

Repository Options in Canada: A Portage Guide

JUNE 2019



www.carl-abrc.ca

portage

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



Table of Contents

Why publish my research data?	2
Where can I publish my research data?.....	3
National repository options.....	4
Dataverse North	5
Federated Research Data Repository (FRDR)	6
How do I decide which repository option to use?.....	7
Resources	10

Why publish my research data?

Publishing research data facilitates data reuse across and within disciplines. Some published data are open for sharing and reuse without restriction (e.g., under [Creative Commons](#) or another open data license); in other cases, it may be appropriate to impose restrictions on sharing and reuse.

Funders and journals increasingly require that data be published in a trustworthy repository as part of a well-developed data management plan (see [Draft Tri-Agency RDM Policy](#)).¹

Publishing your data is also a good way to ensure they remain accessible beyond the life of the study for which they were collected. The ability to find and re-use data is increasingly important for verifying published research findings and supporting new research. More information is available on Portage's [Research Data Management 101 Training Module](#).

¹ The Portage Network's [DMP Assistant](#) is an online, bilingual tool that guides researchers through this process.

Where can I publish my research data?

Canada has a rich data repository landscape operating at national, regional, and local/institutional levels. Data repositories are many and varied, but can be categorized broadly as domain-specific, government, and generalist.

How do I decide which repository type to use?

- Is there a domain-specific repository supported by your discipline?
 - If yes, this may be the best place to store your data as these are often purpose-built to serve specialized disciplinary needs.²
- Is your research being conducted on behalf of government, including federal or provincial agencies and departments, as well as municipalities?
 - If yes, then you could consider either an academic or domain-specific repository, or a government repository, if appropriate.
- If no domain-specific option is appropriate, then you should consider using a generalist repository that typically accepts any and all types of research data.
- There are many generalist repositories to choose from -- some of which operate on a fee-for-service model and others at no direct cost to researchers. This guide will walk you through two general repository options available to Canadian researchers at no cost through partnerships in the academic library community.

Before going further, a good first step is to contact your [local Research Data Management representative](#) to discuss your data publishing needs.

² See <https://www.re3data.org/> for a registry of data repositories.

National repository options

Portage supports two national repository options: the Federated Research Data Repository (FRDR) and Dataverse North. As a part of **Portage's Federated Architecture Model**, we have partnered with Compute Canada, OCUL Scholars Portal, and institutions offering Dataverse to provide hosted repository options that do not require installing, configuring, patching, or upgrading by individual institutions or researchers.

Development efforts are ongoing in terms of funding, features, and services. This document will be updated to reflect these developments as the landscape evolves.

With these repository options, you can:

- Publish your datasets, regardless of file type or format, for discovery and re-use
- Describe datasets using metadata fields appropriate for your discipline
- Immediately obtain a Digital Object Identifier (DOI) for publishing and citing data
- Set the licensing terms specifying how your datasets may be used
- Benefit from secure, Canadian-based repository storage
- Monitor your research impact by tracking use of your published datasets
- Comply with funder and journal requirements for data publishing
- Make your data openly discoverable and reusable to the world

Whether your data is stored in a Dataverse instance, FRDR, or a growing list of domain and government repositories, they are all readily discoverable through **Portage's National Discovery Service** (<http://frdr.ca>). This service harvests and indexes metadata records to make research data from across Canada available from a single web portal. This nationally federated approach encourages a rich repository landscape serving a diversity of needs, drives traffic to host repositories, and helps break down data and disciplinary silos.

Dataverse North

Dataverse North is a community of practice that brings together providers and users of Dataverse in Canada. Dataverse is an open source data repository platform for publishing research data, developed and maintained by Harvard University and contributors from around the world.

Scholars Portal, a service of the Ontario Council of University Libraries (OCUL), provides Dataverse to a [growing number of universities in Canada](#). Scholars Portal has partnered with Portage to provide Dataverse-as-a-service to collaborating Canadian institutions.

Other Dataverse providers may be found [here](#).

Main features:

- **Customizable:** Institutions and research groups can create brandable collections that keep their published research data together in one place.
- **Set policies locally:** Repository policies may be set at the institutional level and can range from mediated deposit to self-publishing models.
- **Data access control:** Researchers can choose the level of 'openness' for their data, and on a per file basis. You can change these settings at any time.
- **A good option in many cases:** Capacity for small-to-medium size data files (currently up to 2 GB) means that Dataverse is well-situated to serve the 'long tail' of data.

More information may be found in Portage's [Dataverse: The Essentials](#) and [Dataverse Primer](#).

Federated Research Data Repository (FRDR)

FRDR leverages Canada's national Digital Research Infrastructure, through a partnership between CARL Portage and Compute Canada, to offer robust, cost-effective repository storage at scale.

Main features:

- **Support for Large Datasets:** FRDR's underlying data transfer technology is purpose-built for large datasets.
- **Maintains file hierarchies:** FRDR allows you to retain the hierarchical relationship between files and folders during transfer.
- **Connect with national digital research infrastructure:** Globus File Transfer is widely used on Compute Canada systems, allowing researchers using those systems as a part of their active research data workflow an easy means of transferring data to FRDR for publication.
- **Large per-researcher allocation:** There is work underway to provide a 'general storage allocation' to all researchers in Canada (potentially up to 3 TB with no application or justification required beyond suitability for publication, curatorial review, and compliance with FRDR's Terms of Service) with mechanisms in place to request additional storage.

More information may be found in Portage's [FRDR: The Essentials](#).

How do I decide which repository option to use?

