# La gestion des données de recherche dans le contexte canadien

# LA GESTION DES DONNÉES DE RECHERCHE DANS LE CONTEXTE CANADIEN

## Un guide pour la pratique et l'apprentissage

SOUS LA DIRECTION DE KRISTI THOMPSON; ELIZABETH HILL; EMILY CARLISLE-JOHNSTON; DANIELLE DENNIE; ET ÉMILIE FORTIN

JENNIFER ABEL; EUGENE BARSKY; MARTIN CHANDLER; LUCIA COSTANZO; DYLANNE DEARBORN; ALISON FARRELL; ÉMILIE FORTIN; JANE FRY; LOUISE GILLIS; MEGHAN GOODCHILD; KARA HANDREN; ELIZABETH HILL; GRANT HURLEY; SHAHIRA KHAIR; DANI KWAN-LAFOND; OLIVER LAPOINTE; AMBER LEAHEY; CYNTHIA LISÉE; DR. RONG LUO; LACHLAN MACLEOD; STEVE MARKS; DR. JOEL T. MINION; JEFF MOON; KAITLIN NEWSON; MIKAYLA REDDEN; CHANTAL RIPP; ÉDITH ROBERT; DR. ALISA BETH ROD; DANY SAVARD; SANDRA SAWCHUK; FELICITY TAYLER; KRISTI THOMPSON; BERENICA VEJVODA; MINGLU WANG; LEE WILSON; TATIANA ZARAISKAYA; ET DR. BIRU ZHOU

Western University, Western Libraries
London, ON

# TABLE DES MATIÈRES

## Partie I. Point de départ en gestion des données de recherche

## Partie III. Méthodes de travail avec les données de recherche

## Partie IV. Types de données de recherche

## Partie V. Perspectives sur la gestion des données de recherche

# COMMENT UTILISER CE MANUEL

## Comment naviguer dans ce manuel

### La table des matières : accès aux sections et aux chapitres

Dans le coin supérieur gauche de l'écran se trouve un onglet noir intitulé « Table des matières ». En cliquant sur l'onglet, un menu déroulant s'ouvre et affiche la table des matières permettant ainsi la navigation vers n'importe quelle section ou chapitre du livre.

En cliquant sur le bouton plus (+) à droite d'une section, vous pourrez l'ouvrir et afficher le titre de chaque chapitre. Ces titres sont cliquables et vous permettent d'accéder directement au chapitre.

### Boutons « précédent » et « suivant »

En bas à gauche ou à droite de chaque page dans Pressbooks (y compris celle-ci !) se trouvent les boutons « précédent » et « suivant. » Ils sont étiquetés avec le titre du chapitre précédent ou suivant. Vous pouvez utiliser ces boutons pour aller directement au chapitre précédent ou suivant sans revenir à la table des matières.

### Glossaire

À la fin de l'ouvrage, vous trouverez un glossaire. S'il y a lieu, les définitions du glossaire ont également été intégrées directement dans les chapitres. Lorsque vous apercevez un terme souligné, cliquez dessus et sa définition va apparaître dans une infobulle.

## Pourquoi un manuel ouvert ?

Avec la récente publication de la Politique des trois organismes sur la gestion des données de recherche, la GDR est devenue d'une importance cruciale. L'ensemble des chercheuses et chercheurs qui demandent des subventions pour financer leur recherche et qui produisent des données doivent désormais satisfaire à des exigences telles que la rédaction de plans de gestion des données et la préparation des données pour

l'archivage. Compte tenu de l'attention accrue portée à la GDR, les besoins éducationnels ainsi que le nombre de cours liés à la GDR sont susceptibles d'augmenter.

Au cours de l'été 2021, plusieurs universitaires et bibliothécaires du Canada, y compris des personnes qui donnent des cours universitaires sur la GDR, ont formé un groupe pour discuter de la création d'un manuel bilingue conçu au Canada. Le groupe a reconnu qu'à l'époque, il n'existait pas de ressources adaptées au contexte réglementaire unique du Canada et appropriées pour une utilisation en classe. Ensemble, les membres du groupe ont décidé qu'une ressource éducative libre (REL) sous la forme d'un manuel serait utile pour la pratique et l'apprentissage des parties prenantes canadiennes, et qu'elle refléterait l'esprit de la GDR en encourageant l'ouverture.

## Qu'est-ce qu'un manuel ouvert ?

Un manuel ouvert est une ressource en ligne accessible au public, gratuite et dotée d'une licence ouverte qui permet à d'autres de réutiliser la ressource, de la conserver, de l'adapter, de la redistribuer et de la réviser. Ce livre a une licence Creative Commons Attribution-Pas d'Utilisation Commerciale (https://creativecommons.org/licenses/by-nc/4.0/deed.fr) (CC BY-NC), qui permet son adaptation et sa redistribution à des fins non commerciales, à condition que la personne créatrice originale soit citée (voir la section « Licence et attribution » ). En plus de la licence ouverte, les autrices et auteurs s'engagent à rendre ce manuel ouvert disponible immédiatement, gratuitement et en permanence à toute personne ayant accès à l'Internet.

Les avantages à l'utilisation de manuels ouverts sont nombreux. Outre le fait qu'ils permettent aux élèves et au corps professoral d'accéder librement à des ressources éducatives de qualité, ce qui représente une économie considérable, les ressources ouvertes garantissent également que les buts éducatifs soient pris en compte. L'objectif de l'ODD 4 de l'UNESCO (https://www.unesco.org/fr/education2030-sdg4) qui vise à « garantir une éducation inclusive et de qualité pour [toutes et] tous et promouvoir l'apprentissage tout au long de la vie » d'ici à 2030 commence par des ressources éducatives libres (REL) accessibles à l'ensemble de la population. La vision précédente selon laquelle l'éducation consiste à diffuser des connaissances a été remise en question par les personnes qui défendent les REL et qui mènent la réforme de l'éducation vers la co-création et le partage des connaissances (Blomgren et Henderson, 2021 ; Cronin, 2017 ; Henderson et Ostashewski, 2018). Outre l'utilisation gratuite d'un manuel ouvert, les ressources ouvertes utilisées pour l'enseignement sont directement en lien avec les objectifs d'un programme d'études et peuvent rester pertinentes dans le domaine grâce à leur adaptation et à leur révision (Hendricks *et al.*, 2017). De plus, elles réduisent la confusion souvent associée à la sélection de ressources avec des licences plus restrictives (Henderson *et al.*, 2018).

