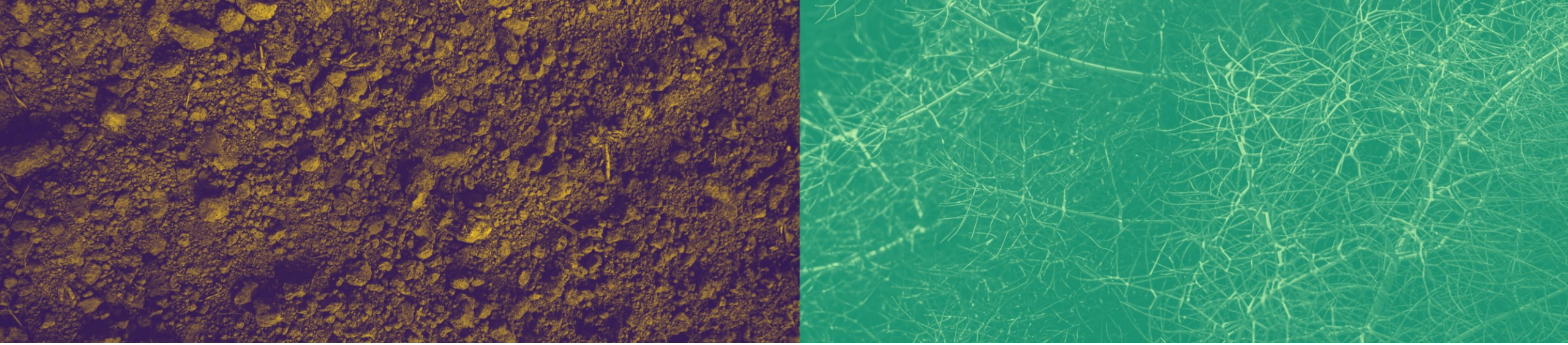


Before You Dig: **Finding and Reusing Datasets**

Isaac Pratt, PhD and Danica Evering, MA

March 15, 2023

Research Data Management Services

McMaster University

McMaster University | Library

McMaster University is located on the traditional Territories of the Mississauga and Haudenosaunee Nations, and within the lands protected by the “Dish With One Spoon” wampum agreement.

Mhsheikholeslami, “Eramosa Karst Conservation Area- Nexus Cave- Stoney Creek- Hamilton-Ontario,” 16 June 2019, Wikimedia Commons - https://commons.wikimedia.org/wiki/File:Eramosa_Karst_Conservation_Area-_Nexus_Cave-_Hamilton-Ontario-2.jpg

Code of Conduct

The Sherman Centre and the McMaster University Library are committed to fostering a supportive and inclusive environment for its presenters and participants.

As a participant in this session, you agree to support and help cultivate an experience that is collaborative, respectful, and inclusive, as well as free of harassment, discrimination, and oppression. We reserve the right to remove participants who exhibit harassing, malicious, or persistently disruptive behaviour.

Please refer to our code of conduct webpage for more information:
scds.ca/events/code-of-conduct/

Session Recording and Privacy

This session is being recorded with the intention of being shared publicly via the web for future audiences. In respect of your privacy, participant lists will not be shared outside of this session, nor will question or chat transcripts.

Questions asked via the chat box will be read by the facilitator without identifying you. Note that you may be identifiable when asking a question during the session in an audio or visual format.

Certificate Program

The Sherman Centre offers a Certificate of Completion that rewards synchronous participation at 7 workshops. We also offer concentrations in Data Analysis and Visualization, Digital Scholarship, and Research Data Management.

Learn more about the Certificate Program: https://scds.ca/certificate-program
If you would like to be considered for the certificate, verify your participation in this form: https://u.mcmaster.ca/verification

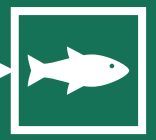
At an unspecified point during the workshop, a code will be read aloud. This is the answer to the third question of the form.

Outline


Why look before you dig? (Limitations + Possibilities)


Secondary Data Sources

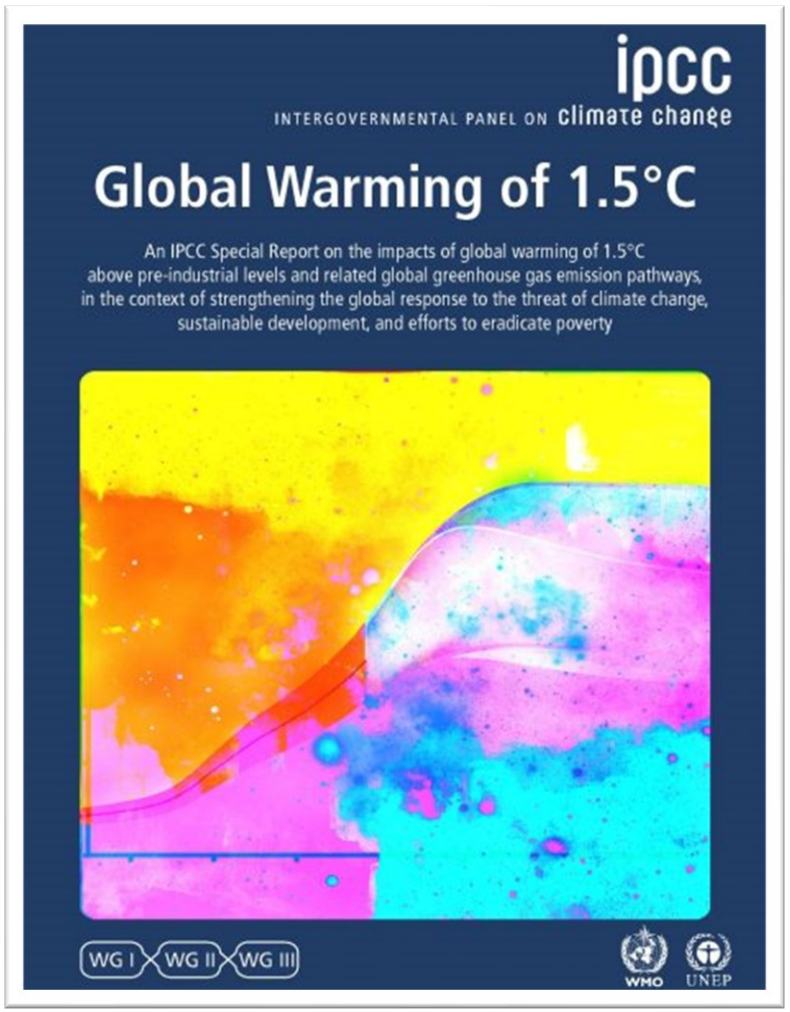

Finding Datasets


Case Studies


Data Access


Reusing Data (Unpacking a Dataset, Data Quality Checking)

