My data collection is complete, now what?
Connecting researchers to Data Repositories that can support Cold Regions Researchers

Cold Regions Research Centre Days Conference
November 27th, 2020

P. Sairam[1], B.D. Persaud[1], M. Steeleworthy[2], and P. Van Cappellen[1]

[1] University of Waterloo
[2] Wilfrid Laurier University

Overview of Presentation

Email: yasairam@uwaterloo.ca

- Background of Project
- Introduction to Repositories
- Benefit of Repositories to Researchers
- FAIR Principles
- Selected Repositories:
  - Federated Research Data Repository (FRDR)
  - Scholars Portal Dataverse
  - DataStream
  - Polar Data Catalogue (PDC)
  - PANGAEA
  - Zenodo
- Discussion

Background of Project

- Canada has many useful, cost-effective (free) repositories
- "Deep-dive" comparison of **33** key characteristics specific to data management
- Make it easier for researchers to select an appropriate repository
  - Researchers may not be aware of which meet their unique needs

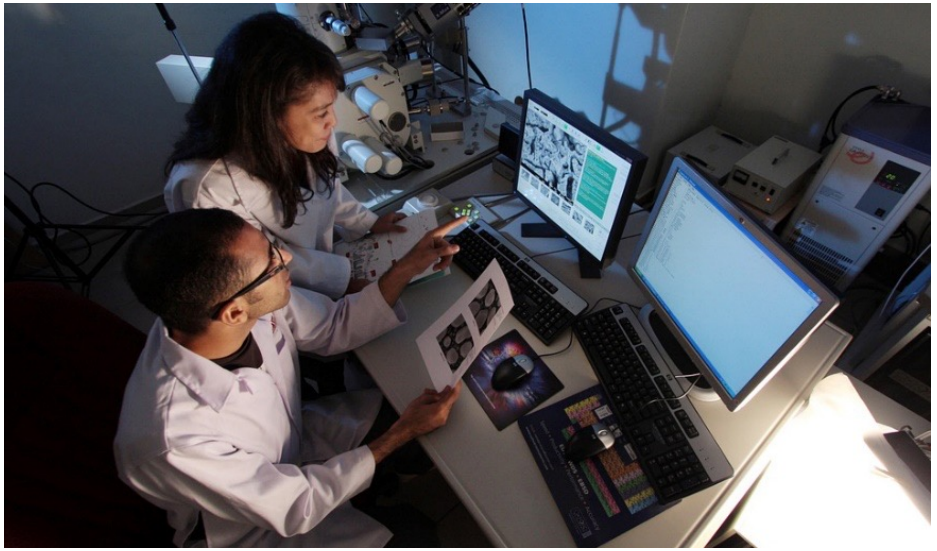


Image from The EU Research & Innovation Magazine by Benedict O'Donnell

What are Repositories?

- A **data repository** is, "a large database infrastructure" that can host several data sets (Brook, 2018)
- Primary purposes include collecting, managing, and storing data sets
- From this consolidation of data, users can create data reports for sharing and analysis

Benefits for Use of Data Repositories

(Blick, 2018)

Open-Access and Increased Visibility	• Researchers and institutions across the world can access and use deposited data for research
Collections	• Curation of data sets for specific branches of research
Preservation	• 'Work-space' for in-progress collaboration, able to use peer-review and embargo periods
Support for Researchers	• Allowing digital access for students and future researchers to be able to further their research

Features for an Ideal Research Repository

(Neuwirth and Alter , 2018)

Ease of Use

Consolidation of data in a single repository, support for many data types; ability to connect to scientific publications and other repositories