Bien qu'il existe de nombreuses maisons d'édition commerciales qui proposent des manuels de qualité

similaire, ceux-ci ont des limites qui réduisent l'impact qu'ils pourraient avoir. En particulier, ils sont rarement permanents ou offerts gratuitement, ce qui limite l'accessibilité de ces ressources à de nombreux élèves, au corps professoral, ainsi qu'aux praticiennes et praticiens. Ce manuel ouvert, intitulé *La gestion des données de recherche dans le contexte canadien : un guide pour la pratique et l'apprentissage*, répond à cet appel à la réforme en éducation en produisant une ressource éducative libre de façon immédiate, gratuite et permanente selon la voie dorée du libre accès. Cette ressource peut être révisée, redistribuée, conservée, adaptée et réutilisée à des fins non commerciales en vertu d'une licence Creative Commons Attribution-Pas d'Utilisation Commerciale Creative Commons Attribution-Pas d'Utilisation Commerciale (https://creativecommons.org/licenses/by-nc/4.0/deed.fr).

Dans la section suivante, « Comment accéder à ce livre et l'utiliser, » nous aborderons les usages prévus de ce manuel.

## Comment accéder à ce livre et l'utiliser ?

Cet ouvrage devrait répondre aux besoins des professeures et professeurs à la recherche de ressources pour soutenir l'enseignement de sujets liés à la GDR, ainsi qu'aux besoins des bibliothécaires, de la population étudiante et des chercheuses et chercheurs qui réclament des documents à jour pour les guider dans leurs pratiques de GDR. En publiant *La gestion des données de recherche dans le contexte canadien : un guide pour la pratique et l'apprentissage* avec une license CC BY-NC, nous souhaitons que ce livre soit adopté dans son intégralité en tant que lecture obligatoire en classe, adapté en partie en tant qu'information complémentaire, ou révisé avec des informations actuelles ou importantes que le manuel pourrait ne pas contenir. Nous sommes ravies de proposer cette ressource éducative libre comme point de départ pour faire progresser le domaine de la GDR en collaboration et au bénéfice des personnes qui travaillent en GDR et nous espérons mettre en lumière le besoin de ressources supplémentaires dans ce domaine ainsi que dans d'autres.

## Licence et attribution



Voici quelques exemples de mentions d'attribution :

## Redistribution du livre au complet :

La gestion des données de recherche dans le contexte canadien : un guide pour la pratique et l'apprentissage créé par Kristi Thompson; Elizabeth Hill; Emily Carlisle-Johnston; Danielle Dennie; and Émilie Fortin publié chez Pressbooks. L'original est disponible gratuitement sous les termes de la licence CC BY-NC 4.0 (https://creativecommons.org/licenses/by-nc/4.0/deed.fr) à l'adresse https://ecampusontario.pressbooks.pub/gdrcanada (https://ecampusontario.pressbooks.pub/gdrcanada).

## Redistribution d'un ou des chapitres :

[Titre du chapitre], [Autrices ou Auteurs], dans La gestion des données de recherche dans le contexte canadien : un guide pour la pratique et l'apprentissage créé par Kristi Thompson ; Elizabeth Hill ; Emily Carlisle-Johnston ; Danielle Dennie ; et Émilie Fortin publié chez Pressbooks. L'original est offert gratuitement selon les termes de la licence CC BY-NC 4.0 (https://creativecommons.org/licenses/by-nc/4.0/deed.fr) à l'adresse https://ecampusontario.pressbooks.pub/gdrcanada (https://ecampusontario.pressbooks.pub/gdrcanada).

## Versions révisées ou adaptées :

Ce document a été adapté/révisé à partir du document La gestion des données de recherche dans le contexte canadien : un guide pour la pratique et l'apprentissage créé par Kristi Thompson ; Elizabeth Hill ; Emily Carlisle-Johnston ; Danielle Dennie ; et Émilie Fortin publié avec Pressbooks. L'original est disponible gratuitement selon les termes de la licence CC BY-NC 4.0 (https://creativecommons.org/licenses/by-nc/4.0/deed.fr) à l'adresse https://ecampusontario.pressbooks.pub/gdrcanada (https://ecampusontario.pressbooks.pub/gdrcanada).

Pour plus d'informations, voir le FAQ Creative Commons Attribution (https://creativecommons.org/faq/fr/#comment-puis-je-attribuer-correctement-un-mat%C3%A9riel-offert-sous-une-licence-creative-commons) et Creative Commons Best practices for attribution (https://wiki.creativecommons.org/wiki/Best_practices_for_attribution) (en anglais seulement).

# Contactez-nous!

Si vous aimez ce manuel et envisagez de l'utiliser, nous serions ravis de le savoir ! Veuillez nous envoyer un message pour nous indiquer comment vous l'utilisez en envoyant un courriel à rdmoerteam@gmail.com (mailto:rdmoerteam@gmail.com).

Vous trouverez la version anglaise de ce manuel à cette adresse : https://ecampusontario.pressbooks.pub/canadardm/ (https://ecampusontario.pressbooks.pub/canadardm/). Si vous souhaitez adapter ou traduire ce travail dans une autre langue, nous serions également ravis d'avoir de vos nouvelles et de répondre à vos questions.