There is no wrong option, but researchers may find that a particular repository is more appropriate for their research or use case (see table below). It is not an either/or choice: you may choose to use Dataverse for your small-to-medium sized datasets, while taking advantage of FRDR's large data capabilities for your 'Big Data.'

Use Case Questions	Dataverse	FRDR
Does my institution have a data repository?	Contact your RDM representative to ask about local options. Check the Portage website to see if your institution is a part of Dataverse North.	Available to all eligible Canadian Principal Investigators and their designates. A good option if you do not have an institutional repository.
Will you need to update your dataset periodically?	Dataverse has robust versioning functionality that allows users to track and download older versions of your data. Updating datasets may be done manually and easily.	Updating datasets in FRDR may be done by contacting a curator. FRDR does not retain multiple versions of your data.
Do you have large data files (over 2GB), large numbers of files, or complex file hierarchies?	Dataverse currently can only handle files up to 2GB. Development efforts are underway to improve scalability and the ability to handle larger file sizes.	FRDR handles large data transfers (in terms of file size and numbers) easily and retains complex file hierarchies. FRDR's storage services are scalable and cost-effective for data-intensive research.
Do you need to control access to your datasets or files?*	Dataverse offers fine-grained control over data access, permissions may be set at the collection-level, dataset-level, or file-level.	In FRDR you can temporarily restrict access to datasets with an embargo, but all files must eventually be made openly available.

<p>What kinds of storage solutions are integrated?</p>	<p>Dataverse instances in Canada are hosted on local institutional storage servers. You can upload to Dataverse via certain U.S.-based cloud storage solutions, including Open Science Framework and Dropbox.</p>	<p>FRDR is connected to Compute Canada's national storage array, which includes remote storage options. Data may be transferred into FRDR from Compute Canada systems, local or institutional servers, or your personal computer or laptop.</p>
<p>Do you require a special collection for your datasets?</p>	<p>Dataverse allows for the creation of customized and branded containers (Dataverses) that store your datasets in a shared, browsable collection. Deposit and download permissions may be set at the collection-level, dataset-level, or file-level.</p>	<p>Research groups may request special collections for their datasets. Collection pages can include branding, information about the project, links to existing web portals, and a searchable list of published datasets. You control who may deposit in a collection.</p>
<p>Is it important to be able to preview or visualize your data?</p>	<p>Dataverse offers data exploration of tabular datasets, providing users with a full list of variables and summary statistics conveniently within a web browser.</p>	<p>In FRDR you can browse folder and file hierarchies before deciding if you want to download a dataset, but preview is limited to files that can be downloaded via HTTPS and viewed in your web browser.</p>
<p>Do you need to collaborate with others in a research team?</p>	<p>Dataverse has a number of features that will support collaborative research. You can track file versions, describe datasets and provide version notes, control access to files within a team easily, assign different access roles such as 'admin' 'contributor' or 'curator' at the Dataverse (collection) or dataset level.</p>	<p>FRDR is a repository for the publication of research data. It does not support active research data management functionality.</p>

<p>Does the system support dataset embargos? Is this automated?</p>	<p>Dataverse does not support automatic embargoing datasets at this time. You can keep a dataset unpublished to restrict access to a dataset and generate a Private URL for sharing if needed for journal or funder review purposes.</p>	<p>FRDR supports embargo periods, typically up to two years. You will receive a notification before your dataset is automatically released, giving you an opportunity to request an extension.</p>
<p>Do you require the use of APIs to integrate or automate aspects of the publishing workflow?</p>	<p>Dataverse has a robust set of APIs for searching, dataset retrieval, and upload. The Native API allows you to automate many aspects of the publishing workflow.</p>	<p>FRDR maintains a full set of APIs for dataset deposit and discovery. More information may be found in the documentation.</p>
<p>Will you need specialized or discipline specific metadata fields for describing your data?</p>	<p>Dataverse supports general citation and description fields, as well as, specialized metadata for the Social Sciences, Geosciences, Health and Life Sciences, and Astronomy.</p>	<p>By default, FRDR uses general standards for data description and discovery. You can request custom metadata fields and web forms specific to your discipline or use case.</p>
<p>What workflows are available to preserve my data for the long-term?</p>	<p>University libraries providing access to Dataverse may have integrated solutions and services available for digital preservation and data archiving. Contact your local RDM representative to inquiry about data archiving.</p>	<p>FRDR offers preservation processing via the creation of Archival Information Packages (AIPs) on datasets of any size. You can request that AIPs be transferred to you or a preservation service provider of your choosing for long-term archiving.</p>

* Neither repository option is suitable for sensitive research data at this time.

Resources

- Draft Tri-Agency RDM Policy: http://www.science.gc.ca/eic/site/063.nsf/eng/h_97610.html
- Federated Research Data Repository: <https://www.frdr.ca/>
- Dataverse: The Essentials: https://portagenetwork.ca/wp-content/uploads/2019/05/Dataverse_Essentials_EN.pdf
- Dataverse Primer: https://portagenetwork.ca/wp-content/uploads/2019/05/DVN_Primer_EN.pdf
- FRDR: The Essentials: https://portagenetwork.ca/wp-content/uploads/2019/05/FRDR_Essentials_EN.pdf
- Portage DMP Assistant: <https://assistant.portagenetwork.ca/>
- Portage Network: <https://portagenetwork.ca/>
- Registry of Research Data Repositories: <https://www.re3data.org/>
- Research Data Management 101 Training Module: <https://portagenetwork.ca/training-resources/portage-training-resources/rdm-101/>
- Scholars Portal Dataverse: <https://dataverse.scholarsportal.info/>