Accessibility and Breadth of Information

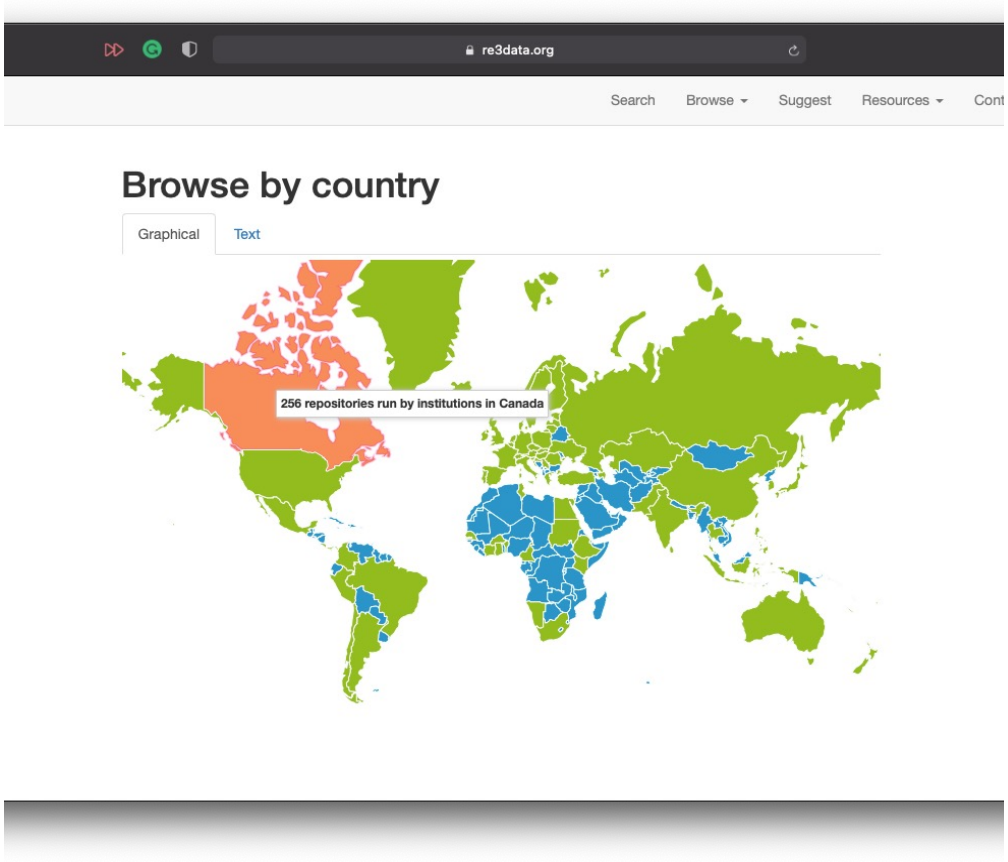
Adherence to **FAIR (Findable, Accessible, Interoperable, Reusable) principles** to access and use data, high reproducibility; scaled support as repository grows

Features

Automated processes and access to advanced analytics of preserved data; does the database support you branch of research?

Methods

- We chose the following repositories to compare:
  - Federated Research Data Repository (FRDR) - Canadian
  - Scholars Portal Dataverse - Canadian
  - DataStream – Canadian
  - Polar Data Catalogue (PDC) - Canadian
  - PANGAEA - German
  - Zenodo - European
- Many Canadian Researchers
- 33 characteristics were compared from repository websites and external journals; database representatives corroborated our information



Federated Research Data Repository (FRDR) – Established in 2016

Summary

Open, national platform for researchers to access and share Canadian research data

Benefit to Researchers

Access to multiple repositories

Stores research data – up to 3 TB per researcher

Support for any type of research data

Support for Researchers

Free creation of DOI

Large support for licensing options

Archivematica for Preservation

UNIVERSITY OF SASKATCHEWAN

GWF
GLOBAL WATER FUTURES

General Repository

Purpose-built for large datasets and own storage group (e.g. GWF, Water Institute)

Supports Embargo Period

28-Oct-2020	Hydrometeorological observations at three boreal forest sites (aspen, jack pine, and black spruce) located in central Saskatchewan, Canada	... G.; Black, T. Andrew
13-Oct-2020	Warm-air entrainment and advection during alpine blowing snow events	Aksamit, Nikolas; Pomeroy, John
8-Oct-2020	EMDNA: Ensemble Meteorological Dataset for North America	Tang, Guoqiang; Clark, Martyn P.; Papalexiou, Simon Michael; Newman, Andrew J.; Wood, Andrew W.; Brunet, Dominique; Whitfield, Paul H.
23-Sep-2020	Atmospheric boundary layer measurements from Athabasca Glacier field campaign, June 2015	Conway, Jono; Helgason, Warren; Pomeroy, John; Sicart, Jean-Emmanuel; Johnson, Bruce

Scholars Portal Dataverse

– Established in 2018

Summary

Open-source, free for researchers, platform for most at Canadian universities to access and share Canadian research data