# Bibliographie

Blomgren, C. et Henderson, S. (2021). Addressing the K-12 open educational resources awareness niche: A virtual conference response. Alberta *Journal of Educational Research, 67*(1), 68-82. https://doi.org/10.11575/ajer.v67i1.56965 (https://doi.org/10.11575/ajer.v67i1.56965)

Cronin, C. (2017). Openness and praxis: Exploring the use of open educational practices in higher education. *The International Review of Research in Open and Distributed Learning, 18*(5), 1-21. https://doi.org/10.19173/irrodl.v18i5.3096 (https://doi.org/10.19173/irrodl.v18i5.3096)

Henderson, S. et Ostashewski, N. (2018). Barriers, incentives, and benefits of the open educational resources (OER) movement: An exploration into instructor perspectives. *First Monday, 23*(12). https://doi.org/10.5210/fm.v23i12.9172 (https://doi.org/10.5210/fm.v23i12.9172)

Hendricks, C., Reinsberg, S. A. et Rieger, G. W. (2017). The adoption of an open textbook in a large physics course: An analysis of cost, outcomes, use, and perceptions. *The International Review of Research in Open and Distributed Learning, 18*(4), 78-99. https://doi.org/10.19173/irrodl.v18i4.3006 (https://doi.org/10.19173/irrodl.v18i4.3006)

# À PROPOS DES ÉDITRICES

Emily Carlisle-Johnston est bibliothécaire chargée de la recherche et de la communication savante à l'Université Western depuis 2020. Elle travaille avec les membres du corps professoral qui souhaitent intégrer des ressources éducatives libres (REL) dans leur enseignement, notamment en les aidant à trouver et à évaluer les REL, en évaluant les options de licence pour la réutilisation et l'adaptation des REL, et en soutenant l'utilisation de logiciels d'édition libre tels que Pressbooks. Auparavant, elle a travaillé à eCampusOntario, où elle a dirigé les processus de travail éditoriaux pour la création de REL. Emily a terminé le programme SPARC Open Education Leadership Fellow en 2022. ORCID : 0000-0002-5391-723X (https://orcid.org/0000-0002-5391-723X)

Danielle Dennie est directrice de la Bibliothèque Vanier à l'Université Concordia depuis 2021. Elle est également bibliothécaire des données de recherche à Concordia depuis 2018. Elle est titulaire d'une maîtrise en microbiologie appliquée de l'INRS-Institut Armand Frappier ainsi que d'une maîtrise en bibliothéconomie et en sciences de l'information de l'Université McGill. Elle était responsable de l'implantation de la stratégie institutionnelle de GDR à l'Université Concordia. ORCID: 0000-0003-3771-2450 (https://orcid.org/0000-0003-3771-2450)

Émilie Fortin est bibliothécaire à la gestion des données de recherche et à la préservation numérique à l'Université Laval depuis 2021. Auparavant, elle occupait le poste de responsable de la production numérique, préservation et conservation des collections. Elle a complété sa maîtrise en science de l'information de l'Université de Montréal en passant une année à la Haute école de gestion de Genève. Impliquée dans le Groupe d'experts de l'Alliance de recherche numérique sur la planification de la gestion des données, dans le groupe de travail sur la gestion des données de recherche du Bureau de coopération interuniversitaire (BCI), elle participe également régulièrement aux conférences de l'iPRES sur la préservation numérique. ORCID: 0000-0002-9717-6840 (https://orcid.org/0000-0002-9717-6840)

Elizabeth Hill est la bibliothécaire des données à l'Université Western. Elle y donne des formations en littératie des données ainsi qu'en accès aux sources de données. Elle agit à titre de conseillère externe auprès de Statistiques Canada. Mme Hill est active dans plusieurs communautés de données et groupes de travail et ce, tant à titre de participante que de cheffe. Ses domaines d'intérêt en recherche comprennent le soutien aux chercheuses et chercheurs. Elle a publié des ouvrages au sujet des systèmes de diffusion de données et de la bibliothéconomie des données au Canada. ORCID : 0000-0002-9715-238X (https://orcid.org/0000-0002-9715-238X)

Kristi Thompson est bibliothécaire en gestion des données de recherche à l'Université Western. Elle a

précédemment occupé les postes de bibliothécaire des données à l'Université de Windsor et de spécialiste des données à l'Université de Princeton. Elle détient un baccalauréat en informatique de l'Université Queen's et une maîtrise en science de l'information de l'Université Western. Kristi soutient des projets de recherche, administre des logiciels d'archivage de données, travaille avec les comités d'éthique de la recherche de l'Université Western et participe au niveau national au développement de l'infrastructure des données de recherche. Elle a coédité le livre *Databrarianship : the Academic Data Librarian in Theory and Practice* et a publié sur des sujets allant des algorithmes d'anonymisation des données à la psychologie intergénérationnelle.
kthom67@uwo.ca | ORCID 0000-0002-4152-0075 (https://orcid.org/0000-0002-4152-0075)

# REMERCIEMENTS

*La gestion des données de recherche dans le contexte canadien* n'aurait pas été possible sans la collaboration et la participation des membres de la communauté universitaire des données au Canada, ainsi que des représentants des agences qui soutiennent la gestion des données de recherche.

L'idée initiale de créer une ressource de ce type est venue de la liste de diffusion canadienne RDM-OER, qui réunit un groupe de personnes qui supportent la gestion des données de recherche au Canada. Lachlan MacLeod a joué un rôle déterminant dans la formation de ce groupe et dans la discussion sur l'élaboration d'un manuel ouvert sur la gestion des données de recherche. Les abonnés de la liste de diffusion RDM-OER ont fourni des informations, des commentaires et un soutien tout au long du projet.

Nous sommes reconnaissantes à Serena Henderson pour le soutien qu'elle nous a apporté durant les phases initiales du projet grâce au soutien financier de l'Université Dalhousie. Pour la version anglaise de ce manuel, Yeliz Cengay nous a aidées à saisir les chapitres dans Pressbooks.

Ce projet n'aurait pas pu voir le jour sans le soutien financier de plusieurs groupes différents. *La gestion des données de recherche dans le contexte canadien* bénéficie en partie du soutien financier du Conseil de recherches en sciences humaines du Canada. Nous remercions également le soutien financier de Compute Ontario, de Western University Research Mobilization, Creation & Innovation Grant, de Western Libraries, Western University Academic Activity Support Fund, de la subvention de recherche de l'Université de Concordia et de la University of British Columbia OER Rapid Innovation Grant. L'Université Dalhousie a apporté son soutien à l'embauche d'une coordonnatrice de projet au début du projet. L'Alliance pour la recherche numérique au Canada a apporté son soutien graphique.