Science built on the power of open data



Research – Reusing Data

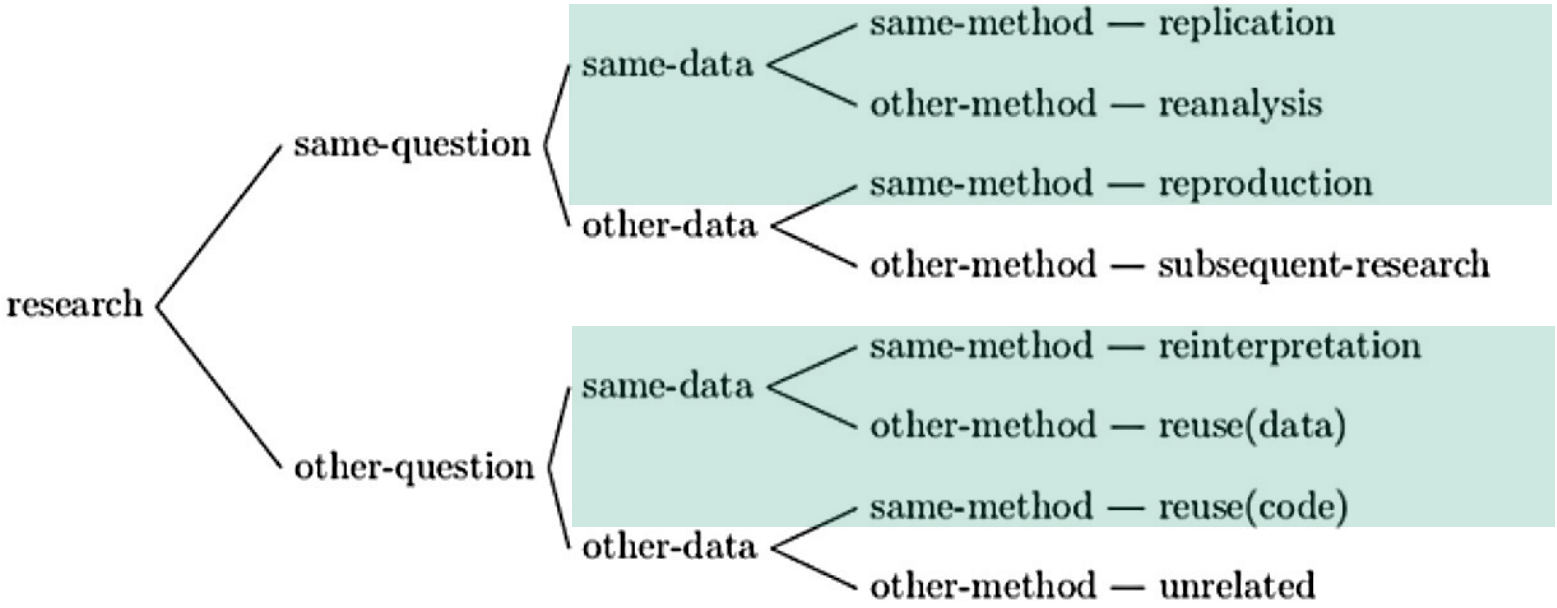


Image from van de Sandt, S., Dallmeier-Tiessen, S., Lavasa, A. and Petras, V., 2019. The Definition of Reuse. *Data Science Journal*, 18(1), p.22. DOI: http://doi.org/10.5334/dsj-2019-022.

Before You Dig:

- **Save Time + Funds**: An existing dataset means you don't spend resources collecting.
- **Larger Datasets**: data may come from many sources, more than is possible to aggregate as an individual.
- **Avoid Duplication:** Ensure you're contributing new research to the field.
- **Combine Datasets:** Add new data to an existing dataset to produce a broader dataset.
- **Reduce "Over-Research"**: Minimize impact on communities made vulnerable by research.

Steven Damron, "Warning! Buried Communication Cable Sign" 17 December, 2008, Wikimedia Commons - https://commons.wikimedia.org/wiki/File:Warning!_buried_communication_cable_sign_%283120060005%29.jpg

Reusing Data – Key Features

- Conducting secondary analysis
- Undertaking follow-up research
- Conducting research reviews
- Scrutinizing findings
- Using data for teaching and learning

Veerle Van den Eynden, "The Research Data Lifecycle," in *Managing and Sharing Research Data: A Guide to Good Practice (2nd Edition)*, ed. Louise Corti, Veerle Van den Eynden, Libby Bishop, and Matthew Woolard (London: SAGE Publications, 2020), 35.

Photo by Killari Hotaru on Unsplash

Reusing Data – **Potential Limitations and Solutions**

Limitations	*Solution*
Lack of available data.	*Increasing practices to make data accessible.*
Data may not meet your exact needs/ research question.	*Look at metadata – "data about data" to determine if this dataset is a good fit.*
Time to understand unfamiliar data or lack of documentation.	*Budget time to locate data sources, check variables and codes, read transcripts.*
Methodological training for secondary analysis.	*Increasing use of secondary data as a subject in methods courses, including reusing data!*
Concerns about ethical reuse.	*Ensuring consent is in place, data is de-identified or anonymized, restricted access.*

Louise Corti, Maureen Haaker, and Veerle Van den Eynden, "Making Use of Other People's Data: Opportunities and Limitations," in *Managing and Sharing Research Data: A Guide to Good Practice (2nd Edition)*, ed. Louise Corti, Veerle Van den Eynden, Libby Bishop, and Matthew Woolard (London: SAGE Publications, 2020), 35.

Photo by Pietro Jeng on Unsplash.

What kinds of **data** are out there?

- **Primary Data**: Data gathered by you or your team. May be surveys, interviews, experiment results, etc.
- **Secondary Data**: Data already collected by someone else which can be analyzed, combined, and added to. Secondary data may be:
  - **Administrative Data:** Government data on Healthcare, Statistics, Environment, Labour, Housing, Education, Economics
  - **Research Data**: Generated by other researchers and may be shared in data repositories.
  - **Other Open Data**: Non-profits, organizations, private sector

Government + Administrative Data

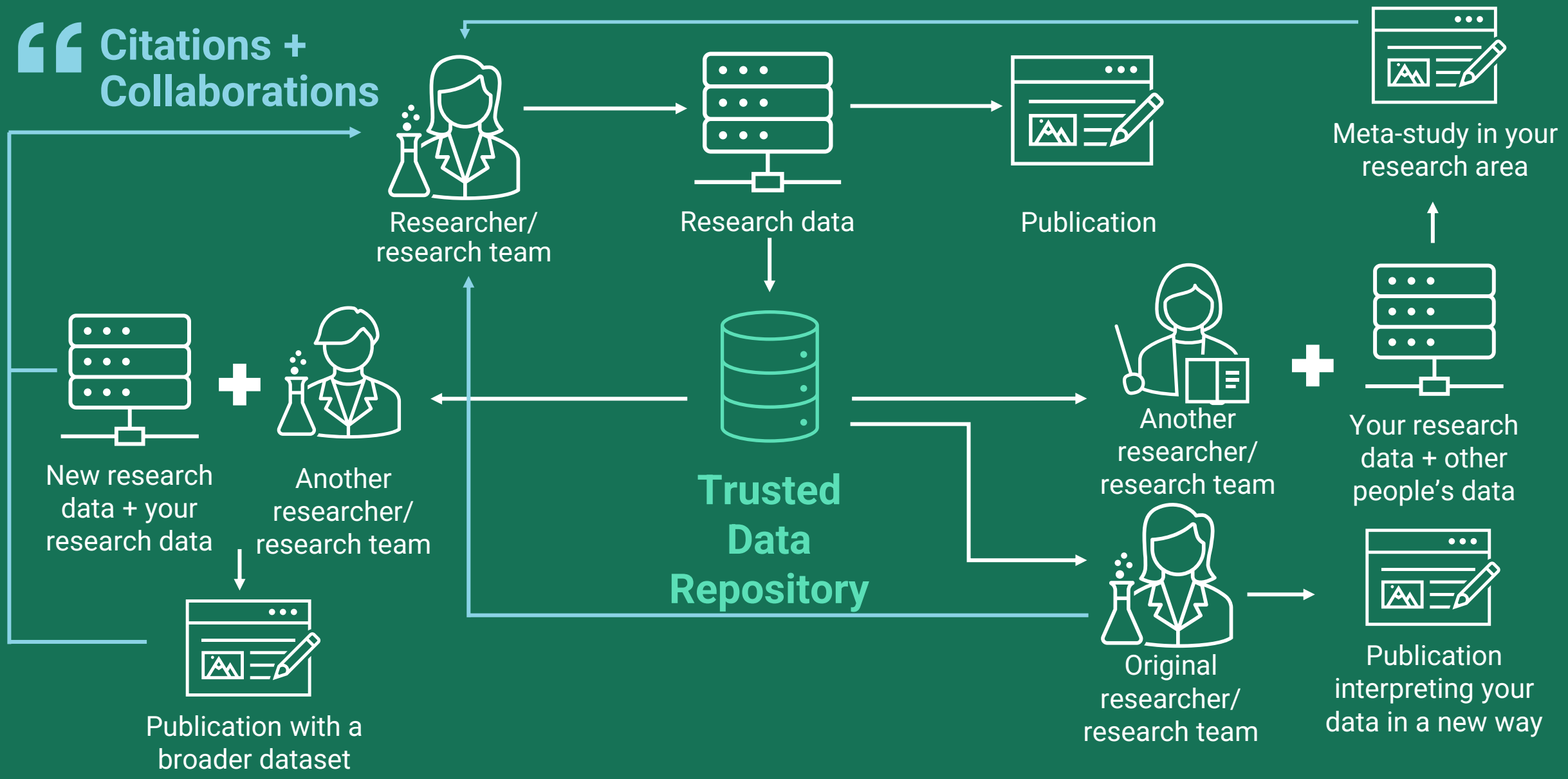
Healthcare: Provincial Governments, Cancer Care Ontario, CIHI, Public Health Ontario Laboratories, etc.