Benefit to Researchers

Focus on long-term preservation

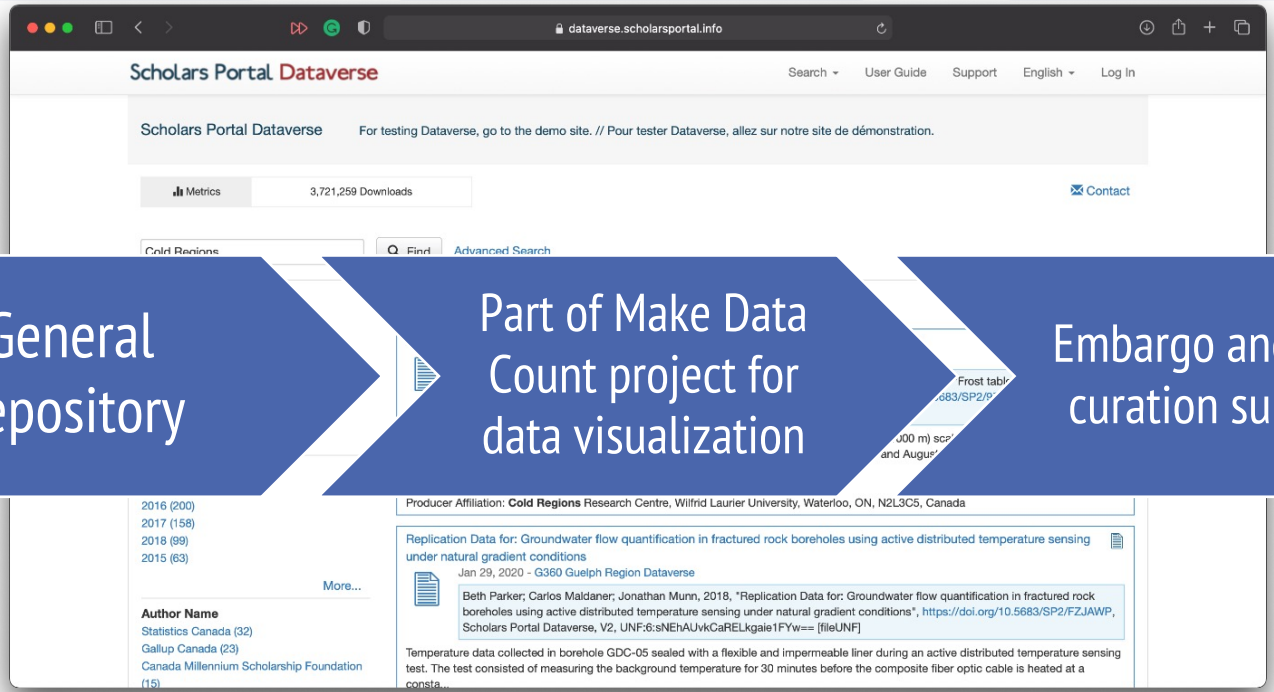
Stores research data – up to 1 TB per researcher

Support for any type of research data

Support for Researchers

Free creation and linking of DOIs

Promotes CC0 licensing option



General Repository

Part of Make Data Count project for data visualization

Embargo and data curation support

DataStream

– Established in 2016

Summary

Open access repository specific for information about local watersheds across Canada

