Groupe d'experts sur la découverte des données – Groupe de travail sur le développement des collections

Rapport de la phase 1

Préparé par le Groupe de travail sur le développement des collections du Groupe d'experts sur la découverte des données du réseau Portage au nom de l'Association des bibliothèques de recherche du Canada (ABRC)

Berenica Vejvoda (McGill University, présidente)
Alison Ambi (Memorial University of Newfoundland)
Eugene Barsky (University of British Columbia)
Kevin Lindstrom (University of British Columbia)
Heather MacDonald (Carleton University)
Kathleen Matthews (University of Victoria)
Michael Moosberger (Dalhousie University)
Lisa O'Hara (University of Manitoba)
Susan Powelson (University of Calgary)
Kimberly Silk (Canadian Research Knowledge Network)
Allison Sivak (University of Alberta)
Kristi Thompson (University of Windsor)

JUILLET 2017

Réseau Portage
Association des bibliothèques de recherche du Canada
portage@carl-abrc.ca

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

CARL ABRC
CANADIAN ASSOCIATION OF RESEARCH LIBRARIES
ASSOCIATION DES BIBLIOTHÈQUES DE RECHERCHE DU CANADA

www.portagenetwork.ca

Table des matières

Contexte

Ce groupe de travail est l'un des deux groupes créés pour épauler le Groupe d'experts sur la découverte des données du réseau Portage. L'objectif du Groupe de travail sur le développement des collections consiste à veiller à ce que les données de recherche canadiennes soient complètement incluses et indexées dans le Dépôt fédéré de données de recherche (DFDR) et dans d'autres outils de repérage afin de favoriser leur découverte et leur réutilisation[1].

Définition de données de recherche

Pour les besoins du présent projet, le Groupe de travail a élaboré la définition de données de recherche suivante :

> *[TRADUCTION] Données employées comme sources principales de soutien aux enquêtes techniques ou scientifiques, à la recherche, aux études et aux activités artistiques; elles servent comme preuve dans le processus de recherche et/ou sont généralement considérées comme nécessaires pour valider les constatations et les résultats des recherches[2].*

Cette définition est fondée sur celle proposée dans le CASRAI Dictionary[3]. La définition du CASRAI Dictionary a été modifiée afin d'en amoindrir la portée et d'éliminer l'inclusion de la notion de contenu numérique et non numérique qui était simplement [TRADUCTION] « susceptible de devenir des données de recherche ».

---

[1] Mandat : https://portagenetwork.ca/wp-content/uploads/2017/04/GEDD-CollectionsGT-Mandat-FR.pdf
[2] CASRAI Dictionary, « Research Data (Données de recherche) », http://dictionary.casrai.org/Research_data (en anglais).
[3] Ibid.

Type de dépôts de données de recherche

En plus de la définition de données de recherche, une typologie[4] des dépôts de données de recherche a été établie pour faciliter le recensement de possibles dépôts qui pourraient être inclus dans le DFDR. La liste ci-après est établie selon le degré d'investissement en matière de conservation, de temps et d'archive, qui progresse du degré minimal au degré maximal d'investissement.

1. **Des dépôts intermédiaires** qui sont utilisés avant le transfert des données vers un autre dépôt et qui soutiennent souvent un projet de recherche actif, comme Globus (p. ex., DSpace utilisé à titre de dépôt intermédiaire jusqu'à ce qu'une meilleure option soit trouvée).
2. **Des dépôts de données de recherche** conçus à partir d'un dépôt institutionnel qui sert principalement aux publications (p. ex., des ensembles de données de recherche provenant de dépôts institutionnels, comme DSpace, qui sont clairement désignés comme des données – dc.type:Dataset –[5]).
3. **Des dépôts de données de recherche qui fournissent principalement un accès** aux données de recherche sans nécessairement archiver les données.
4. **Des dépôts numériques génériques** appartenant à des éditeurs, comme Dryad ou Figshare, ou conservant des collections numériques spécialisées, comme Canadiana.
5. **Des dépôts d'ensembles de données liés à des publications** qui soutiennent la reproduction des résultats de recherche dans les publications (p. ex., Mendeley Data).
6. **Des dépôts de domaine** qui conservent des données de recherche qui sont échangées dans un domaine précis (p. ex., Centre canadien de données astronomiques).

