

Consortium Dataverse Canada :

Une année en revue

Le Consortium Dataverse Canada (CDC) est une collaboration des quatre consortiums régionaux de bibliothèques universitaires au Canada (CBUA, BCI, CBUO et COPPUL) pour développer une plateforme de gestion et de publication de données partagées basée sur Dataverse, un logiciel de dépôt ouvert. Le CDC ouvre l'accès au dépôt original du Dataverse de Scholars Portal aux autres bibliothèques universitaires au Canada.

En 2020-2021, année inaugurale du CDC, le service est passé à 58 établissements, avec plusieurs nouveaux abonnés du CBUA et de COPPUL. La bibliothèque de l'Université de l'Alberta (UAL) est un nouvel abonné cette année et le personnel du CDC travaille en étroite collaboration avec l'équipe d'Edmonton pour transférer les données vers l'instance nationale.

À mesure que de nouveaux établissements se joignent au CDC, elles entrent dans une communauté de pratique en croissance qui se réunit tous les mois. La réunion de février 2021 a attiré plus de 50 participants, représentant un groupe important et diversifié de bibliothécaires et de gestionnaires de dépôts de GDR à travers le Canada. De mars 2020 à mars 2021, la liste de diffusion communautaire a reçu plus de 550 messages, y compris des questions et réponses de répondants institutionnels, du personnel des CDC et de certains utilisateurs finaux¹.

Cette année, plus de 1 100 chercheurs ont créé de nouveaux comptes dans Dataverse. Les téléversements dans le dépôt ont augmenté de plus de 1 000 ensembles de données en deux ans, pour atteindre 3 286 ensembles de données publiés contenant près de 70 000 fichiers de données. La taille des collections varie d'un établissement à l'autre, les grandes universités de recherche contribuant à la majorité des plus grandes collections combinées avec des contributions d'établissements de toutes tailles.

Mise à niveau des services

La plateforme du CDC a été mise à niveau à la version 5.1.1 de Dataverse en décembre 2020. Cette mise à niveau comprenait une nouvelle application web dorsale appelée Payara Server, ajoutant à la vitesse, à la stabilité et aux fonctionnalités de l'installation du CDC.

Les nouvelles fonctionnalités offertes par la mise à niveau comprennent :

- Un nouveau design pour les pages d'ensembles de données
- Une amélioration de la réactivité sur les petits écrans
- Une amélioration de l'expérience d'utilisation et de la performance lors du téléchargement de fichiers en lot
- La capacité de téléverser des fichiers sans extension
- La capacité de téléverser des fichiers avec le même nom

¹ Le soutien aux usagers n'est pas offert de façon centralisée; chaque établissement gère son propre soutien aux usagers localement.

En 2020, le personnel du CDC a effectué une mise à jour du bloc de métadonnées des sciences de la vie et biomédicales pour le compte du groupe de travail sur les métadonnées de Dataverse Nord, sur la base des recommandations fournies à l'équipe IQSS de l'Université Harvard.

De plus, la mise en œuvre de la catégorie Research and Scholarship (R&S) de la Fédération canadienne d'accès (FAC) dans plus de 20 universités permet aux membres de la communauté recherche d'utiliser leur identifiant institutionnel pour accéder au CDC, simplifiant et sécurisant la gestion des comptes (<https://www.canarie.ca/fr/identite/gfi/categories-entites/>).

Le personnel du CDC a également développé et publié de nouvelles versions de l'Outil de curation et de l'Explorateur de données cette année. Ces applications ouvertes font désormais partie du code principal de Dataverse et sont utilisées pour prendre en charge le balisage de métadonnées DDI amélioré, la visualisation des données et l'analyse de tabulation croisée pour les données tabulaires dans le dépôt.

Explorateur de données 2.0

The screenshot shows the Dataverse 2.0 interface for the dataset "The COVID-19 International Border Surveillance Cohort Study Public Dataset". The interface includes a search bar, a list of variables, and options to view statistics or download data. The variables listed are:

ID	Nom	Libellé	Poids	Afficher les statistiques sommaires	Afficher les catégories	Afficher les questions
<input type="checkbox"/>	v541695	patient_id	patient_id	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/>	v541694	gender_imp	gender_imp	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/>	v541691	gender	gender	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/>	v541677	age_category_imp	age_category_imp	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/>	v541696	age_category	age_category	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/>	v541669	countryoforigin_imp	countryoforigin_imp	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/>	v541686	countryoforigin	countryoforigin	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Projets en cours pour 2021-2022