La couverture a été conçue par CC Goodwin Consulting.

Les services de révision des chapitres originaux en anglais ont été assurés par Paula Chiarcos et Amanda Feeney de Colborne Communications. La révision des chapitres originaux en français a été assurée par Suzanne Aubin de Colborne Communications et Jonathan Dorey. La traduction du français vers l'anglais a été effectuée par Jonathan Dorey et Amanda Feeney. La traduction de l'anglais vers le français a été effectuée par Manon St-Jules et Suzanne Aubin. Une révision supplémentaire de la version française du chapitre 3 « La souveraineté des données autochtones » a été assurée par Wintranslation.

Nous tenons tout particulièrement à souligner les efforts des personnes responsables de l'évaluation par les

pairs qui ont contribué à garantir l'intégrité académique et la qualité du manuel. Les personnes suivantes ont apporté leur aide :

Jennifer Abel
Fatoumata Bah
Lacey Cain
Alicia Cappello
Erin Clary
Mathieu Clouthier
Alexandra Cooper
Lyne Da Sylva
Sarah Forbes
Jane Fry
Meghan Goodchild
Monique Grenier
Alex Guindon
Melissa Helwig
Laurence Horton
Jasmine Hoover
Fiona Inglis
Erin Johnson
Sandra Keys
Marjorie Mitchell
Nora Mulvaney
Kaitlin Newson
Paul R. Pival
Isaac Pratt
Kharah Ross
Kimberly Silk
Tara Stieglitz
Robyn Stobbs
Carolyn Sullivan
Felicity Tayler
Arielle Vanderschans
Minglu Wang
Susie Wilson
Shiloh Williams
Nadia Zurek

# AVANT-PROPOS : RÉFLEXIONS SUR UNE CARRIÈRE DE BIBLIOTHÉCAIRE DE DONNÉES

Jeff Moon

Au cours des dernières années, la reconnaissance de la gestion des données de recherche (GDR) en tant que pilier essentiel des activités de recherche est montée en flèche, et ce, grâce aux efforts des bibliothécaires, des spécialistes des données, des responsables du développement en recherche, des autorités politiques, des organismes de financement, des maisons d'édition, des administrations dans les établissements d'enseignement supérieur et d'un nombre croissant de chercheuses et chercheurs de première ligne. Comment y sommes-nous arrivés? En réfléchissant à mes 36 années d'expérience dans ce domaine, la réponse m'apparaît évidente : grâce à la communauté. La nature collégiale et collaborative de la communauté canadienne des données nous a permis au fil des décennies d'atteindre cette reconnaissance, car nous croyons tous qu'ensemble, nous pouvons faire mieux. Pour mettre en contexte l'origine et l'objectif de cette nouvelle ressource éducative libre (REL), j'ai revisité ce parcours. Mon récit de notre histoire commune prendra une couleur personnelle et s'avérera sélectif; vous pouvez approfondir le sujet en consultant les excellents travaux de Gray et Hill (2016) et de Humphrey (2020).

Je suis arrivé à l'Université Queen's en 1987, fort d'une formation en biologie, d'un diplôme en bibliothéconomie et de connaissances de base en statistiques et en ordinateurs centraux. C'est d'ailleurs grâce à ces derniers que j'ai été engagé comme premier bibliothécaire de données à Queen's. Je crois qu'à l'époque, l'Université était l'un des six établissements canadiens à avoir des bibliothécaires de données. Très tôt, j'ai appris que cette profession était une activité aérobique : apporter les rubans de données à neuf pistes au centre informatique, revenir à la bibliothèque, accomplir une tâche en lot sur l'ordinateur central, retourner au centre informatique pour récupérer les résultats imprimés, revenir à la bibliothèque, trouver et corriger les erreurs; répéter le processus. Je n'ai jamais été aussi en forme!

À cette époque, le gouvernement fédéral imposait des mesures de recouvrement des coûts qui décuplaient le prix pour les données de Statistiques Canada, soit de 25 $ à 2 500 $ par fichier, rendant les données inaccessibles pour les chercheuses, les chercheurs et les universités. Laine Ruus, une bibliothécaire de données de longue date à l'Université de Toronto, était d'avis qu'ensemble, nous pouvions faire mieux. En collaboration avec l'Association des bibliothèques de recherche du Canada (ABRC), elle a mené des négociations afin d'acheter un ensemble de fichiers de données du recensement auprès de Statistiques Canada afin de les copier puis de les partager sous licence avec les établissements participants. La tâche gargantuesque

et totalement altruiste de copier et d'envoyer des centaines de rubans magnétiques à l'échelle du pays a fait en sorte que les données demeuraient abordables et accessibles pour les 25 établissements participants.

Ce succès comprenait son lot de défis : que devaient faire les bibliothèques universitaires avec ces rubans? La plupart du temps, les bibliothécaires étaient responsables des documents gouvernementaux et se voyaient attribuer le rôle de « bibliothécaire de données », bien que la plupart n'avaient pas de formation dans le domaine. C'est ainsi qu'en 1988 est née l'Association canadienne des utilisateurs de données publiques (ACUDP) dont l'un de ses principaux mandats était de former ses membres. Parmi les personnes à la tête de cette formation, mentionnons Wendy Watkins (Université Carleton) et Laine Ruus. D'abord offerte de manière non formelle, souvent individuellement, la formation est devenue plus formelle à l'occasion de diverses conférences.