---

[4] La typologie des dépôts de données de recherche a été fournie par Chuck Humphrey, directeur de Portage.
[5] http://dublincore.org/documents/2008/01/14/dcmi-type-vocabulary/ (en anglais)

7. **Des dépôts de données de recherche offrant accès aux données de recherche et permettant leur archivage** (p. ex. Scholars Portal Dataverse ou UBC Abacus Dataverse Network).

Collection de dépôts et critères de développement

Le dépôt Re3.data[6] a servi de source initiale pour les dépôts de données de recherche, puisqu'il donne accès à l'une des listes les plus exhaustives. Des dépôts enregistrés auprès de DataCite Canada[7] ont également été inclus. Une recherche ciblée pour les dépôts de données de recherche en sciences humaines a été effectuée pour compléter la liste. Sur la base des commentaires des membres du Groupe d'experts sur la découverte des données, la liste a été limitée aux dépôts de données de recherche hébergés au Canada. La liste initiale résultante se composait d'environ 170 dépôts de données de recherche.

Une liste de critères, qui a été précisée selon un consensus établi en discussion de groupe, a été élaborée d'après les filtres de Re3.data. La liste de critères comprend ce qui suit :

- le type de dépôt de données de recherche;
- le fait de savoir si le dépôt est exploité par le gouvernement, une université ou une autre entité;
- le fait de savoir si les coordonnées du service de soutien technique sont facilement présentées;
- le fait de savoir si le dépôt de données de recherche a mis en œuvre un modèle de métadonnées normalisé et observable;
- le fait de savoir si une API (interface de programmation d'applications) pour recueillir les métadonnées est observable;

[6] http://www.re3data.org/ (en anglais).
[7] https://search.datacite.org/ (en anglais).

- le fait de savoir si le dépôt utilise des identifiants permanents comme les systèmes Handle ou DOI (identifiant d'objet numérique);
- le fait de savoir si un filtre limitant les données était à la disposition des dépôts qui comportaient d'autres documents.

Recensement de dépôts de données de recherche pilotes

Avec l'application de ces critères à la liste des dépôts de données de recherche, deux critères se sont révélés particulièrement pertinents, soit le type de dépôt de données de recherche et le modèle de métadonnées. Des difficultés à évaluer certains de ces critères à partir du site Web destiné au public ont été éprouvées, p. ex. API visible ou filtre de données disponible. Parmi les critères qui n'ont pas été déterminés a priori, mais qui ont émergé naturellement et qui ont aussi guidé l'analyse, citons le fait de savoir si le dépôt de données était toujours actif, le caractère récent des mises à jour du site Web du dépôt, et le fait de savoir s'il était effectivement possible ou non de télécharger les ensembles de données. Les dépôts de données qui semblaient exiger un mot de passe pour y accéder ont été exclus.

Les dix candidats ci-après ont été recensés.

Dépôts de données de recherche	Coordonnées
Canadian Opinion Research Archive (http://www.queensu.ca/cora) (en anglais)	Canadian Opinion Research Archive, School of Policy Studies Queen's University, Kingston (Ontario) K7L 3N6 Courriel : cora@queensu.ca
Ocean Networks Canada (http://www.oceannetworks.ca) (en anglais)	Ocean Networks Canada University of Victoria C.P. 1700 STN CSC Victoria (C.-B.) V8W 2Y2 Téléphone : 250-472-5400 Courriel : info@oceannetworks.ca

Polar Data Catalogue (https://www.polardata.ca) (en anglais)	Gestionnaire de données Gabrielle Alix Téléphone : 519-888-4567 x 37572 Courriel : galix@uwaterloo.ca
Toutes les instances de Dataverse dans les universités canadiennes	Dalhousie University (en cours de conception, aucun lien pour l'instant) Ontario Council of University Libraries Courriel : dataverse@scholarsportal.info University of Alberta Libraries' Dataverse Network (https://dataverse.library.ualberta.ca) (en anglais) University of Manitoba (en cours de conception, aucun lien pour l'instant) University of British Columbia, Simon Fraser University, University of Victoria and University of Northern British Columbia (http://dvn.library.ubc.ca/dvn) (en anglais) Personne-ressource : eugene.barsky@ubc.ca
Mouse Atlas of Gene Expression (http://www.mouseatlas.org/mouseatlas_index_html) (en anglais)	BC Cancer Agency / Michael Smith Genome Sciences Centre Téléphone : 604 707-5900
Centre mondial de données sur l'ozone et le rayonnement ultraviolet (http://woudc.org)	Service météorologique du Canada Environnement et Changement climatique Canada 4905, rue Dufferin Toronto (Ontario) M3H 5T4 Formulaire électronique : http://woudc.org/contact.php?lang=fr