Education: number of students, achievement and grades.

Municipal Data: housing, arenas, cemeteries, hospitals, lakes, museums, bikeways, waste management, transit

Census: population, demographics, immigration, net revenue, marital status, etc.

Secondary Data – **Open Research Data**



Secondary Data – Other Open Data

Data Is Plural — 2022.09.21 edition

JS Jeremy Singer-Vine <data-is-plural@buttondown.email>
To Danica Evering

Reply | Reply All | Forward

Unsubscribe

Biodiversity trends. Maria Dornelas et al.'s BioTIME project has collected and standardized data from hundreds of studies examining ecolo time. You can browse and search the studies by year, taxa, species, and biome. You can also downlo full dataset, which provides information about each study (biome, start/end years, number of species tallied, and much more) ar sample collected (date, location, species, abundance, and biomass). **As seen in**: "Economic Production and Biodiversity in the Un States," by Yuanning Liang et al.

academia

university-based open access research database

Probabilistic predictions. Metaculus is a forecasting platform whose community has registered more than 1 million predictions on quest a major nuclear power plant in Germany be operational on June 1, 2023?" The website's API provides data on questions posed, user rankings, and other aspects of the platform. For each question, you can see its phrasing, date posed, creator, prediction type, the distribution of predictions, and more. **Related**: Zoltar, a forecast archive assembled by Nicholas G. Reich et al. **Previously**: FiveThirtyEight's assessment of its own predictions (DIP 2019.04.10).

industry

forecasting platform

Work Endowment for the Arts regularly produces statistical profiles of the arts in the United States. The latest, "Arti tional and State Estimates for 2015-2019," is tabulated from the Census Bureau's American Community Survey. It provides employment and earning estimates by artistic occupation and demographic. Additional tabulations, including for the country's 25 largest metro areas, are available through the National Archive of Data on Arts and Culture. [h/t Gary Price]

journalism

national funder

Atari emails. A couple of decades ago, Jed Margolin posted a cache of electronic mail messages from his time as a video game hardware engin Games, a successor company). In 2017, with Margolin's permission, Vikram Oberoi scraped the 4,000+ emails and built atariemailarchive.org, which groups the messages into threads, categories, and a list of favorites. The project also includes a database file containing each message's sender, recipients, timestamp, subject, body, and Oberoi's thread grouping. **Related**: "How I made atariemailarchive.org."

non-profits

independent enthusiastic person

Exploring existing data sources.

DataCite Search

Works People Repositories Members Support Danica Evering

rural family medicine Search

5,874 Works

Data from: What influence do courses at medical school and personal experience have on interest in practicing family medicine? – results of a student survey in Hessia

Antonia Bien, Gisela Ravens-Taeuber, Maria-Christina Stefanescu, Ferdinand M. Gerlach & Corina Güthlin
Dataset published via Dryad
Aim: Against the background of an impending shortage of family practitioners, it is important to investigate the factors influencing the choice to become one. The aim of this study was to identify factors that encourage medical students to choose to practice family medicine. Method: Using a questionnaire, students in the fourth and fifth years of their studies in the Federal State of Hesse were asked about the factors that had influenced their choice of medical...

1 citation 64 views 7 downloads

https://doi.org/10.5061/dryad.74tk6cr Cite Add to ORCID record

Assessment of antenatal care utilization and c and urban areas of Kathmandu District of Nepal

Prakash Prasad Shah
Dataset published via Chulalongkorn University
Background: Antenat components of mate Region, MMR is still higher in Nepal therefore SLTHP targets to increase in the percentage of pregnant women attending a minimum of four ANC visit to 80%. This study aimed to assess the antenatal care utilization and client satisfaction...

Registration Year

2023	482
2022	1,420
2021	1,137
2020	1,263
2019	925
2018	445
2017	3
2016	181
2015	4
2014	11

Resource Types

Dataset	5,874

Wisconsin-Madison

Finding Data – **General Search**

DataCite: https://search.datacite.org/

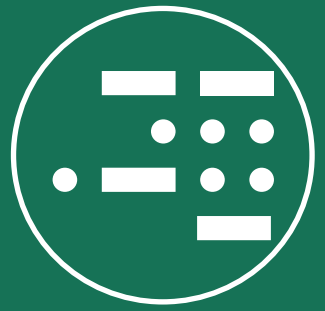
Google Dataset: https://datasetsearch.research.google.com/

Finding Data: **Open Data Repositories**

In addition to sharing data, open data repositories are also a great place to find data. There are thousands of repository options

Domain Specific Repositories
Focus on certain types of data such as genomic information or astronomical information.

General Repositories
Accept broader types of research data. *ex. McMaster Dataverse (part of Borealis) and Canada's Federated Research Data Repository (FRDR).*



Code Repositories
There are also software repositories like GitHub.

Repository Finder
This tool by DataCite can help you find an appropriate repository to deposit your research data: https://repositoryfinder.datacite.org

McMaster University | Research & High Performance Computing

McMaster University | Library

Finding Data - **Government + Administrative Data**

McMaster Library Data Services for DLI
https://library.mcmaster.ca/services/data-services

Statistics Canada Microdata through McMaster RDC https://rdc.mcmaster.ca/

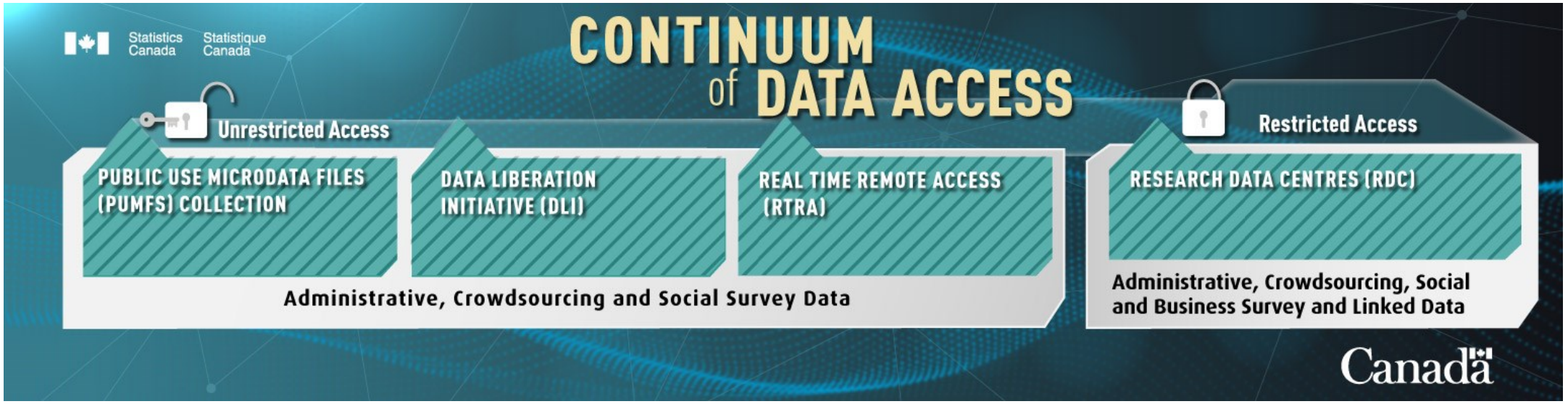


Image by Technicians Make it Happen, *This is Engineering* on Flickr, https://www.flickr.com/photos/thisisengineering/48315633117

