Canadian Dataverse Consortium:
A Year in Review

The Canadian Dataverse Consortium is a collaboration of the four regional academic library consortia in Canada (CAUL, BCI, OCUL and COPPUL) to develop a shared data management and publishing platform based on the Dataverse open source repository software. The CDC extends the original Scholars Portal Dataverse repository to university libraries across Canada.

In 2020-2021, the inaugural year of the Canadian Dataverse Consortium (CDC), the service grew to 58 institutions, with several new subscribers from CAUL and COPPUL. The University of Alberta Library (UAL) is a new subscriber this year, and staff have been working closely with the team in Edmonton to migrate data into the national instance.

As new institutions join the CDC, they become part of a growing community of practice that meets monthly. The February 2021 meeting drew over 50 attendees, representing a large and diverse group of RDM librarians and repository managers across Canada. Between March 2020-2021, the community listserv received over 550 messages, including questions and responses from institutional contacts, CDC staff, and some end-users[1].

This year, more than 1,100 researchers registered new accounts in the CDC. Deposits in the repository grew by over 1,000 datasets in a two-year period, standing now at 3,286 published datasets containing close to 70,000 data files . Collection sizes vary from institution to institution, with large research institutions contributing the majority of the larger collections combined with contributions from institutions of every size.

Service upgrades

The CDC platform was upgraded to release version 5.1.1 of Dataverse in December 2020. This upgrade included a new web application backend called Payara Server, adding to the speed, stability and functionality of the CDC installation.

Additional features related to the upgrade included the following:

- Redesigned dataset pages
- Improved responsiveness on small screens
- Improved user experience and performance when downloading many files at a time
- Ability to upload files without extensions
- Ability to upload files with the same name

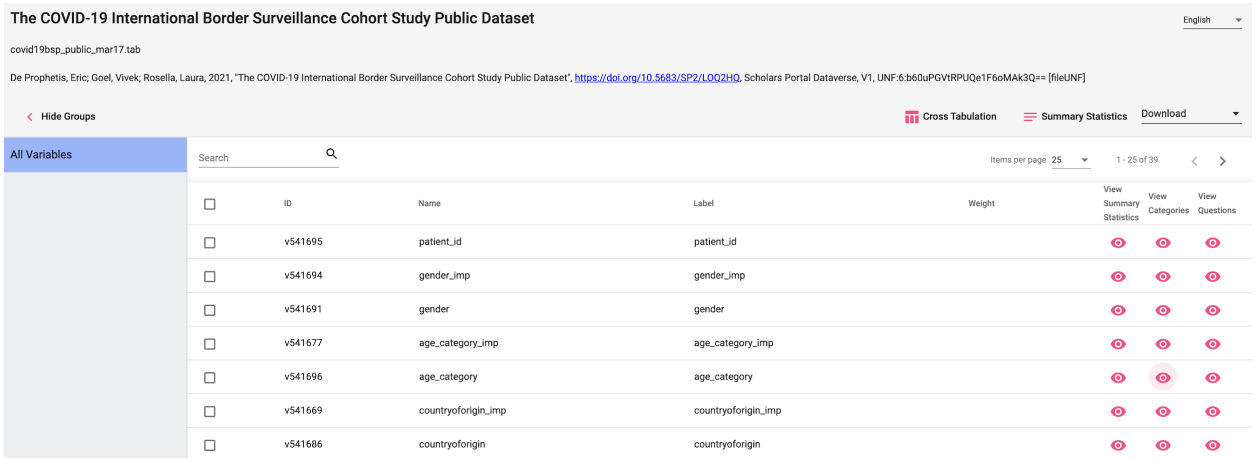
[1] End-user support is not captured centrally, each institution manages their own user support locally.

In 2020, CDC staff contributed an update to the Life Science/Biomedical Metadata Block on behalf of the Dataverse North Metadata WG based on recommendations provided to the IQSS team at Harvard University.

Furthermore, implementation of the Canadian Access Federation (CAF) Research and Scholarship (R&S) Entity Category at over 20 universities allows researchers to use their institutional computer credentials to access the CDC, simplifying and securing account management. (https://www.canarie.ca/identity/fim/research-and-scholarship-entity-category/)

CDC staff also developed and released new versions of the Data Curation and Data Explorer Tools this year. These open-source applications are now part of the core Dataverse code and are used to support enhanced DDI metadata markup, data visualization, and cross-tabulation analysis for tabular data in the repository.

Data Explorer 2.0



Current Projects for 2021-2022

CoreTrustSeal Pilot with Portage

CoreTrustSeal (CTS) is an internationally recognized standard for trustworthy certification of data repositories. Portage is sponsoring a project to encourage Canadian data repositories to secure certification under CTS. Unlike other trusted digital repository certification standards, CTS certification applies only to data collections curated by individual institutions rather than to aggregators of such collections, such as the CDC. The approach CDC staff are taking toward certification through this Portage program is to partner with institutional subscribers of the CDC looking to secure certification of their local data collections. CDC will provide support to these institutional subscribers by documenting our management/preservation practices and policies for the shared infrastructure, hosting cohort support meetings, and providing templates to support institutional CTS applications. In the future, once a

CoreTrustSeal certification process is developed for aggregators, CDC staff will have an opportunity to secure certification for the repository service as a whole if that is deemed worthwhile by CDC members.

