SMART WATER MANAGEMENT SYSTEM

GUIDED BY
R. SIVASAKTHI

S. SNEHA
R. INDHIRA
M. DEVA
V. NIRESHKUMAR

DEVELOPING THE DATA SHARING PLATFORM

Data Sharing Platforms

It is highly recommended for researchers to share data associated with the findings and results of publications as this promotes the cluster's aim to further engage in open science. In addition to providing a data archival solution whenever appropriate, data sharing platforms have other benefits such as:

- Reduced overall cost of storage - the use most of these platforms are free of charge
- Backups are created by the platforms and data integrity is highly assured
- Meeting data management requirements, such as number of backups, backup location and mode, and long-term preservation
- Embargo features allows fresh data to be backed up and released publicly after publication

The use of APIs provided by these platforms can be used to automate the upload of large quantities of data to these platforms during the initial process of archival and standardised in-lab procedures can then be set up to make archival on such platforms a frequent activity.

Data sharing flowchart

Recommended Data Sharing Platforms

Platform	Cost	Total Storage Quotas	Dataset Size Limits	Curation	Shareability	Maintains File Hierarchy	Backed up
Teamshare	CAD 0.15 per gigabyte year	Total Storage Purchased	None	Internal	UBC only	Yes	Yes
Compute Canada	Free	On default allocation, 50 Gb per user and 1 TB per group per Compute Server. More available through Resource Allocation Competitions.	None	Internal	Compute Canada Users Only. Full access control possible with a Globus Subscription	Yes	Yes
Scholars Portal Dataverse	Free	None within reasonable usage	2.5 GB for binary files, 500 MB for tabular data formats	Internal	Full access control possible	Only if double zipped	Yes
Federated Research Data Repository	Free	None within reasonable usage	None within reasonable usage	External	Open	Yes	Yes
Open Science Framework	Free (with free and paid add-ons available)	None within reasonable usage	5.1 GB on OSF storage	Internal	Full access control possible	Must be done manually on OSF dashboard	Yes

- Borealis: The Canadian Dataverse Repository
  - Storage and Backup
    - Location
    - Size Limits
  - User Permissions
  - Additional Features
- Federated Research Data Repository (FRDR)
  - Storage and Backup
    - Location
    - Size limits
  - Changes
  - Curation
  - Data Sharing and Collaboration
  - Deposit and Download
    - Globus Connect Personal

        - Deposit
        - Download
- Sharing on Compute Canada
    - Internal Sharing
    - External Sharing
        - Quick Reference
            - Sharing a Single file
            - Sharing a folder
- Open Science Framework
    - OSF Institutions
    - Storage and Backup
        - Location
        - Size limits
    - Add-ons
        - Citation add-ons
        - Storage add-ons
    - Collaboration
        - Within a lab
        - Across labs
        - Public Access
    - Version Control
        - Registration
- Bespoke solutions
    - Compute Canada Cloud
    - EduCloud
- Canadian Open Neuroscience Platform (CONP)
    - CONP Portal
    - NeuroLibre

Other Data Sharing Platforms

Platform	Cost	Total Storage Quotas	Dataset Size Limits	Curation	Shareability	Maintains File Hierarchy	Backed up
Zenodo	Free	None	50 GB	Internal (Changes have to be requested)	Open	Yes	Yes
DataLad	Self-Hosted (Dependent on system configuration)	None	None	Internal	Open	Yes	Optional

- Zenodo
  - Location
  - Size Limits
  - Changes
  - Collaboration and Data Sharing
- DataLad
  - Installation: Linux

Discovering Open Data

The CONP portal promises to deliver a central repository aggregating metadata from multiple platforms, allowing researchers to discover open data. As the portal is still under development, we have compiled a list of resources that can be used for this purpose:

- Registry of Research Data Repositories - Discover open data repositories across the globe

- bioCADDIE - Biomedical and healthcare data discovery and indexing ecosystem
- Federated Research Data Repository - FRDR crawls through multiple other Canadian data repositories including dataverses, FRDR's own data and government open data (view a full list here), allowing users to search for and discover data from multiple sources
- CRCNS - Collaborative Research in Computational Neuroscience. Hosts data from projects funded by their funding program.
- Center for Open Neuroscience - open software frameworks, platforms, data and methodologies for neuroscience
- OpenNeuro - free and open platform for sharing MRI, MEG, EEG, iEEG, and ECoG data

Lists that have been compiled by other sources:

- Awesomedata - A list of open neuroscience data platforms
- Opening up: open access publishing, data sharing, and how they can influence your neuroscience career -