Statistics Canada Research Data Centre at McMaster (RDC)

- Part of the Canadian Research Data Centre Network (CRDCN)
- Apply through Microdata Access Portal - https://www.statcan.gc.ca/en/microdata/data-centres/access
- Become a “Deemed Employee” of Statistics Canada: Obtain security clearance, declare conflicts of interest, sign an agreement for data access
- Peter Kitchen and Li Wang, Analysts - rdc@mcmaster.ca
- https://rdc.mcmaster.ca/

McMaster University | Library

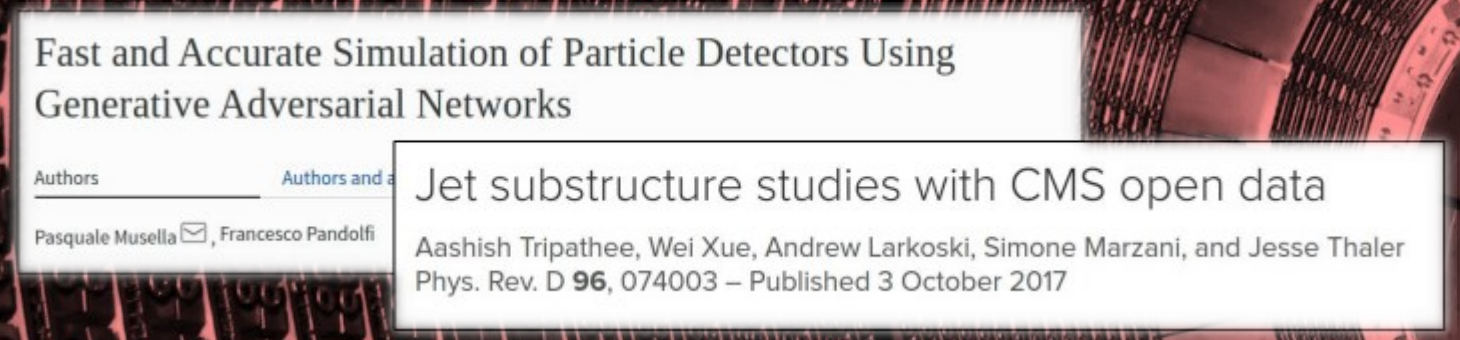
Finding Data: **Domain Specific Repositories**

- MIRA Open access data repositories: https://mira.mcmaster.ca/research/open-access-data-repositories
- Nature – Repositories by discipline: https://www.nature.com/sdata/policies/repositories
- NIH Recommended Repositories: https://sharing.nih.gov/data-management-and-sharing-policy/sharing-scientific-data/repositories-for-sharing-scientific-data
- Data Deposit Recommendation Service for Humanities: https://ddrs-dev.dariah.eu/ddrs/
- <odesi> https://search1.odesi.ca/#/

Photo by Kier In Sight on Unsplash.

Case Study: **CERN open data https://opendata.cern.ch/**

- CERN Open data – access point to a growing range of data produced through the research performed at CERN
- Explore more than three petabytes of open data from particle physics
- Datasets are shared under open licenses, and they are issued with a Digital Object Identifier (DOI) to make them citable objects.



- Photograph: Langstaff, R; Liquid Argon End Cap Cryostat https://cds.cern.ch/record/1005535

Case Study: **AquaMaps**

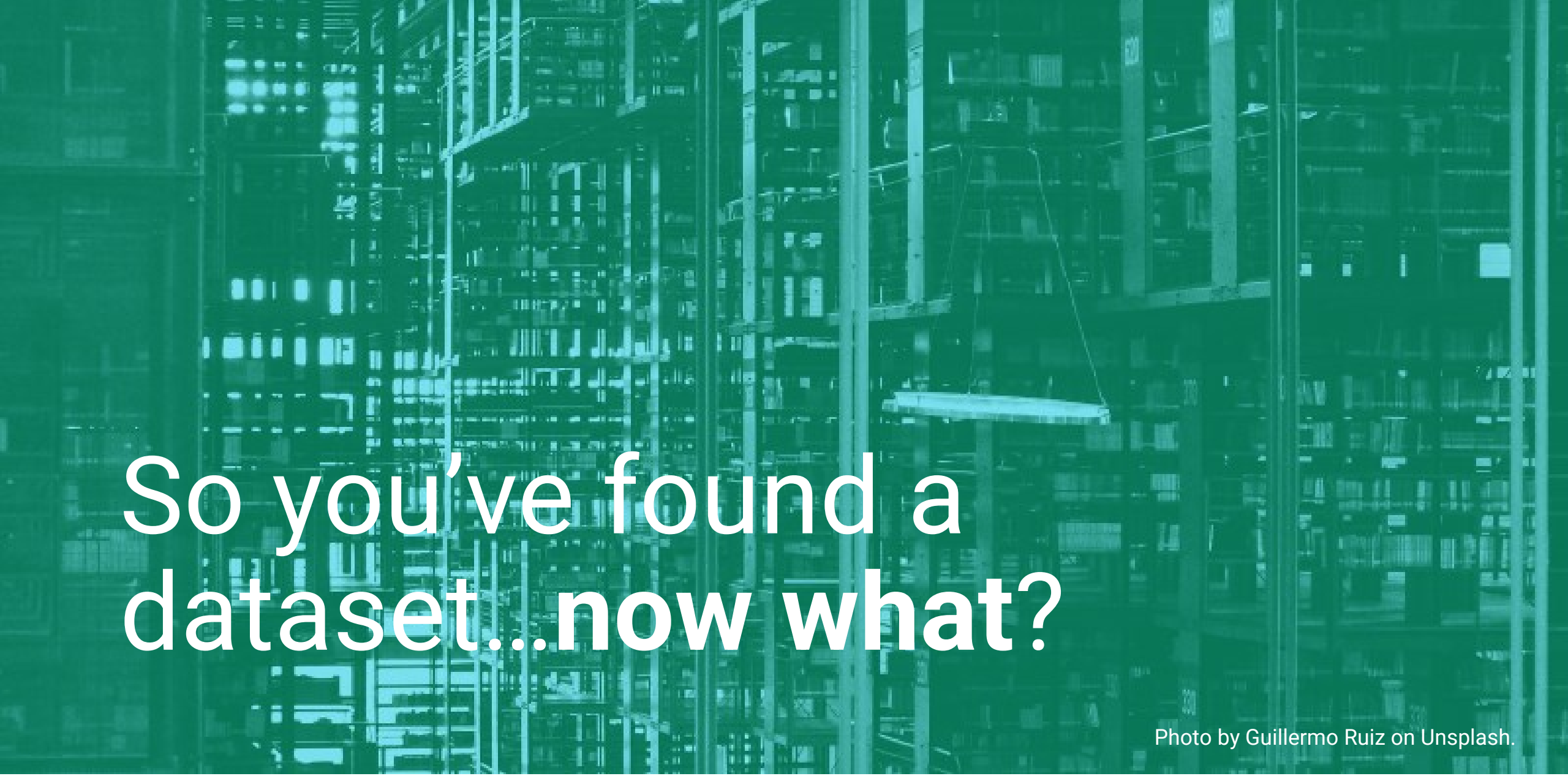
- **FishBase** – Gathers data on 34,300 fish species from different data sources – grey literature, books, journals, symposia, reports. Raw data are released on a website with a CC-BY-NC license.
- Used by consortium of 9 international institutions – 2275 citations
- Data processed by new algorithm developed by CNR-ISTI, available as a web service and producing a **new dataset**.
- **AquaMaps** – using new dataset, integrating with other data into a tool to generate predictions about marine species based on environmental and climate data; used for research into climate change's impact on marine species.