Plus tard, Wendy s'est jointe à Ernie Boyko de Statistiques Canada pour entreprendre un projet de grande envergure : élaborer ce qui est devenu l'Initiative de démocratisation des données (IDD) et trouver les ressources nécessaires pour ce modèle de service des données national conçu pour fournir un accès aux données de Statistiques Canada et, surtout, une formation ciblée moyennant le paiement de frais d'abonnement annuels fixes et abordables. Cette réussite a nécessité beaucoup d'adhésion, de temps et d'effort. En 1995, dans un rapport régional au ICPSR (https://iassistdata.org/about/regional-report-1994-1995-canada/) (en anglais uniquement), Wendy indiquait qu'à ce jour, toutes les parties étaient enthousiastes, mais qu'il manquait des engagements fermes en matière de financement. Au moment du lancement en 1996, plus de 50 établissements s'étaient joints à l'initiative en tant que « représentants de l'IDD » et bénéficiaient du double avantage des économies et de l'indispensable  formation. Un autre avantage, moins tangible, à émerger de l'IDD a été une communauté de pratique dans le cadre de laquelle les bibliothécaires de données plus habiles offraient du soutien, une direction et des encouragements à un nombre croissant de nouvelles recrues dans le domaine des données au Canada. Ce réseau d'expertise et de mentorat a de facto aidé à créer des liens, à bâtir la confiance et la crédibilité et constitue un modèle de développement communautaire dont nous tirons profit aujourd'hui.

Avançons rapidement dans le temps : je vois le progrès depuis les rubans magnétiques jusqu'aux cartouches puis aux CD-ROM, à la fois individuels et dans des « tours » réseautées, jusqu'à l'émergence des données livrées par Internet par site FTP puis par le Web. Plusieurs services canadiens de livraison de données fondés sur le Web ont vu le jour au cours de cette période; leurs noms abstraits rappelleront des souvenirs aux bibliothécaires d'un certain âge : IDLS, Equinox, QWIFS, LANDRU, ISLAND, Sherlock et SDA. L'IDD a souvent offert une formation régionale au sujet d'un ou de plusieurs de ces services. Ce pot-pourri de systèmes a servi de terrain d'essai pour des solutions nationales plus ambitieuses; plusieurs de ces plateformes offraient un accès par abonnement aux établissements d'un océan à l'autre.

Fait important : au cours de cette période, le concept de gestion des données est apparu et s'est développé, bien que lentement. Plusieurs bibliothécaires de données ont commencé à participer à des « sauvetages de

données », montrant ainsi le risque de perdre les fichiers de données produits par le gouvernement par cause d'ignorance, d'absence de financement ou par négligence. Statistiques Canada a souvent demandé à Laine Ruus, une collectionneuse hors pair, si elle avait conservé (géré) une copie des données dont ils avaient besoin et qu'ils ne trouvaient plus. Le rapport régional de l'ICPSR a illustré la situation en mentionnant les activités de la Data Library de l'Université de l'Alberta qui a réussi à sauver 20 années de données de l'étude Albert Hail après la fermeture d'un programme gouvernemental provincial. Il est aujourd'hui possible d'accéder à ces données dans Borealis (https://borealisdata.ca/dataverse/dv?q=alberta+hail+), le dépôt Dataverse canadien.

Au fur et à mesure que la technologie a évolué, ainsi en a-t-il été de l'importance d'effectuer des recherches numériques. Comme dans les cas d'initiatives de sauvetage de données mentionnées précédemment, la valeur, combinée à la fragilité, des données générées par les chercheuses et les chercheurs a émergé dans les consciences. Au cours de la dernière décennie, le gouvernement fédéral et ses trois organismes de financement ont publié une gamme de documents politiques fondamentaux qui définissent leur position par rapport à la science ouverte et à l'importance de la transparence, de la reproductibilité, de la vérification et de la réutilisation des données. Les bibliothèques aussi, guidées par l'Association des bibliothèques de recherche du Canada (ABRC) dirigée de main de maître par la directrice générale Susan Haigh, se sont beaucoup intéressées à la GDR. Grâce au soutien des directrices et directeurs de bibliothèque de l'ABRC et du leadership visionnaire de Charles (Chuck) Humphrey (Université de l'Alberta), une feuille de route de la GDR au Canada a vu le jour, aboutissant en 2015 à la création du réseau Portage de l'ABRC. En 2017, j'ai accepté de relever le défi de prendre le relais de Chuck lorsqu'il a pris sa retraite. Je me suis alors joint à Lee Wilson, alors gestionnaire de service de Portage, dans le but de continuer à développer le réseau pancanadien d'expertes et experts (signe de tête de reconnaissance à l'IDD), lequel a été mis sur pied pour, à partir de zéro, développer et coordonner la capacité de GDR et la formation au Canada. Ensemble, nous avons supervisé l'intégration du réseau Portage dans l'Alliance de recherche numérique du Canada (l'Alliance). L'équipe de GDR à l'Alliance et le réseau d'expertes et experts, désormais dirigés par Lee Wilson, a poursuivi le travail de Portage en étroite collaboration avec d'autres dans l'écosystème d'infrastructure de recherche numérique afin d'améliorer les pratiques de gestion de données, les plateformes, les services et la formation à l'échelle du Canada.

Peu après le lancement du réseau Portage, on m'a demandé de préparer un programme de GDR de premier cycle pour l'école de bibliothéconomie de l'Université Western. Au bout d'une recherche méticuleuse, j'ai fini par choisir un manuel rédigé au Royaume-Uni comme base pour le cours. Admirablement rédigé et exhaustif, il ne portait toutefois que sur les outils, cadres politiques et exemples britanniques et européens. Si plusieurs aspects de la GDR dépassent les frontières nationales, faire comprendre son contexte local à la communauté étudiante canadienne aurait eu une grande valeur. D'autres personnes ont exprimé une frustration semblable alors qu'elles cherchaient un soutien local à la GDR qui faisait autorité en la matière.

Portage, et maintenant l'Alliance, a fait beaucoup pour aborder les besoins de formation en GDR au Canada,

entre autres en collaborant étroitement avec le réseau d'expertes et experts en GDR. Le Groupe d'experts national sur la formation (GENF) a été particulièrement impliqué pour créer une vaste gamme de webinaires, de modèles, de guides, de glossaires, de vidéos et de cours préparatoires – tous offerts gratuitement sur le site Web d'alliancecan.ca (https://alliancecan.ca/fr/services/gestion-des-donnees-de-recherche/apprentissage-et-ressources). En même temps, d'autres membres de la communauté de GDR ont convenu qu'il était possible d'en faire plus. Mentionnons Lachlan MacLeod de l'Université Dalhousie qui a lancé la discussion à propos de la création d'un manuel ouvert sur la GDR, organisé des appels communautaires et établi une liste de distribution pour les personnes intéressées par le sujet. Une équipe de rédaction nationale a été mise sur pied; elle est composée d'Elizabeth (Liz) Hill, de Kristi Thompson et d'Emily Carlisle-Johnston, toutes de l'Université Western (anglais) et de Danielle Dennie (Université Concordia) ainsi qu'Émilie Fortin (Université Laval) (français).