Ocean Tracking Network (OTN) (http://oceantrackingnetwork.org) (en anglais)	Lenore Bajona Director of Data Management Téléphone : (902) 494-7893 Courriel : lenore.bajona@dal.ca
Hakai Institute (https://www.hakai.org) (en anglais)	Liste du personnel : https://www.hakai.org/people/hakai-staff
BC Conservation Data Centre (http://www2.gov.bc.ca/gov/content/environment/plants-animals-ecosystems/conservation-data-centre) (en anglais)	Gouvernement de la Colombie-Britannique Téléphone : 250-356-0928 Courriel : cdcdata@gov.bc.ca

Limitations et problèmes constatés

- Initialement, il s'est avéré difficile de définir et de comprendre clairement la portée du présent groupe de travail et les attentes à son endroit.
- De nombreux portails n'hébergent pas de données de recherche : bon nombre de portails enregistrés auprès de Re3data.org semblaient être des portails de recherche hébergés par des chercheurs, mais ne sont pas des dépôts de données en soi.
- De différents types de données dans un seul dépôt avec divers modèles de métadonnées ont été rencontrés.
- Il s'est avéré difficile d'isoler les ensembles de données des non-ensembles de données (à cause de l'absence d'un champ de métadonnées distinct qui permettrait de filtrer le type d'ensemble de données de recherche).
- Il s'est avéré difficile de trouver des dépôts de données de recherche en sciences humaines (par rapport à des dépôts comportant des publications en sciences humaines); cette difficulté pourrait être attribuable au fait que les données en sciences humaines sont textuelles plutôt que numériques, ce qui pourrait nécessiter une approche différente.

- Des collections n'ayant pas de modèles de métadonnées standardisés énoncés expressément ont été rencontrées.
- Diverses méthodes d'accès contrôlés ont été rencontrées.

Recommandations

Le Groupe de travail sur le développement des collections recommande :

- que la liste de critères ci-après soit utilisée pour déterminer les dépôts de données de recherche qui devraient être inclus dans le DFDR :
    - le type de dépôt de données de recherche;
    - le fait de savoir si le dépôt est exploité par le gouvernement, une université ou une autre entité;
    - le fait de savoir si les coordonnées du service de soutien technique sont aisément présentées;
    - le fait de savoir si le dépôt de données de recherche a mis en œuvre un modèle de métadonnées normalisé et observable;
    - le fait de savoir si une API (interface de programmation d'applications) pour recueillir les métadonnées est observable;
    - le fait de savoir si le dépôt utilise des identifiants permanents comme les systèmes Handle ou DOI (identifiant d'objet numérique);
    - le fait de savoir si un filtre limitant les données était à la disposition des dépôts qui comportaient d'autres matériels;
    - les travaux ultérieurs devraient porter sur la redondance des données, les multiples DOI, et la duplication des métadonnées;

- que, conformément à la définition de données de recherche :
    - les données qui sont « susceptibles de devenir » des données de recherche soient exclues puisque presque tout a le potentiel de devenir des données.

Prochaines étapes

Le Groupe de travail sur le développement des collections recommande les prochaines étapes suivantes :

- poursuivre l'examen des dépôts de données de recherche qui sont hébergés à l'extérieur du Canada, mais qui comportent un important contenu canadien (p. ex., Dryad et PANGAEA);
- revoir l'inclusion des dépôts de données de recherche hébergés par le gouvernement canadien qui ont été exclus dans la première phase afin de gérer la portée;
- entrer en contact avec les dépôts sélectionnés afin d'approfondir et de vérifier les critères obtenus à partir de sites Web destinés au public;
- entrer en contact avec les dépôts sélectionnés afin de déterminer et de confirmer expressément la capacité de recueillir les métadonnées et de limiter la cueillette au contenu des données de recherche;
- peaufiner et mettre au point les critères selon les résultats des contacts avec les dépôts pilotes, avant de procéder au recensement d'autres dépôts.