Gina Pavone, "Data Reuse Stories: Some concrete cases involving several institutions and consortia in Europe," *OpenAIRE*, 30 November 2020, https://www.openaire.eu/blogs/data-reuse-stories-some-concrete-cases-involving-several-institutions-and-consortia-in-europe

Photo by Adobe Stock Images.

RDM Certificate Program

- Certificate you can add to your CV
- Attend 7 RDM workshops to receive the certificate!
- Go to this website to verify today's session: https://u.mcmaster.ca/verification
- Learn more about the Certificate Program: https://scds.ca/certificate-program



So you've found a dataset...**now what**?

Photo by Guillermo Ruiz on Unsplash.

McMaster University | Library

Meteorological data for three atmospheric river case studies and Python programs for calculating column relative humidity and primary condensation rate

Contact Dataset Administrator

Description: Atmospheric rivers (ARs) are long, narrow, and transient corridors of strong horizontal water vapor transport that frequently lead to heavy precipitation where they are forced upward. The presence and strength of ARs are often described using the integrated water vapor (IWV) and the integrated vapor transport (IVT). However, the associated precipitation is not directly correlated with these two variables. Instead, the intensity of precipitation is mainly determined by the net convergence of moisture flux and the initial degree of saturation of the air column. The column relative humidity (CRH) and primary condensation rate (PCR) are two supplements to the standard AR analysis to focus attention on the heavy precipitation potential. Datasets and two Python programs presented here can be used, for demonstration purposes, to calculate and verify the CRH and PCR in three case studies of the AR events in 2020.

Authors: Mo, Ruping; Environment and Climate Change Canada; https://orcid.org/0000-0002-0284-0439

Keywords: column relative humidity
principal condensation rate
atmospheric river
heavy precipitation
integrated water vapor
integrated vapor transport

Field of Research: Earth and related environmental sciences > Atmospheric sciences > Meteorology and weather

Date: 2021-06-23

Photo by Adobe Stock Images.

Restricted Access Data

- Dataset may have an application process to verify usage conforms with requirements
- Application may ask you for information including:
  - Your name and affiliation,
  - Research project description,
  - A data security plan,
  - A data use/sharing agreement,
  - Ethics approval or application,
  - Qualifications and relevant certifications
- Sensitive data - qualitative data, genomics data, endangered species, traumatic brain injury research, personal information, and more.

Data Sharing Agreements

- An agreement between a researcher/research group and the data source organization
- **Documents** what data will be shared, how it must be stored and secured, and how it will be used.
- McMaster Research Ethics Board – Simple Data Sharing Agreement
- McMaster Industry Liaison Office – Material Transfer Agreement Templates

Photo by Adobe Stock Images.

Secure Empirical Analysis Lab (SEAL)

- **High Security Computer Lab** within Spark at McMaster
- Safe place to **store** and **access confidential data**
- **Companies, governments, researchers, community orgs** – may contain personal information or intellectual property.
- Secure remote access via MobiKey
- Lily Wang – SEAL Manager - https://seal.mcmaster.ca/

Research Ethics for Secondary Data

- Research Ethics is still required for secondary data analysis of data from human subjects.
- If data contains non-identifiable information, you need to apply for review but do not need to seek consent from the original participants.
- Consent is needed when:
  - Information can be linked to individuals
  - Possibility for identification in published reports or through data linkage.
- Data Linkage – two combined datasets. New research opportunities but potential for re-identification with anonymized data.

Photo by Clint Adair on Unsplash.

Indigenous Data Sovereignty

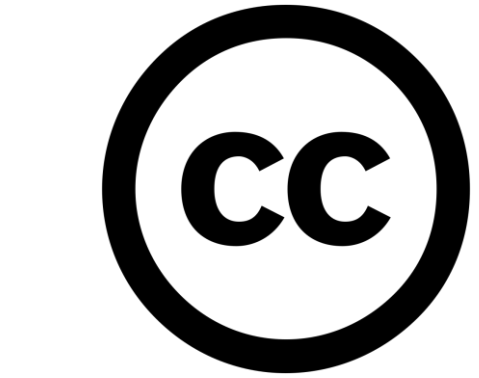
- Determined by the kind of data you hope to access.
- First Nations Information Governance Centre - regional health, early childhood, education, and employment data.
- Promoting advancement of FN health and wellness by facilitating research that will benefit FN people within a context.
- OCAP principles, Métis principles, Inuit Qaujimajatuqangit, Global Indigenous Data Alliance's CARE principles.

Image by BirdingInSpain, *Scanned primary feather of Short-toed Eagle Circaetus gallicus*. CC BY-SA 3.0 via Wikimedia Commons.

McMaster University | Library

Understanding Data Licensing

- **Creative Commons (CC) -** (creativecommons.org)
  - CC0 – public domain dedication
  - CC-BY – require attribution
  - There are further restrictions that may be added such as NC
- **Open Data Commons** (opendatacommons.org)
  - Similar licenses to CC but built for data
  - PDDL - Public Domain Dedication and License
  - ODC-By – require attribution
  - ODbL – attribution and share alike
- **Traditional Knowledge (TK):** In addition to the First Nations Information Governance Centre's OCAP® training, communities may also use TK licenses - **localcontexts.org/licenses**



In our Stó:lō culture, certain types of knowledge are restricted in some way. This knowledge is considered sacred, secret, potent and/or private, and only certain people or families can and should have access to them. We call this xa:xa in our language. This label indicates that there is additional knowledge about a certain subject that cannot be shared on the website.

Images from Sq'éwlets, "Traditional Knowledge Labels," http://digitalsqewlets.ca/traditional-knowledge_connaissances_traditionnelles-eng.php and Creative Commons, fixed by Quibik.

Reusing datasets.

Project: Kristin's important chemistry project
Date: June 2013-April 2014
Description: Description of my awesome project here
Funder: Department of Energy, grant no: XXXXXX
Contact: Kristin Briney, kristin@myemail.com

ORGANIZATION

All files live in the 'ImportantProject' folder, with content organized into subfolders as follows:

– 'RawData': All raw data goes into this folder, with subfolders organized by date
– 'AnalyzedData': Data analysis files
– 'PaperDrafts': Draft of paper, including text, figures, outlines, reference library, etc.
– 'Documentation': Scanned copies of my written research notes and other research notes
– 'Miscellaneous': Other information that relates to this project

NAMING

Raw data files will be named as follows:

"YYYYMMDD_experiment_sample_ExpNum"
(ex: "20140224_UVVis_KMnO4_2.csv")

STORAGE

All files will be stored on my computer and backed up daily to

Kristin Briney, "README.TXT," Data Ab Initio, February 25, 2014 http://dataabinitio.com/?p=378

Reusing Data - **Unpacking a Dataset**

- **Data:** Data files – tabular data, images, audio files, code.
- **README**: "Passport" for data - simple document that should help you understand the project. Contains folder hierarchy and file organization, description of important file contents.
- **Data Dictionaries**: A document for tabular data that describing names, labels, units, and constraints.
- **Codebooks:** Like data dictionaries but for survey or statistical data—includes the survey layout and structure, and codes for questions and answers.