L'équipe de rédaction anglophone a préparé le concept initial pour l'élaboration du manuel, la levée de fonds et la révision des soumissions en anglais. Liz Hill apporte une grande expérience en données et GDR, une connaissance approfondie de l'histoire des services de données au Canada (consultez l'article mentionné ci-après ainsi que le chapitre sur l'histoire compris dans ce manuel) et connaît presque tous les membres de l'écosystème de données canadien – autant de membres qui la connaissent en retour. Elle a su rallier habilement les gens et les relations pour ce projet. Kristi Thompson apporte une expérience en sciences informatiques et en analyse quantitative au projet, ce dont, en plus de son expérience en rédaction, elle a tiré profit pour réviser le contenu technique du manuel. Elle est reconnue pour son travail d'anonymisation des données (consultez le chapitre sur les données sensibles), sa capacité de lecture de textes au contenu quantitatif ainsi que sa participation au « sauvetage de données », le tout bien ancré dans une expertise solide en GDR. Kristi a aussi mené des efforts de collecte de fonds fructueux pour le projet. L'équipe de rédaction a bénéficié du travail d'Emily Carlisle-Johnston qui dispose d'une expertise essentielle en REL, en révision et en élaboration de manuels. Ses connaissances de la plateforme de publication ouverte Pressbooks, son plaidoyer à l'égard de l'ouverture tout au long du processus du projet et son expérience à diriger le processus de rédaction de ressources éducatives libres (REL) alors qu'elle travaillait chez eCampusOntario ont fait d'elle une ressource parfaite pour ce projet.

L'équipe de rédaction francophone était responsable de superviser la traduction, de réviser les contributions en français et de diriger la production d'une édition entièrement en français. Émilie Fortin dispose d'une expérience variée et d'une formation en préservation en plus d'avoir rédigé des documents essentiels sur les métadonnées et les formats pour ce manuel. Elle travaille en GDR depuis 2021. Danielle Dennie a une formation en bibliothéconomie scientifique et en GDR et elle a occupé plusieurs postes de direction de bibliothèque. Danielle est la coordonnatrice principale entre les aspects anglophones et francophones du projet; elle est l'agente de liaison avec l'équipe anglophone et supervise le travail des réviseurs et des traducteurs. Danielle et Émilie ont toutes deux contacté la communauté des données francophone et ont traduit des communications pour le projet.

Cette équipe de rédaction nationale dispose d'une vaste gamme de compétences et de niveaux d'expérience; chaque membre apporte une contribution distincte et complémentaire. En fin de compte, leurs efforts conjoints ont attiré plus de 50 membres de la communauté canadienne des données pour des rôles d'édition, de rédaction, de révision, de collecte de fonds, entre autres contributions à ce projet. Cette équipe pancanadienne élargie partage une appréciation de la valeur et de l'importance d'encadrer la formation et les ressources en GDR dans un contexte canadien et a décidé de combler ce besoin. Résultat : ce manuel sur la GDR bilingue, entièrement canadien, *La gestion des données de recherche dans le contexte canadien : un guide pour la pratique et l'apprentissage* (https://ecampusontario.pressbooks.pub/gdrcanada/).

Il est passionnant de penser à quel point ce travail promet d'être précieux et apprécié dans le cadre d'un arsenal toujours plus grand de ressources canadiennes de formation en matière de GDR. Ce manuel vise les chercheuses et les chercheurs et les spécialistes à tous les niveaux et de toutes les disciplines. Il présente un fort potentiel d'utilisation dans les contextes suivants :

- En tant que matériel éducatif (cours, ateliers, école de bibliothéconomie, etc.);
- En tant que source de référence (pour les chercheuses et chercheurs et spécialistes en GDR – novices ou avec de l'expérience);
- En administration, pour les gestionnaires qui souhaitent en savoir davantage sur les aspects politiques et réglementaires de la GDR;
- En tant que moteur de changement pouvant être mis en application dans des discussions sur les politiques, leur élaboration et leur mise en place.

Le fait que ce manuel soit en ligne et libre facilite son accessibilité et ses possibilités d'amélioration continue. Le paysage de la GDR évolue sans cesse grâce aux progrès réalisés sur la scène locale, régionale, nationale et internationale. Autant de travaux qui peuvent nourrir et améliorer cette référence au fil du temps.

Au fond, ce manuel incarne un océan de changements dans l'écosystème canadien des données. Nous témoignons et participons à élargir notre objectif collectif national qui ne se limite plus à faciliter l'accès et l'utilisation des données existantes, mais à développer activement le contenu disponible en favorisant et en soutenant la FAIR (https://www.frdr-dfdr.ca/docs/fr/principes_fair/)-isation des données générées par les chercheuses et chercheurs selon les moyens décrits dans ce manuel. Les pratiques exemplaires, les conseils, l'orientation, les discussions sur les politiques et les exemples renforceront certainement les efforts déployés pour normaliser l'attention nécessaire et croissante portée aux principes FAIR. Je choisis le verbe « normaliser », car nous devons faire des pratiques exemplaires relatives à la gestion des données de recherche une normalité. Nous devons aussi nous attendre à ce qu'elles soient intégrées aux mentalités et aux processus de travail des différentes communautés de recherche,  et ce, non seulement en réaction à des impératifs politiques, mais parce que les chercheuses et chercheurs reconnaissent et valorisent les bienfaits des données bien gérées – pour leur discipline, leur réputation, la réutilisation et vérification futures, et la société dans son

ensemble. Ce manuel nous aidera à atteindre cet objectif. Ne sous-estimez jamais le pouvoir d'une communauté dévouée à l'action.