Benefit to Researchers

Open-access data

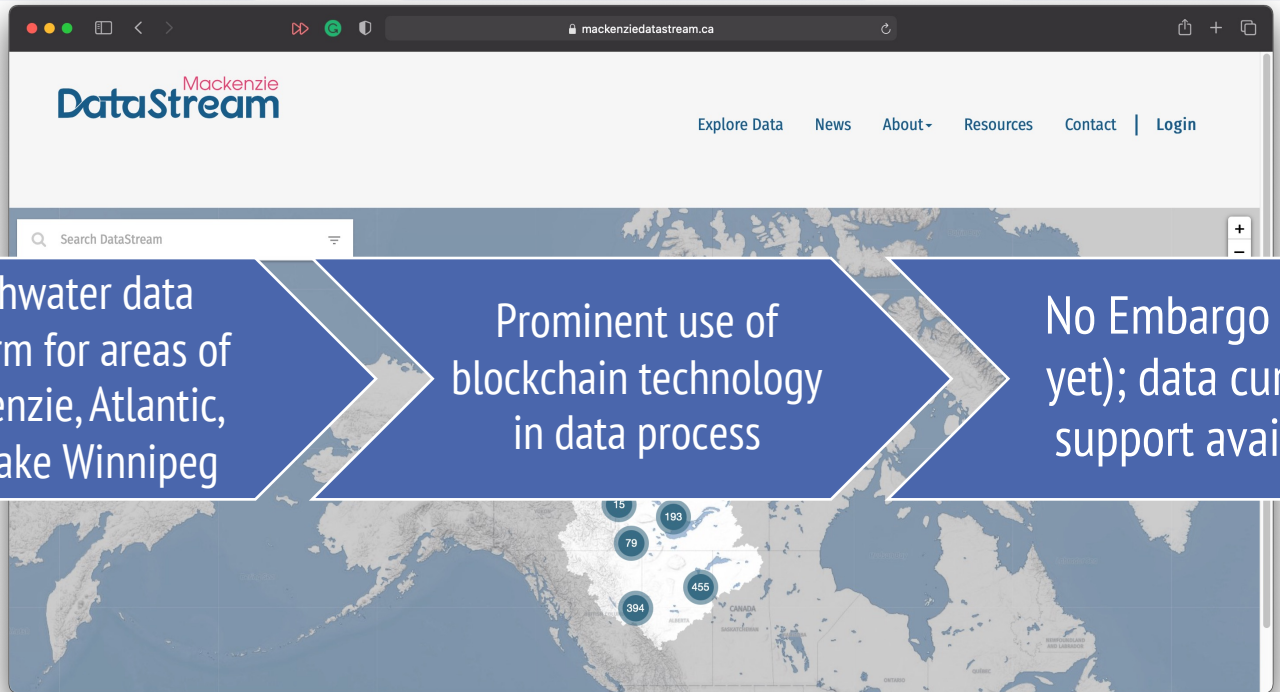
Built-in visualization metrics for tabular data

Support for .csv files

Support for Researchers

Free creation of DOI

Promotes ODC-BY v1.0 and PDDL licensing options



Freshwater data platform for areas of Mackenzie, Atlantic, and Lake Winnipeg

Prominent use of blockchain technology in data process

No Embargo (as of yet); data curation support available

Polar Data Catalogue (PDC)

– Established in 2007

Summary

Free, open access repository for researchers accessing datasets generated by Arctic and Antarctic researchers