McMaster University | Library

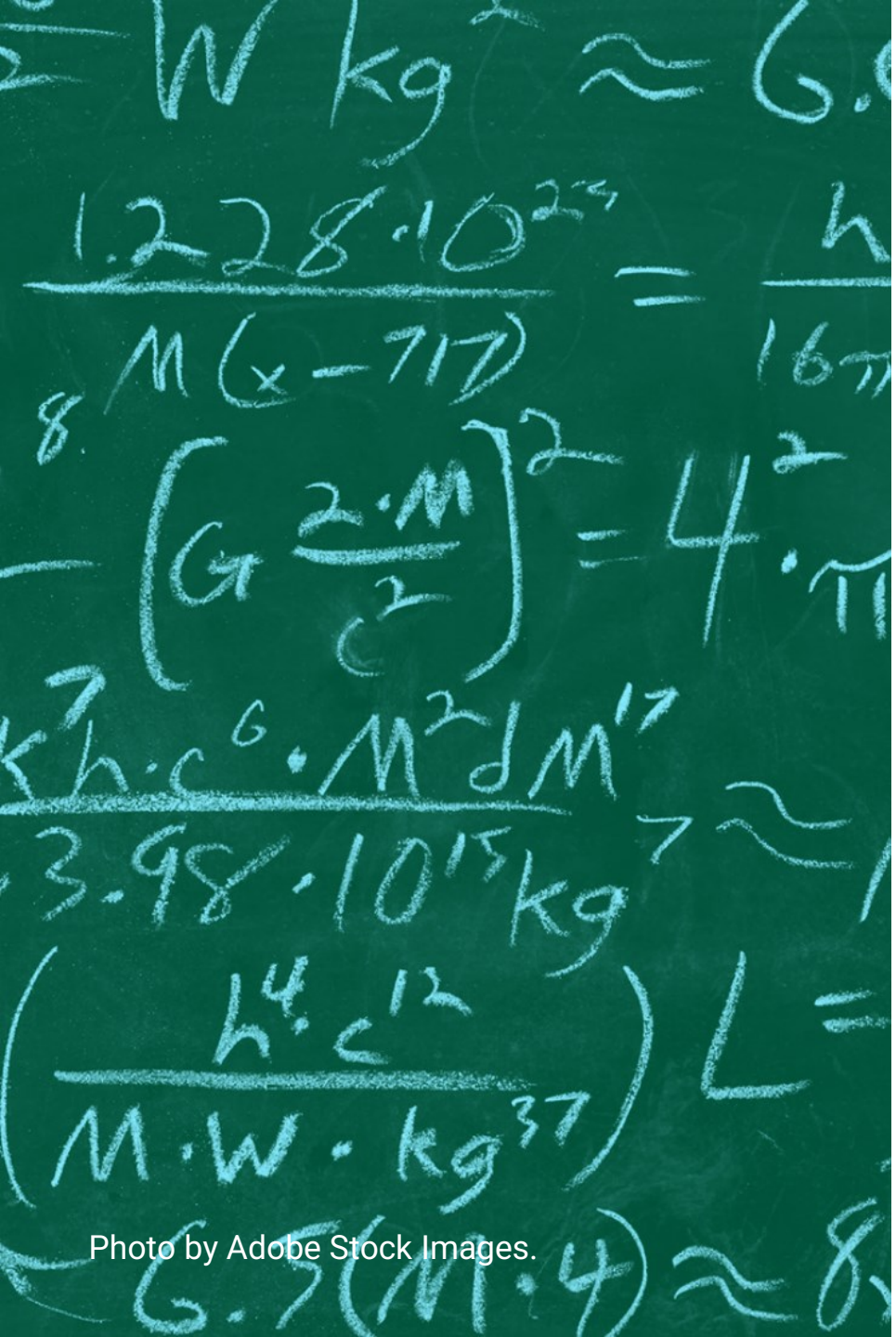


Photo by Adobe Stock Images.

Data Quality Checking

- Compare dataset to the README file – does the data appear as expected?
- Spot check for duplicates, table errors, missing data, etc.
- **Everyone is responsible for quality** - share any questions and report concerns with the researcher and data repository.

Reusing Data - **Unpacking a Dataset**

Meteorological data for three atmospheric river case studies and Python programs for calculating column relative humidity and primary condensation rate

Contact Dataset Administrator

Description:	Atmospheric rivers (ARs) are long, narrow, and transient corridors of strong horizontal water vapor transport that frequently lead to heavy precipitation where they are forced upward. The presence and strength of ARs are often described using the integrated water vapor (IWV) and the integrated vapor transport (IVT). However, the associated precipitation is not directly correlated with these two variables. Instead, the intensity of precipitation is mainly determined by the net convergence of moisture flux and the initial degree of saturation of the air column. The column relative humidity (CRH) and primary condensation rate (PCR) are two supplements to the standard AR analysis to focus attention on the heavy precipitation potential. Datasets and two Python programs presented here can be used, for demonstration purposes, to calculate and verify the CRH and PCR in three case studies of the AR events in 2020.
Authors:	Mo, Ruping; Environment and Climate Change Canada; https://orcid.org/0000-0002-0284-0439
Keywords:	column relative humidity principal condensation rate atmospheric river heavy precipitation integrated water vapor integrated vapor transport
Field of Research:	Earth and related environmental sciences > Atmospheric sciences > Meteorology and weather
Date:	2021-06-23
Publisher:	Federated Research Data Repository / dépôt fédéré de données de recherche
URI:	https://doi.org/10.20383/102.0472

Geographic Coverage:		
	Place Name	Global
	Country	Canada
	Country	United States

Files in Dataset - Total Size - 619.03 MB

ar_glbhyb_2020081500_000.nc	103.16 MB
ar_glbhyb_2020112600_000.nc	103.16 MB
ar_glbhyb_2020112612_024.nc	103.16 MB
ar_glbhyb_2020112700_000.nc	103.16 MB
ar_glbhyb_2020112712_000.nc	103.16 MB
ar_glbhyb_2020112800_000.nc	103.16 MB
frdr-checksums-and-filetypes.md	3.64 KB
hrly_pr_15Aug2020.csv	4.13 KB
hrly_pr_28Nov2020.csv	4.26 KB
hrly_pr_time_series_26_28Nov2020.csv	1.12 KB
LICENSE.txt	13 KB
P1_IWV_IVT_CRH.py	3.44 KB
P2_CRH_PCR.py	2.25 KB
Port_Hardy_YZT_Sounding_2020112712.csv	6.83 KB
ReadMe.txt	12.56 KB

Download Dataset

Access to this dataset is subject to the following terms:

Citation

Mo, R. (2021) Meteorological data for three atmospheric river case studies and Python programs for calculating column relative humidity and primary condensation rate. Federated Research Data Repository. https://doi.org/10.20383/102.0472

Get Citation

Reusing Data - **Unpacking a Dataset**

McMaster University | Library

README – General + Sharing/Access Information

```
This README.txt file was generated on 2021-06-16 by Ruping Mo, Environment
and Climate Change Canada [ruping.mo@ec.gc.ca]

--------------------
GENERAL INFORMATION
--------------------

1. Title of Dataset:
      Meteorological data for three atmospheric river case studies and Python
      programs for calculating column relative humidity and primary
      condensation rate.

2. Author Information
      Name: Mo, Ruping
      Email: ruping.mo@ec.gc.ca
      Institution: Environment and Climate Change Canada (ECCC)

3. Date of data collection:
      2020-08-20--2021-0501

4. Geographic location of data collection:
      Numerical Weather Prediction (NWP) model data in the netCDF files
      cover the globe at horizontal spacings about 15 km. Observed
      precipitation data were collected at weather stations across British
      Columbia, Canada.

---------------------------
SHARING/ACCESS INFORMATION
---------------------------

1. Licenses/restrictions placed on the data:
      These data are available under a CC BY 4.0 license
      [https://creativecommons.org/licenses/by/4.0/]

2. Links to publications that cite or use the data:
      These data and programs are used in Mo (2020) and Mo et al. (2021).

      Mo, R. (2020). Diagnosing primary condensation rate attributed to the
      moisture convergence: Applications to atmospheric river analysis and
      extratropical storm classification. A117-0004, presented at 2020 Fall
      Meeting, AGU, 1-17 December. doi: 10.1002/essoar.10505440.1

      Mo, R., So, R., Brugman, M. M., Mooney, C., Liu, A. Q., Jakob, M.,
      Castellan, A., & Vingarzan, R. (2021). Column relative humidity and
      primary condensation rate as two useful supplements to atmospheric river
      analysis. Water Resources Research. In revision.
```