Mars 2023

Gray, S. V. et Hill, E. (2016). The Academic Data Librarian Profession in Canada: History and Future Directions. Dans L. Kellam et K. Thompson (dir.), *Databrarianship: The Academic Data Librarian in Theory and Practice* (p. 321-334). Association of College and Research Libraries. http://ir.lib.uwo.ca/wlpub/49 (http://ir.lib.uwo.ca/wlpub/49)

Humphrey, C. (2020). The CARL Portage Partnership Story. *Partnership: The Canadian Journal of Library and Information Practice and Research, 15*(1). https://doi.org/10.21083/partnership.v15i1.5825

## À propos de l'auteur

### Jeff Moon

Jeff Moon est directeur de la stratégie et des services de données chez Compute Ontario.

PARTIE I

# POINT DE DÉPART EN GESTION DES DONNÉES DE RECHERCHE

1.

# LES RUDIMENTS: UNE INTRODUCTION À LA GESTION DES DONNÉES DE RECHERCHE

Kristi Thompson

**Objectifs d'apprentissage**

À la fin de ce chapitre, vous pourrez:

1. Définir les termes « données de recherche, » « gestion des données de recherche » et « plan de gestion des données. »
2. Décrire les trois éléments de la Politique de 2021 des trois organismes sur la gestion des données de recherche.
3. Comprendre le lien entre la gestion des données de recherche et la réplicabilité de la recherche.
4. Énumérer les éléments courants d'un plan de gestion des données et expliquer leur importance.

## Introduction

En 2021, les trois agences fédérales de financement de la recherche au Canada – les Instituts de recherche en santé du Canada (IRSC), le Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG) et le Conseil de recherches en sciences humaines (CRSH) – ont publié la **Politique des trois organismes sur la gestion des données de recherche**. L'objectif de la politique est d'assurer que « les données recueillies par la recherche au moyen de fonds publics [soient] gérées de manière responsable et sûre. Elles doivent aussi,

lorsque les obligations éthiques, juridiques et commerciales le permettent, être disponibles pour être réutilisées par d'autres » (Gouvernement du Canada, 2021a). Les agences de financement de plusieurs autres pays ont émis des politiques semblables.

Dans ce chapitre, nous discuterons de quelques-unes des questions fondamentales en lien avec la gestion des données de recherche (GDR) au Canada: d'où viennent ces efforts de formalisation de la GDR? En quoi consistent les données de recherche dans le contexte de cette politique et de façon générale? Quelles sont les exigences d'une bonne gestion des données?

**Les agences fédérales de financement de la recherche au Canada**

Le Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG), le Conseil de recherches en sciences humaines (CRSH) et les Instituts de recherche en santé du Canada (IRSC) représentent les trois agences fédérales de financement de la recherche du Canada. Collectivement, elles sont parfois appelées les trois organismes ou trois conseils; tout au long du texte, nous emploierons souvent le terme « **organismes subventionnaires** » pour les désigner collectivement. Ils sont à la source d'une importante proportion des fonds de recherche au Canada et sont donc en mesure d'établir des politiques qui ont un grand impact sur la façon dont les recherches sont menées au Canada. En plus de la Politique des trois organismes sur la gestion des données de recherche, ils sont également responsables de l'*Énoncé de politique sur l'éthique de la recherche avec des êtres humains* (**EPTC 2**), la Politique sur le libre accès aux publications et autres (https://science.gc.ca/site/science/fr/financement-interorganismes-recherche/politiques-lignes-directrices). Leurs politiques ne constituent pas des lois. Les organismes subventionnaires peuvent décider d'accorder ou non des fonds à certaines chercheuses ou certains chercheurs, mais ils peuvent aussi interdire à un établissement entier de gérer des fonds de recherches, rendant ainsi chaque chercheuse et chercheur de cet établissement inéligible à soumettre des demandes de fonds. Les organismes subventionnaires ont donc une influence énorme sur la façon dont les recherches sont menées au Canada.

## En quoi consistent les données de recherche?

Pour bien comprendre les exigences en GDR, vous devez comprendre la définition des **données de recherche**. Le terme « données de recherche » combine deux concepts : la recherche et les données. La

recherche peut être décrite comme étant un processus d'enquête systématique, un moyen d'en apprendre plus sur des phénomènes variés. La recherche transforme l'information en connaissances et constitue un moyen par lequel nous découvrons le monde. Les données peuvent représenter une part importante de cette découverte de connaissances. Les données constituent des types d'informations ou de preuves qui servent de base à une recherche. Mais ce ne sont pas toutes les informations incluses dans un projet de recherche qui sont des données.

La foire aux questions (2021) des trois organismes (https://science.gc.ca/site/science/fr/financement-interorganismes-recherche/politiques-lignes-directrices/gestion-donnees-recherche/politique-trois-organismes-gestion-donnees-recherche-foire-aux-questions#1b) du Canada établit que « la définition des données de recherche pertinentes est très souvent contextuelle et la détermination de ce qui compte comme tel devrait être guidée par les normes disciplinaires » (Gouvernement du Canada, 2021b). Autrement dit, le contexte est important; les données de recherche ne peuvent être définies sans savoir de quelle façon elles seront générées et utilisées. La section de la FAQ qui traite des liens entre les documents de recherche et les données de recherche se penche sur cette question : « les matériaux de recherche font l'objet d'une enquête – de nature scientifique, universitaire, littéraire ou artistique – et sont utilisés pour créer des données de recherche. Ils sont transformés en données par la méthode ou la pratique. »

Cette transformation est fondamentale pour séparer les informations générales des données de recherche. Les données sont le résultat de la collecte d'informations brutes issues d'une source quelconque (p. ex., des réponses à un sondage, des données d'archives ou bibliographiques, des médias sociaux, des instruments scientifiques, des documents textuels) et de l'assemblage de cette information en une forme structurée qui peut servir de base à des recherches éventuelles. En raison du travail nécessaire pour structurer, annoter et organiser les données de recherche, elles peuvent aussi être considérées comme des résultats de recherche, au même titre que les livres, les articles et autres éléments créés par des chercheuses et chercheurs. Les données de recherche constituent une source vitale d'informations, mais elles demeurent souvent inaccessibles. Si elles sont publiées ou partagées, d'autres chercheuses et chercheurs peuvent les consulter et elles peuvent être citées comme tout autre résultat de recherche.