Benefit to Researchers

Focus on polar data

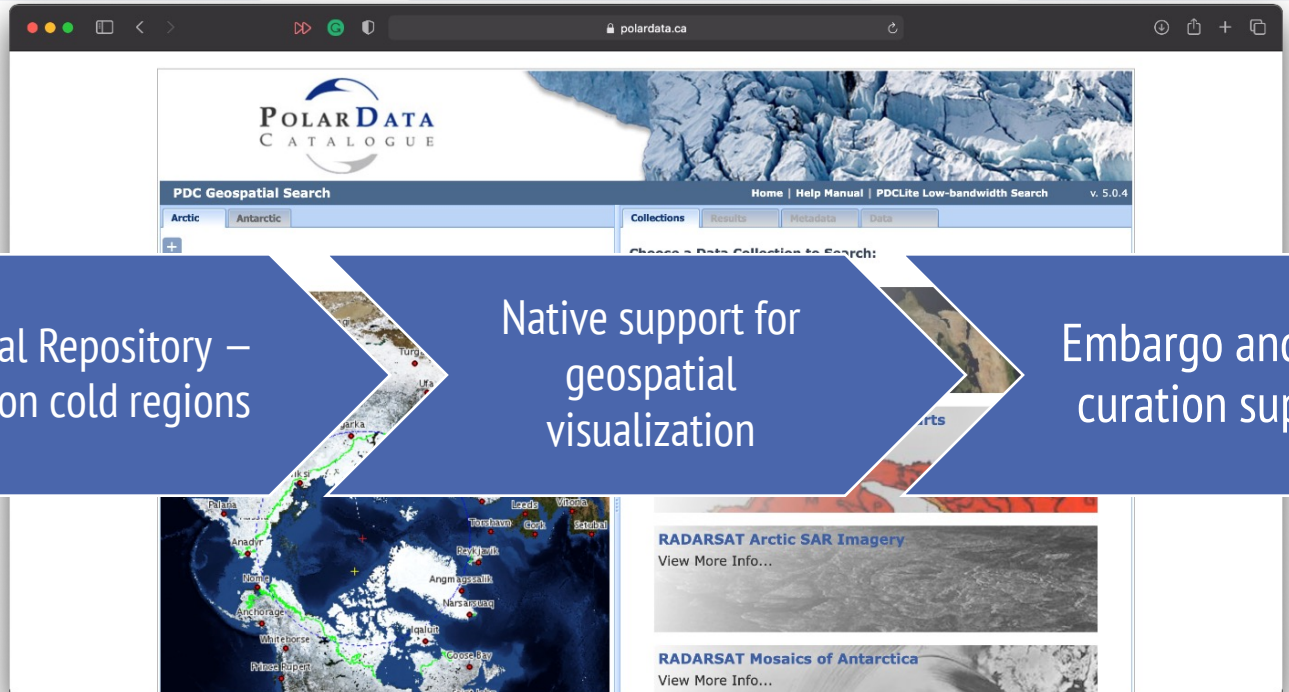
Mediated by data manager

Large amounts of polar data

Support for Researchers

Free creation and linking of DOIs

Many supported Metadata schemas including ISO 19115 standard



General Repository – focus on cold regions

Native support for geospatial visualization

Embargo and data curation support

PANGAEA

– Established in 1995

Summary

Open-access database with a focus for storing and accessing georeferenced data in a pursuit of long-term availability

Benefit to Researchers

Focus on long-term preservation

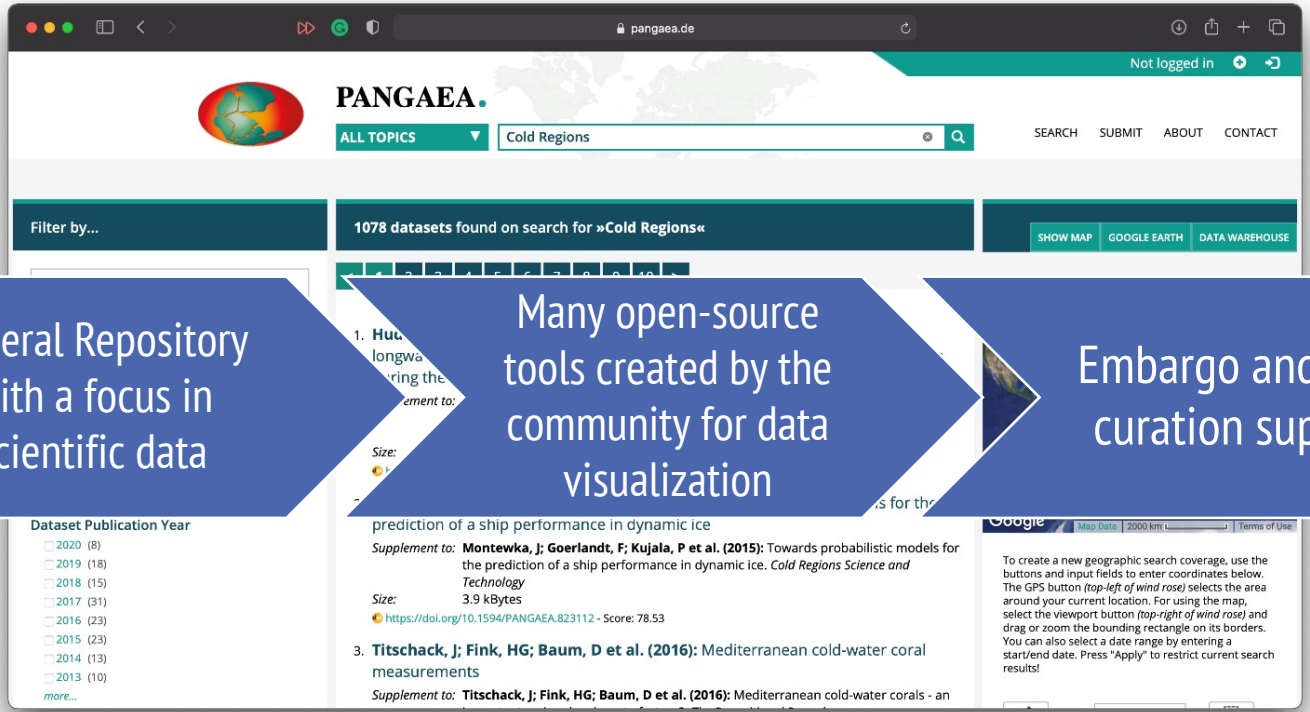
ORCID ID author identification support

Native support of R and Python for data visualization

Support for Researchers

DOI and International GeoSample Identifiers

Promotes many CC licensing options



General Repository with a focus in scientific data

Many open-source tools created by the community for data visualization

Embargo and data curation support

Zenodo

– Established in 2013

Summary

Open-source repository built for researchers in mind hosted by CERN with specialization in scientific datasets