Files in Dataset - Total Size - 619.03 MB

- ar_glbhyb_2020081500_000.nc
- ar_glbhyb_2020112600_000.nc
- ar_glbhyb_2020112612_024.nc
- ar_glbhyb_2020112700_000.nc
- ar_glbhyb_2020112712_000.nc
- ar_glbhyb_2020112800_000.nc
- frdr-checksums-and-filetypes.md
- hrly_pr_15Aug2020.csv
- hrly_pr_28Nov2020.csv
- hrly_pr_time_series_26_28Nov2020.csv
- LICENSE.txt
- P1_IWV_IVT_CRH.py
- P2_CRH_PCR.py
- Port_Hardy_YZT_Sounding_2020112712.csv
- ReadMe.txt

Download Dataset

Access to this dataset is subject to the following terms:

Creative ...ons.org/license

Citation

Mo, R. (2021) Meteorological data for three atmospheric river case studies and Python programs humidity and primary condensation rate. Federated Research Data Repository. https://doi.org/10

Reusing Data - Unpacking a Dataset

1. File List

1) ar_glbhyb_2020112712_000.nc:
A NetCDF data file containing some atmospheric variables valid at 1200 UTC 27 Nov 2020, based on the GDPS analysis (0-hour forecast).

2) ar_glbhyb_2020112612_024.nc:
A NetCDF data file containing some atmospheric variables valid at 1200 UTC 27 Nov 2020, based on the GDPS 24-hour forecast with model initialization at 1200 UTC 26 Nov 2020.

3) ar_glbhyb_2020112600_000.nc:
A NetCDF data file containing some atmospheric variables valid at 0000 UTC 26 Nov 2020, based on the GDPS analysis (0-hour forecast).

4) ar_glbhyb_2020112700_000.nc:
A NetCDF data file containing some atmospheric variables valid at 0000 UTC 27 Nov 2020, based on the GDPS analysis (0-hour forecast).

5) ar_glbhyb_2020112800_000.nc:
A NetCDF data file containing some atmospheric variables valid at 0000 UTC 28 Nov 2020, based on the GDPS analysis (0-hour forecast).

6) ar_glbhyb_2020081500_000.nc:
A NetCDF data file containing some atmospheric variables valid at 0000 UTC 15 Aug 2020, based on the GDPS analysis (0-hour forecast).

7) Port_Hardy_YZT_Sounding_2020112712.csv:
The sounding data of Port Hardy, British Columbia, Canada, valid at 1200 UTC 27 Nov 2020.

8) hrly_pr_time_series_26_28Nov2020.csv:
Time series (1200 UTC 26--1200 UTC 28 Nov 2020) of hourly precipitation amounts observed at three weather stations in British Columbia, Canada.

9) hrly_pr_28Nov2020.csv:
Hourly precipitation amounts valid at 0000 and 0100 UTC 28 Nov 2020, observed at 101 weather stations in British Columbia, Canada.

10) hrly_pr_15Aug2020.csv:
Hourly precipitation amounts valid at 0000 and 0100 UTC 15 Aug 2020, observed at 98 weather stations in British Columbia, Canada and Alaska, USA.

11) P1_IWV_IVT_CRH.py:
A Python program used to calculate IWV, IVT, and CRH from a given sounding data file.

12) P2_CRH_PCR.py:
A Python program used to calculate CRH and PCR based on the GDPS output. ar_glbhyb_2020081500_000.nc

13) ReadMe.txt

2. Relationship between files, if important:

1) The Python program P1_IWV_IVT_CRH.py reads input data from Port_Hardy_YZT_Sounding_2020112712.csv.

2) The Python program P2_CRH_PCR.py reads input data from ar_glbhyb_2020081500_000.nc

3. Are there multiple versions of the dataset? no

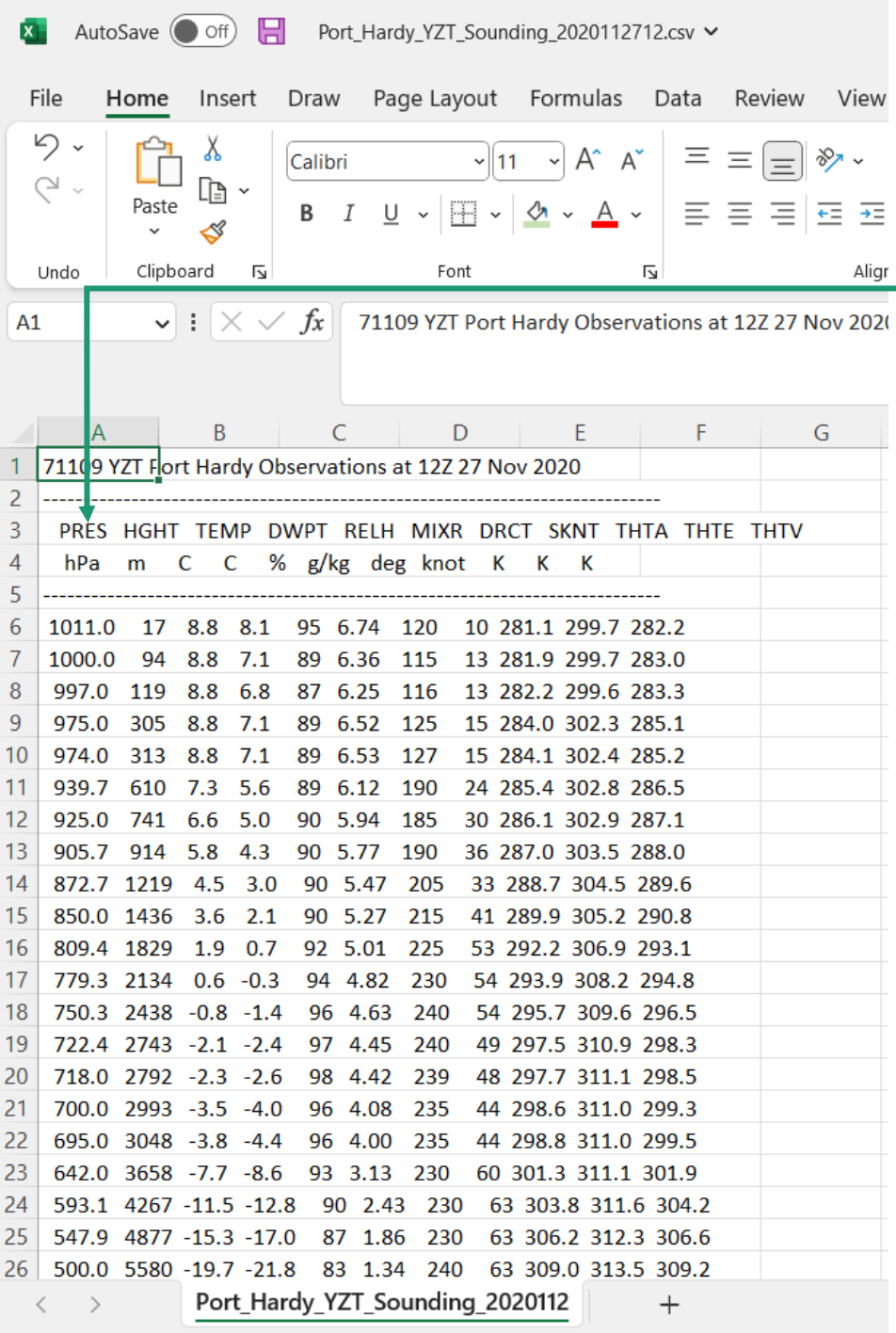
METHODOLOGICAL INFORMATION

---

1. Description of methods used for collection/generation of data:
The methodology for data collection and generation is described in Mo, R., So, R., Brugman, M. M., Mooney, C., Liu, A. Q., Jakob, M., Castellan, A., & Vingarzan, R. (2021). Column relative humidity and primary condensation rate as two useful supplements to atmospheric river analysis. Water Resources Research. In revision.