Highlights of Key Points Made by Individual Speakers

- Data-sharing platforms should be flexible to meet the needs of individual users; data governance models should incorporate input from a variety of stakeholders. (Li)
- A nonprofit organization that manages a data-sharing platform can help enable data sharing, but other stakeholders in the broader research ecosystem (e.g., funders, publishers, academia) hold the major levers to reward investigators for sharing data. (Li)
- Data-sharing platforms have already presented new opportunities for conducting studies, as well as direct collaborations between researchers and original data contributors. (Ross)
- Concerns about potential unintended consequences of data sharing have not materialized (e.g., participant privacy breaches, undermining of original study findings, commercial or litigious use of data). (Ross)
- The supply of data available for sharing is much greater than the demand for it, and there is a need to raise researcher awareness of the rich data resources that are available to them. (Li, Ross)
- The majority of data requesters are from academia (Rockhold, Ross) while much of the clinical trial data sharing thus far has been carried out by industry. (Rockhold)
- To foster open science and data-sharing academia needs to find a way to provide academic credit for researchers who share original data. (Rockhold)
- To conduct a study across multiple clinical trials that are housed in different data-sharing platforms, researchers often must submit multiple separate data requests, conduct parallel analyses in the different platforms, and then attempt to combine the results. (Shaunessy)
- Data-sharing platforms and companies should work together to identify solutions that allow studies to combine data across platforms. (Shaunessy)
- Data use agreement (DUA) negotiations can cause significant delays during the data request process. A single, harmonized master DUA could improve the data user experience by increasing the efficiency of the data request process. (Li, Ross, Shaunessy)

6 MIN READ

5 Key Components of a Data Sharing Platform

Picture of Lewis Wynne-Jones Lewis Wynne-Jones

Data Management Data Strategy Data Catalog Data Sharing

Increasingly, companies are focused on finding ways to connect to new and valuable sources of data in order to enhance their analytical capabilities, support data governance initiatives, and deliver more insight to their business units.

Due to the increased demand for new data sources, companies are also looking at their internal data differently. Organizations that have access to valuable datasets are starting to explore opportunities that will let them share and monetize their data outside the four walls of their business.

The data sharing landscape today

Data sharing has become top of mind for organizations reassessing their data strategy. The ability to distribute data effectively - both within an ecosystem and to external parties - represents a new and exciting value proposition. Data "monetization" (the strategy of extracting more value from data you own, whether by reselling it or using it more broadly within your organization) is becoming prominent because it represents the elusive return on investment that data divisions have historically struggled with. Recently, Venturebeat featured data sharing specifically as part of their **2021 Machine Learning, AI, and Data**

Google launched **Analytics Hub** in May 2021 as a platform for combining datasets and sharing data, dashboards, and machine learning models. Google also launched **Datashare**, developed more for financial services and based on Analytics Hub.

Databricks announced **Delta Sharing** on the same day as Google's release. Delta Sharing is an open source protocol for secure data sharing across organizations.

The changing needs for modern companies

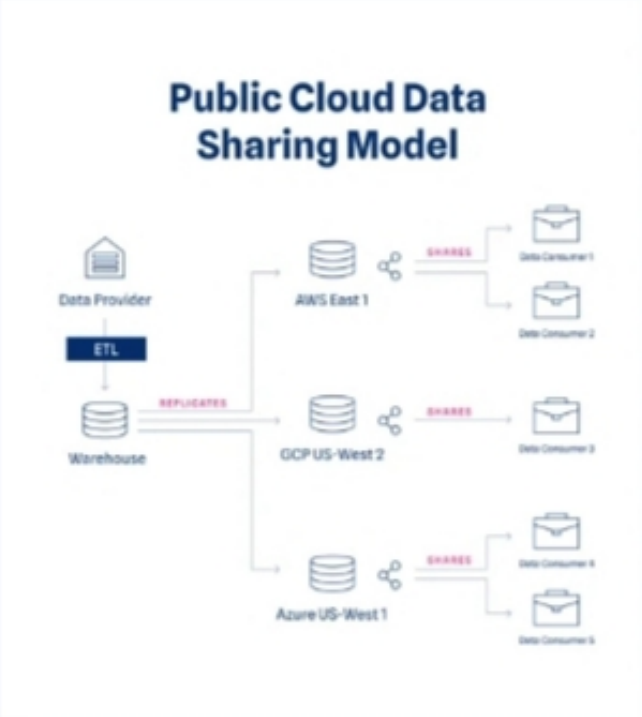
There are some things to consider when it comes to sharing data. Choosing who you will let use your data, and for what purposes, is an important question that organizations should think through from both a strategic and governance perspective. But outside of the rules about who will use your data and why there are also logistical problems involved with sharing data that are more technological than legal in nature.

Owner-defined access

Often, when organizations want to share the data they have, they have to slice and dice the datasets to remove sensitive information or create customer-facing copies of the dataset. Aside from the overhead of building a client-friendly version of your datasets, this is bad practice from a data governance perspective.

Consistent and configurable metadata

Most datasets aren't built to be consumer-facing. If the data is generated as a by-product of a core business process, it's probably particular to a very specific use case that may not be shared by the end user who wants to use the data.



Public Cloud Data Sharing Model

- data redundancy across multiple warehouses
- overhead in new tools, training, and time
- burdensome operational complexity

Private Cloud Data Sharing Model

- single source of truth for data
- sharing enabled on existing infrastructure
- simple and low-maintenance data access

THANK YOU