Benefit to Researchers

Focus on long-term preservation

Stores research data – no limit

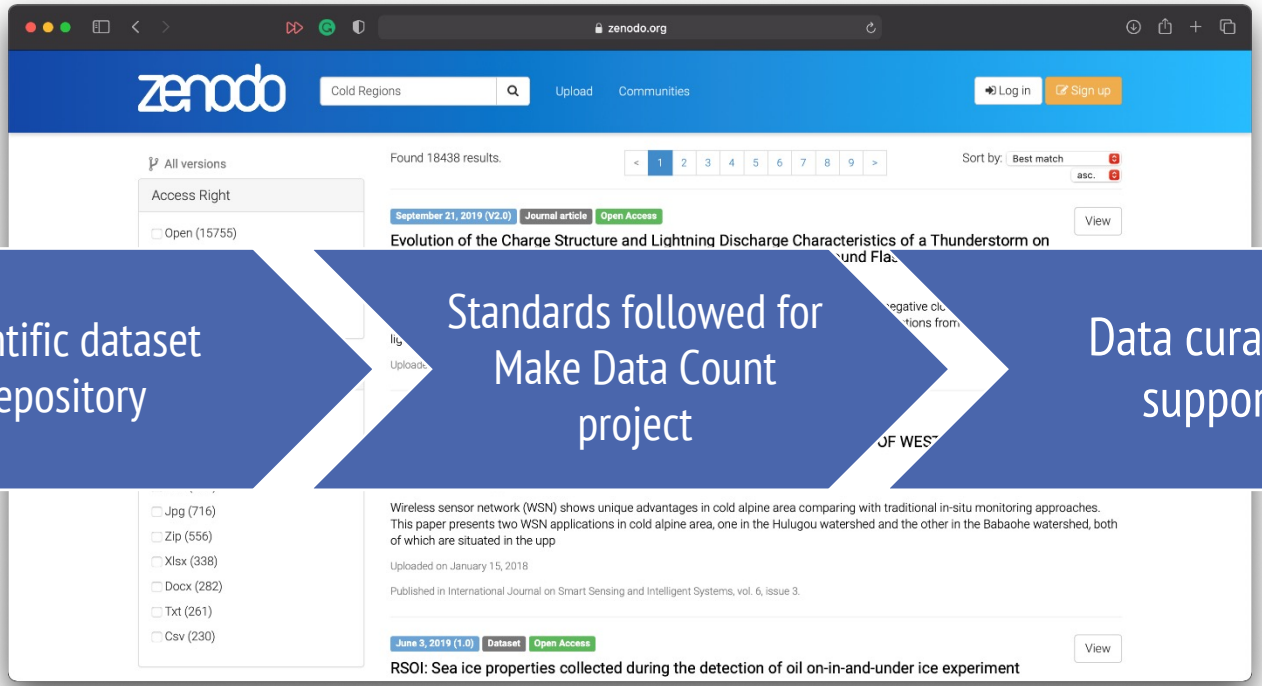
Support ORCID ID author identification

Support for Researchers

Free creation of DOIs

Support for 400 open licensing options

Native GitHub support for programming



Scientific dataset Repository

Standards followed for Make Data Count project

Data curation support

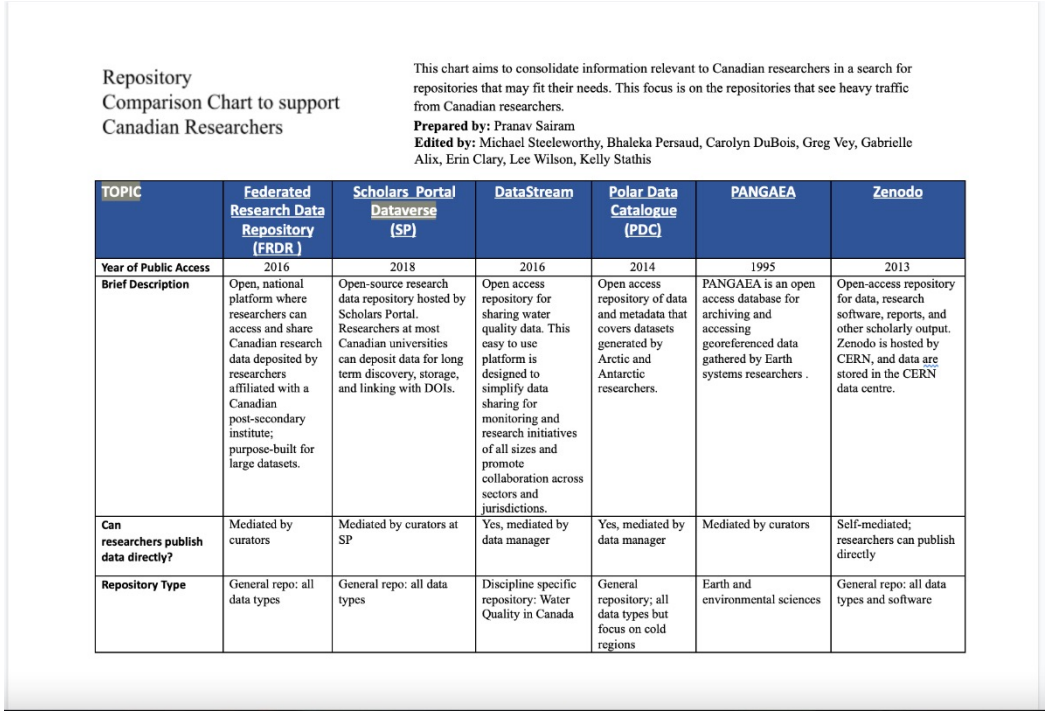
Recommendations for Future Repository Development

- Of the explored repositories, DataStream appears to have comprehensive visualization on the data despoited
- As a researcher, can we get repository developers to focus more on visualization support?
- Areas for Repository Development: Outdated design philosophy, no heavy emphasis on data visualization tools for many repositories, integration with GitHub
  - How can researchers benefit from a greater focus in this aspect?

Image from Mackenzie DataStream, published on May 9, 2019 entitled 'Data Specialist Intern joins DataStream team!'

My Next Steps

- Share information to a wider audience via GWF/UW website
- Our project can be viewed at the following link: https://drive.google.com/file/d/1MB1h0wfCP_2QUFw-Zp-mXFF_ug7NMUN/view?usp=sharing



Repository
Comparison Chart to support
Canadian Researchers

This chart aims to consolidate information relevant to Canadian researchers in a search for repositories that may fit their needs. This focus is on the repositories that see heavy traffic from Canadian researchers.

Prepared by: Pranav Sairam
Edited by: Michael Steeleworthy, Bhaleka Persaud, Carolyn DuBois, Greg Vey, Gabrielle Alix, Erin Clary, Lee Wilson, Kelly Stathis