2. Methods for processing the data:
In the netCDF files, IWV, ISWV, ICW, IVTU, IVTV, ICTU, and ICTV are vertically integrated variables; others are raw model variables of the operational GDPS output. The vertical integration is carried out from the Earth's surface up to the 200 hPa level.

Reusing Data - **Unpacking a Dataset**



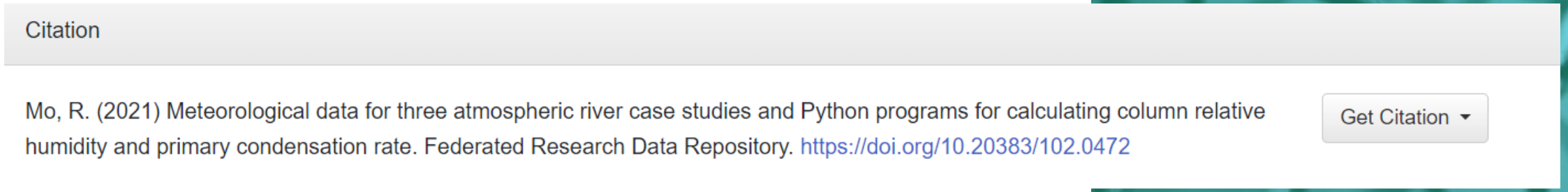

```
------------------------------------------------------------------
DATA-SPECIFIC INFORMATION FOR: Port_Hardy_YZT_Sounding_2020112712.csv
------------------------------------------------------------------

1. Number of variables: 13

2. Missing data codes: Empty spaces

3. Variable List:

Variable:       PRES
Description: Atmospheric pressure
Units:   hPa

Variable:       HGHT
Description: Geopotential height
Units:   m

Variable:       TEMP
Description: Temperature
Units:   C

Variable:       DWPT
Description: Dewpoint temperature
Units:   C

Variable:       RELH
Description: Relative humidity
Units:   %

Variable:       MIXR
Description: Mixing ratio
Units:   g/kg

Variable:       DRCT
Description: Wind direction
Units:   deg

Variable:       SKNT
Description: Wind speed
Units:   knot

Variable:       THTA
Description: Potential temperature
Units:   K

Variable:       THTE
Description: Equivalent potential temperature
Units:   K

Variable:       THTV
Description: Virtual potential temperature
Units:   K
```

Reusing Data - **Unpacking a Dataset**

McMaster University | Library

Reusing Data – **Citation**

- Proper attribution and credit
- Connect original publications and supporting data with secondary work
- Citation makes it easier to find datasets
- Most Data Repositories will include a citation box for you to easily cite data in your publication:



- Some guidance on formatting:

https://guides.library.columbia.edu/datacitation

Photo by Clint Adair on Unsplash.

McMaster RDM webpage: rdm.mcmaster.ca

Contact RDM services at: rdm@mcmaster.ca

Upcoming RDM webinars: rdm.mcmaster.ca/events

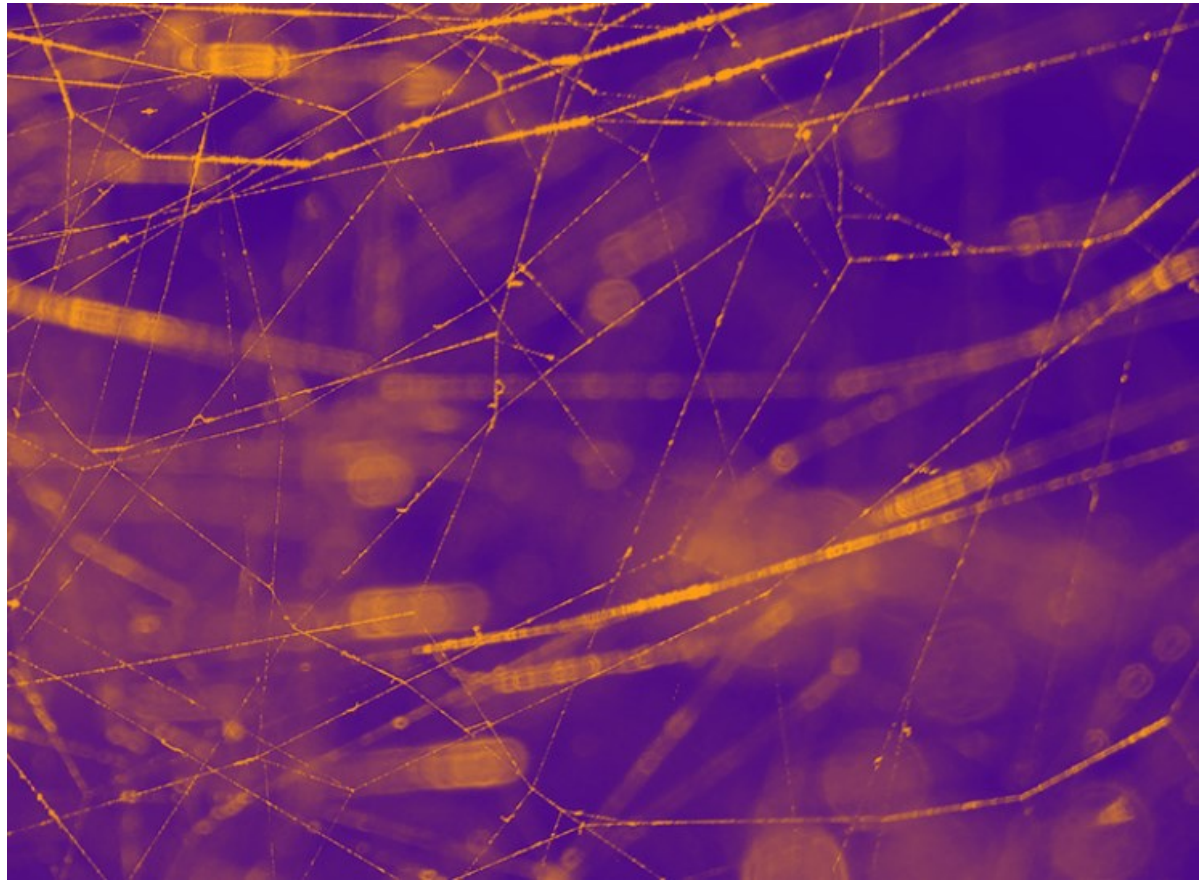
Recorded RDM webinars: u.mcmaster.ca/learn-rdm

Make an appointment with a Research Data Management Specialist:
u.mcmaster.ca/rdm-appointments

McMaster University | Library

RDM Community of Practice

- Monthly meetings of people interested in RDM at McMaster
- **Thursday March 30th – 11 AM**
  - Dr. Claudia Emerson discuss data ethics and Research Data Management (RDM). Dr. Emerson's work connects bioethics, epidemiology, public health, law, anthropology, and philosophy
- Connect with other researchers practicing RDM across the university!
- https://u.mcmaster.ca/rdm-community



McMaster University | Library

RDM

April 5 | 10:30-11:30am

Hybrid Workshop

Qualitative Data: Practices for RDM Planning & Sharing

u.mcmaster.ca/scds-events

SCDS

Library | McMaster University Library