https://dataoneorg.github.io/Education/bestpractices/ for ideas on elements of RDM that may be useful to document
  - StreamPulse examples: Click “Show SOPs“ button on https://data.streampulse.org/ for SOP examples.

Resources list, continued

- Lab Notebooks / Field Notes
  - Harvard University, Harvard Biomedical Data Management (2020, Feb 27) Electronic Lab Notebooks. https://datamanagement.hms.harvard.edu/electronic-lab-notebooks
  - Kwok R (2018) How to pick an electronic laboratory notebook. Nature 560: 269-270. https://doi.org/10.1038/d41586-018-05895-3
  - Pain E (2019, Sep 3) How to Keep a Lab Notebook. Science. https://doi.org/10.1126/science.caredit.aaz3678
  - University of Southern California Libraries (2020, Aug 6) Research Guides, Organizing Your Social Sciences Research Paper, Writing Field Notes. https://libguides.usc.edu/writingguide/fieldnotes
- File Organization and File Naming
  - Smithsonian Libraries. Best Practices for Filenaming, Organizing, and Working with Data https://library.si.edu/research/best-practices-filenaming-organizing-and-working-data
  - The National Institute of Standards and Technology. Electronic File Organization Tips. https://www.nist.gov/system/files/documents/pml/wmd/labmetrology/ElectronicFileOrganizationTips-2016-03.pdf
  - Stanford Libraries (Research Support) Best Practices for File Naming. https://library.stanford.edu/research/data-management-services/data-best-practices/best-practices-file-naming
  - University of British Columbia. File Naming Guidelines. https://researchdata.library.ubc.ca/files/2019/01/FileName_Guidelines_20140410_v03.pdf

Resources list, continues

- READMEs
  - Brigham D (2020, Apr 1) Quick Guide: Creating a README for your dataset. https://doi.org/10.5281/zenodo.4058971
    - En français https://doi.org/10.5281/zenodo.4058961
  - Cornell University. Research Data Management Service Group. Guide to writing "readme" style metadata. https://data.research.cornell.edu/content/readme (Template available for download).
- Metadata Standards
  - List of disciplinary metadata standards, brief descriptions, and tools
    - The Research Data Alliance Metadata Standards Directory Working Group. https://rd-alliance.github.io/metadata-directory/
    - Digital Curation Centre. Disciplinary Metadata. https://www.dcc.ac.uk/guidance/standards/metadata
  - Search for standards and policies by domain or subject.
    - FAIRSharing. https://fairsharing.org/standards/
- Licensing
  - For Data. Choosing a Creative Commons License. https://chooser-beta.creativecommons.org/
  - Choose an Open Source License. https://choosealicense.com/
  - For Software. Open Source Initiative. Licenses and Standards. https://opensource.org/licenses