TOPIC	Federated Research Data Repository (FRDR)	Scholars Portal Dataverse (SP)	DataStream	Polar Data Catalogue (PDC)	PANGAEA	Zenodo
Year of Public Access	2016	2018	2016	2014	1995	2013
Brief Description	Open, national platform where researchers can access and share Canadian research data deposited by researchers affiliated with a Canadian post-secondary institute; purpose-built for large datasets.	Open-source research data repository hosted by Scholars Portal. Researchers at most Canadian universities can deposit data for long term discovery, storage, and linking with DOIs.	Open access repository for sharing water quality data. This easy to use platform is designed to simplify data sharing for monitoring and research initiatives of all sizes and promote collaboration across sectors and jurisdictions.	Open access repository of data and metadata that covers datasets generated by Arctic and Antarctic researchers.	PANGAEA is an open access database for archiving and accessing georeferenced data gathered by Earth systems researchers .	Open-access repository for data, research software, reports, and other scholarly output. Zenodo is hosted by CERN, and data are stored in the CERN data centre.
Can researchers publish data directly?	Mediated by curators	Mediated by curators at SP	Yes, mediated by data manager	Yes, mediated by data manager	Mediated by curators	Self-mediated; researchers can publish directly
Repository Type	General repo: all data types	General repo: all data types	Discipline specific repository: Water Quality in Canada	General repository; all data types but focus on cold regions	Earth and environmental sciences	General repo: all data types and software

What Repository is Best?