Pilote CoreTrustSeal avec Portage

[CoreTrustSeal](#) (CTS) est une norme internationalement reconnue pour la certification fiable des dépôts de données. Portage parraine un projet visant à encourager les dépôts de données canadiens à obtenir la certification CTS. Contrairement à d'autres normes de certification de dépôts numérique de confiance, la certification CTS s'applique uniquement aux collections de données gérées par des établissements individuels plutôt que par les agrégateurs de telles collections, comme le CDC. L'approche adoptée par le personnel du CDC en vue de la certification par le biais de ce programme Portage consiste à établir un partenariat avec les abonnés institutionnels du CDC qui cherchent à obtenir la certification de leurs collections de données locales. Le CDC fournira un soutien à ces abonnés institutionnels en documentant ses pratiques et politiques de gestion/préservation pour l'infrastructure partagée, en organisant des réunions de soutien pour la cohorte et en fournissant des modèles pour soutenir les applications CTS institutionnelles. À l'avenir, une fois qu'un processus de certification CoreTrustSeal sera développé pour les agrégateurs, le personnel du CDC aura la possibilité d'obtenir une certification pour le service de dépôt dans son ensemble si cela est jugé approprié par les membres du CDC.

Intégration Globus

[Globus](#) est une application de transfert de fichiers avec des racines dans la communauté des supercalculateurs qui prend en charge le transfert sécurisé de grands ensembles de données au-delà de ce qui est possible en utilisant des protocoles web normaux tels que HTTP. Globus fait partie de l'infrastructure informatique nationale actuelle soutenue par Calcul Canada et implémentée dans le service DFDR de Portage. Cette année, le personnel du CDC publiera une nouvelle version de Dataverse avec prise en charge de Globus. Les membres de la communauté de la recherche pourront télécharger des fichiers volumineux et des ensembles de données contenant des milliers de fichiers individuels vers Dataverse à partir de n'importe quel ordinateur de bureau ou serveur départemental exécutant le logiciel client gratuit Globus. Les tests commenceront à l'été 2021. L'intégration de Globus et Dataverse est un projet de développement du personnel du CDC en collaboration avec l'IQSS de l'Université de Harvard et fera partie du code de base de Dataverse dans les prochaines versions.

Migration au stockage en nuage

Avec le soutien financier de Portage, l'Université de Toronto, au nom de CDC, a acheté 300 To de stockage à l'automne 2020 pour augmenter la capacité du dépôt CDC et fournir à chaque abonné institutionnel une quantité de stockage de base dans les frais d'abonnement. Le nouveau stockage CDC sera intégré au Ontario Library Research Cloud (OLRC), une infrastructure de stockage en nuage collaborative et géographiquement distribuée hébergée et gérée par Scholars Portal. Cela connectera l'infrastructure CDC à un service de stockage qui peut s'étendre pour répondre aux besoins des membres.

Politiques de service Dataverse

En collaboration avec le groupe d'experts sur la préservation de Portage/NOIRN et Dataverse Nord, le personnel du CDC élabore des politiques et des lignes directrices pour favoriser une meilleure compréhension et une plus grande transparence autour du dépôt pour les abonnés institutionnels et les membres de la communauté de recherche. Ces politiques et documents d'orientation comprendront les éléments suivants :

- Une politique de dépôt
- Des directives pour le dépôt
- Une politique de sécurité de l'information
- Un plan de préservation
- Un guide d'utilisateur mis à jour (y compris le nouveau contenu pour les répondants institutionnels)
- Les exigences de certification CoreTrustSeal

Le personnel travaille également avec la communauté du CDC pour élaborer un modèle d'entente qui peut être utilisé par les abonnés institutionnels pour étendre le CDC à des projets de recherche multi-institutions et aux éditeurs de revues universitaires canadiennes. Cela permettra à un établissement abonné de parrainer de tels groupes tout en fournissant à ces groupes une présence indépendante à des fins de marque au sein du CDC.

Feuille de route trimestrielle – 2021-2023

Il s'agit d'une feuille de route à long terme pour les futures versions de Dataverse qui fournit le contexte pour les initiatives du CDC prévues au cours des deux prochaines années.