Globus Integration

Globus is a file transfer application with roots in the supercomputing community that supports the secure transfer of large datasets beyond what is possible using normal web protocols such as HTTP. Globus is part of the current national computing infrastructure supported by Compute Canada and implemented in the Portage FRDR service. This year, CDC staff will release a new version of Dataverse with support for Globus. Researchers will be able to upload large files and datasets with thousands of individual files to Dataverse from any desktop computer or departmental server running the free Globus client software. Testing will begin in summer 2021. The integration of Globus and Dataverse is a development project of CDC staff in collaboration with IQSS at Harvard University and it will become part of the core Dataverse code in future releases.

Cloud Storage Migration

With funding support from Portage, the University of Toronto, on behalf of CDC, purchased 300TB of storage in the fall of 2020 to expand the capacity of the CDC repository and provide each institutional subscriber with a base amount of storage as part of each institution's annual subscription fee. The new CDC storage will be integrated with the Ontario Library Research Cloud (OLRC), a collaborative and geographically distributed cloud storage infrastructure hosted and managed by Scholars Portal. This will connect the CDC infrastructure with a storage service that can expand to meet member needs.

Dataverse Service Policies

Working in collaboration with the Portage/NDRIO Preservation Expert Group and Dataverse North, CDC staff are developing policies and guidelines to support greater understanding and transparency around the repository for institutional subscribers and researchers. These policies and guideline documents will include the following:

- Submissions Policy
- Deposit Guidelines
- Information Security Policy
- Preservation Plan
- Updated User Guide (including new content for institutional contacts)
- CoreTrustSeal certification requirements

Staff are also working with the CDC community to develop a template agreement that can be used by institutional subscribers to extend CDC to multi-institution research projects and Canadian academic journal publishers. This will allow a subscribing institution to sponsor such groups while providing those groups with an independent presence for branding purposes within the CDC.

Quarterly Roadmap - 2021-2023

This is a long-term roadmap for future Dataverse releases which provides context for planned CDC initiatives in a two-year context.

Timeline	Planned Projects (major effort)	Dataverse Release Roadmap
2021 Q1	**Planned Projects:** ● Service-level policy development (ongoing) ● Globus integration for large file support (ongoing) ● CoreTrustSeal Pilot Project (initiate) ● Data storage migration to library cloud (initiate) ● Preservation system development and workflows (initiate)	**Release 5.1.1 (current CDC DV version)** **Features/ release goals:** ● Data Explorer 2.0 release ● File Preview updates ● Data licensing display improvements
2021 Q2	**Planned Projects:** ● Service-level policy development (finalize) ● Data storage migration to library cloud (finalize) ● CoreTrustSeal Pilot Project (ongoing) ● Globus integration for large file support (ongoing) ● Preservation system development and workflows (ongoing) ● Accessibility improvements (initiate) ● Planning for future release 5.x in fall 2021 (initiate)	**Release 5.1.1 (current CDC DV version)** **Features/ release goals:** ● Service-level policies launch: ○ IT/Security ○ Data Deposit/Collections ○ Preservation ● Community testing for large file support (Globus) ● Data storage migration completed ● Metrics Report for institutions & Make Data Count
2021 Q3	**Planned Projects:** ● CDC Dataverse upgrade ● CoreTrustSeal Pilot Project (ongoing) ● Preservation system development and workflows (ongoing) ● Sensitive data support (initiate)	**Release 5.x TBD** **Features/ release goals:** ● CDC Dataverse Upgrade to 5.x TBD (update of core code features) ● Large file support (Globus integration) release ● Accessibility improvements
2021 Q4	**Planned Projects:** ● Sensitive data support (ongoing) ● Geospatial File Preview / Geodisy integration (initiate) ● Data Curation Tool improvements	**Release 5.x TBD** **Features/ release goals** ● CoreTrustSeal Pilot Project completed ● Preservation system release

	(initiate)	
2022 Q1	**Planned Projects:** ● Sensitive data support (ongoing) ● Geospatial File Preview / Geodisy integration (ongoing) ● Data Curation Tool improvements (ongoing) ● Discovery interface (initiate) ● Licensed/commercial data support (initiate)	**Release 5.x TBD** **Features/ release goals** ● Sensitive data support community testing ● Data Curation Tool update testing
2022 Q2	**Planned Projects:** ● Discovery interface integration (ongoing) ● Licensed/commercial data support (ongoing) ● Geospatial File Preview / Geodisy integration (ongoing) ● Planning for future release 5.x or 6.x in fall 2022 (initiate)	**Release 5.x TBD** **Features/ release goals** ● Sensitive data support release ● Data Curation Tool 2.0 release ● Geospatial File Preview testing
2022 Q3	**Planned Projects:** ● Discovery interface integration (ongoing) ● Licensed/commercial data support (ongoing)	**Release 5.x or 6.x TBD** **Features/ release goals** ● Upgrade to Dataverse 5.x or 6.x TBD ● Geospatial File Preview / Geodisy integration release ● Licensed/commercial data support testing
2022 Q4	**Planned Projects:** ● **TBD**	**Release 5.x or 6.x TBD** **Features/ release goals** ● Discovery interface integration release ● Licensed / commercial data support release