- **Depends on your needs, there is a repository out there for you**
- FRDR is a national platform built specifically for large dataset support and data curation & for embargo
- Dataverse has great embargo and data curation support
- DataStream can be used for Canadian institute water quality data and you can support a broader national initiative
- PDC has a specific focus on Arctic and Antarctic data deposits & for embargo
- **Support local if you can, they all have free data curation support to help ensure that your data deposited following establish standards and best practices**
- PANGAEA is a generalist repository with great community features/support
- Zenodo is repo that offers integration with GitHub so very useful for researchers doing lots of coding, and easy to navigate but limited checks on data quality (on your own)

"Why Should Cold Regions Researchers Think About Repositories/Open Access data?"

- Journals and Funding Agencies are asking

DATA POLICY

First adopted by Publications Committee November 1993 [Revi

AGU affirmed in its 2015 position statement that "Earth and sp
sponsoring institutions." Following this statement and to advan
evaluate, replicate, and build upon the reported research must l

For the purposes of this policy, data include, but are not limited

- Data used to generate, or be displayed in, figures, graphs
- New protocols or methods used to generate the data in a
- New code/computer software used to generate results o
- Derived data products reported or described in a paper.

AGU encourages authors to identify and archive their data in ap
authors are expected to curate the above data for at least 5 yea
be reliably made available to anybody requesting data may not

Hydrology and Earth System Sciences

An interactive open-access journal of the European Geosciences Union

| EGU.eu | EGU Publications | EGU Highlight Articles | Contact | Imprint | Data protection |

Data policy

The output of research is not only journal articles but also data sets, model code, samples, etc. Only the entire network of interconnected information can guarantee integrity, transparency, reuse, and reproducibility of scientific findings. Moreover, all of these resources provide great additional value in their own right. Hence, it is particularly important that data and other information underpinning the research findings are "findable, accessible, interoperable, and reusable" (FAIR) not only for humans but also for machines.

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

Therefore, Copernicus Publications requests depositing data that correspond to journal articles in reliable (public) data repositories, assigning digital object identifiers, and properly citing data sets as individual contributions. Please find your appropriate data repository in the registry for research data repositories: re3data.org. A data citation in a publication resembles a bibliographic citation and needs to be included in the publication's reference list. To foster the accessibility as well as the proper citation of data, Copernicus Publications requires all authors to provide a statement on the availability of underlying data as the last paragraph of each article (see section data availability). In addition, data sets, model code, video supplements, video abstracts, International Geo Sample Numbers, and other digital assets should be linked to the article through DOIs in the assets tab. With Earth System Science Data (ESSD) Copernicus Publications provides a journal dedicated to the publication of data papers, including peer review of data sets. Authors should consider submitting a data paper to ESSD in addition to their research paper in another journal published by Copernicus Publications.

Best practice following the Joint Declaration of Data Citation Principles initiated by FORCE 11:

COPDESS

In addition to promoting these data citation principles, Copernicus Publications is a signatory of the Coalition on Publishing Data in the Earth and Space Sciences (COPDESS) commitment statement and the Enabling FAIR Data Commitment Statement in the Earth, Space, and Environmental Sciences.

Statement on the availability of underlying data

Acknowledgements

THE GORDON FOUNDATION

POLAR DATA CATALOGUE

CCIN

LAURIER
Inspiring Lives.

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

Any Questions?

Email: yasairam@uwaterloo.ca

References and Links to Repositories for Further Information

References

Blick, B. (2020, June 18). Academic Works: Your Institutional Repository: Benefits of Digital Repositories. Retrieved from https://qcc.libguides.com/c.php?g=30171&p=2019401

Brook, C. (2018, December 05). What is a Data Repository? Retrieved from https://digitalguardian.com/blog/what-data repository

Neuwirth, E., & Alter, A. (2018, December 13). 6 Features of an Ideal Research Repository [Web log post]. Retrieved from https://www.exlibrisgroup.com/blog/6 features-ideal-research-repository/

Repositories Explored

https://www.frdr-dfdr.ca/repo/

https://dataverse.scholarsportal.info/

https://gordonfoundation.ca/initiatives/datastream/

https://www.polardata.ca/

https://www.pangaea.de/

https://zenodo.org

Thank You

Discussion: How do you use repositories?

- Mentimeter Poll Code: **58 55 69 4** at **menti.com**
  - https://www.menti.com/2fna523my5
- Is there a specific preference for specialist data bases, as opposed to generalist, when applicable?
- How do you choose to deposit data? Would you like to have full control to this process or prefer having access to database administration?
- Do you have any suggestions as to how you would use data repositories more often?