Terme	Projets prévus (jalons importants)	Feuille de route des versions de Dataverse
2021 T1	Projets planifiés : <ul style="list-style-type: none"> • Développement de politiques de niveaux de service (en continu) • Intégration Globus pour le soutien aux fichiers volumineux (en continu) • Projet pilote CoreTrustSeal (amorcé) • Migration du stockage des données en infonuagique (amorcé) • Développement et processus du système de préservation (amorcé) 	Version 5.1.1 (version actuelle de DV CDC) Fonctions/objectifs de la version : <ul style="list-style-type: none"> • Explorateur de données 2.0 publié • Mises à jour du prévisualisateur de fichiers • Améliorations de l'affichage de la licence des données
2021 T2	Projets planifiés : <ul style="list-style-type: none"> • Développement de politiques de niveaux de service (fin) • Migration du stockage des données en infonuagique (fin) • Projet pilote CoreTrustSeal (en continu) • Intégration Globus pour le soutien aux fichiers volumineux (en continu) • Développement et processus du système de préservation (en continu) • Améliorations de l'accessibilité (amorcé) • Lancement prévu de la version 5.x à l'automne 2021 (amorcé) 	Version 5.1.1 (version actuelle de DV CDC) Fonctions/objectifs de la version : <ul style="list-style-type: none"> • Publication des politiques de niveaux de service : <ul style="list-style-type: none"> ○ TI/sécurité ○ Dépôt de données/Collections ○ Préservation • Test par la communauté du soutien aux fichiers volumineux (Globus) • Fin de la migration du stockage des données • Rapport de métriques d'utilisation pour les établissements et projet Make Data Count
2021 T3	Projets planifiés : <ul style="list-style-type: none"> • Mise à niveau de Dataverse du CDC • Projet pilote CoreTrustSeal (en continu) • Développement et processus du système de préservation (en continu) • Soutien aux données sensibles (amorcé) 	Version 5.x à déterminer Fonctions/objectifs de la version : <ul style="list-style-type: none"> • Mise à niveau de Dataverse du CDC 5.x à déterminer (mise à jour des fonctions du code de base) • Soutien aux fichiers volumineux (intégration Globus) • Améliorations de l'accessibilité
2021 T4	Projets planifiés : <ul style="list-style-type: none"> • Soutien aux données sensibles (en continu) • Prévisualisation des fichiers géospatiaux/intégration de Geodisy (amorcé) 	Version 5.x à déterminer Fonctions/objectifs de la version : <ul style="list-style-type: none"> • Fin du projet pilote CoreTrustSeal • Lancement du système de préservation

	<ul style="list-style-type: none"> Améliorations de l'outil de curation des données (amorcé) 	
2022 T1	Projets planifiés : <ul style="list-style-type: none"> Soutien aux données sensibles (en continu) Prévisualisation des fichiers géospatiaux/intégration de Geodisy (en continu) Améliorations de l'outil de curation de données (en continu) Interface de découverte (amorcé) Soutien pour les données sous licence/commerciales (amorcé) 	Version 5.x à déterminer Fonctions/objectifs de la version : <ul style="list-style-type: none"> Test par la communauté du soutien aux données sensibles Test de la mise à jour de l'outil de curation de données
2022 T2	Projets planifiés : <ul style="list-style-type: none"> Intégration de l'interface de découverte (en continu) Soutien pour les données sous licence/commerciales (en continu) Prévisualisation des fichiers géospatiaux/intégration de Geodisy (en continu) Planification de la mise à niveau de 5.x ou 6.x à l'automne 2022 (amorcé) 	Version 5.x à déterminer Fonctions/objectifs de la version : <ul style="list-style-type: none"> Lancement du soutien aux données sensibles Lancement de l'outil de curation de données 2.0 Test de la prévisualisation des fichiers géospatiaux
2022 T3	Projets planifiés : <ul style="list-style-type: none"> Intégration de l'interface de découverte (en continu) Soutien pour les données sous licence/commerciales (en continu) 	Version 5.x ou 6.x à déterminer Fonctions/objectifs de la version : <ul style="list-style-type: none"> Mise à niveau de Dataverse 5.x ou 6.x Lancement de la prévisualisation des fichiers géospatiaux/intégration de Geodisy Test du soutien pour les données sous licence/commerciales
2022 T4	Projets planifiés : <ul style="list-style-type: none"> À déterminer 	Version 5.x ou 6.x à déterminer Fonctions/objectifs de la version : <ul style="list-style-type: none"> Lancement de l'intégration de l'interface de découverte Lancement du soutien pour les données sous licence/commerciales