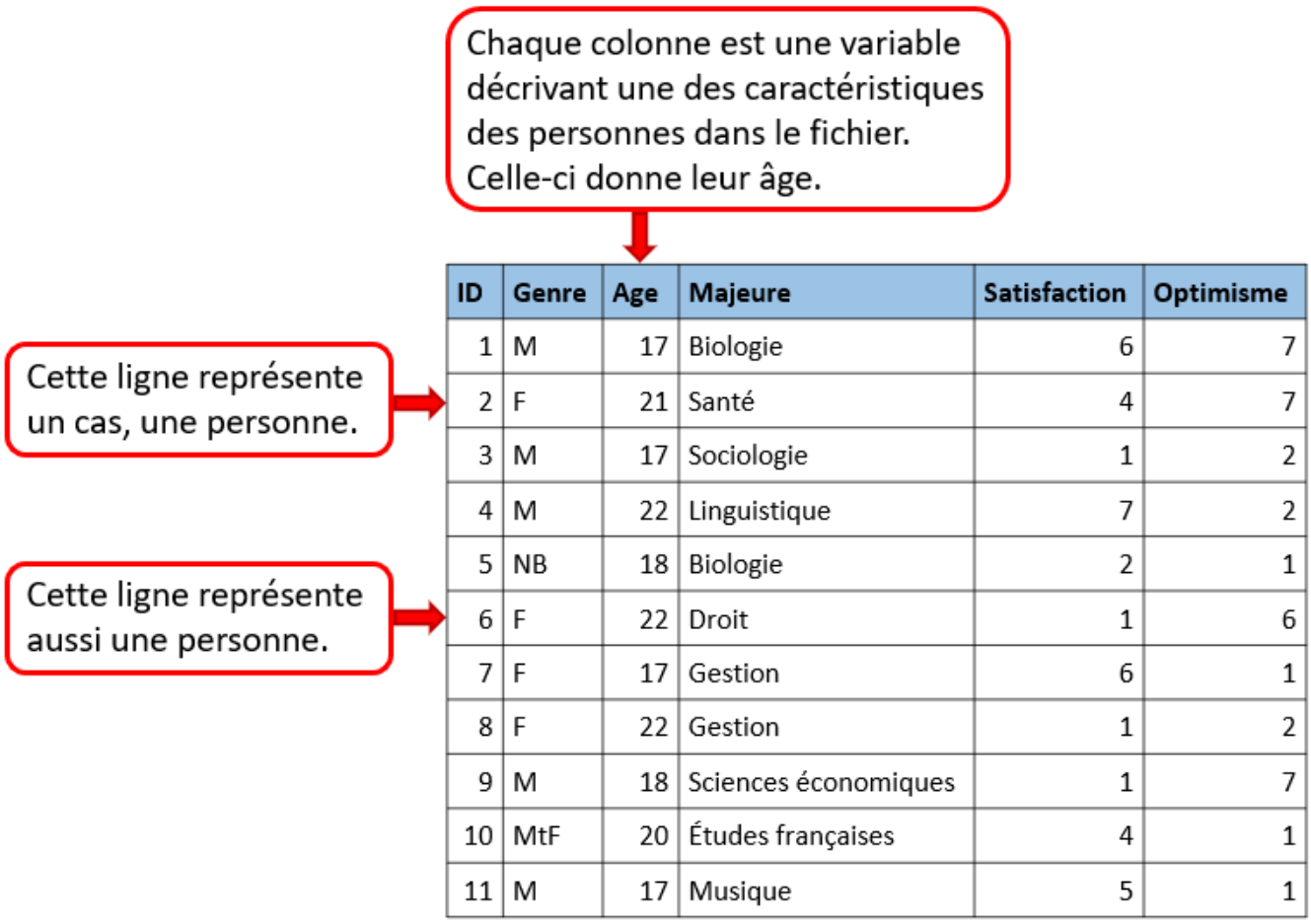
Par exemple, un chercheur peut utiliser une série d'articles de recherche comme point de départ pour sa recherche. S'il en fait simplement la lecture et se rapporte à leurs contenus par le biais de citations pour appuyer d'autres idées, les articles servent de matériaux de recherche et non de données de recherche. Si toutefois ce chercheur utilise la même série d'articles, les importe dans un logiciel, les étudie et les annote sous une forme structurée pour ensuite formuler une conclusion globale sur l'ensemble des articles, ces articles deviennent alors un jeu de données et représentent des données de recherche.

Les données de recherches peuvent être des données secondaires, ce qui implique que la chercheuse ou le chercheur n'a pas recueilli ou assemblé les matériaux lui-même. Dans ce cas, le travail fait pour structurer ou peaufiner les données pour qu'elles servent d'intrant peut avoir été fait par quelqu'un d'autre. Ou encore, les

données peuvent déjà arriver avec une structure s'il s'agit de **données administratives** (extraites, par exemple, d'une base de données d'un bureau d'admission). Mais un ensemble structuré d'informations qui est affiné lors de la recherche par le biais d'une analyse représente tout de même des données de recherche.

## La structure des données

Utilisé pour les tableurs et fichiers statistiques, le rectangle est une structure courante pour les données. À l'intérieur de ce format, les données sont organisées en rangées et en colonnes. Chacune des rangées contient un cas – soit une unité simple de l'objet étudié (p. ex., une personne dans une enquête ou une mouche à fruits dans une expérience). Chacune des colonnes sera utilisée pour stocker une variable ou caractéristique pour chacun des cas, tels que l'âge de chaque personne (ou des mouches à fruits) dans l'étude.




| ID | Genre | Age | Majeure | Satisfaction | Optimisme |
|---|---|---|---|---|---|
| 1 | M | 17 | Biologie | 6 | 7 |
| 2 | F | 21 | Santé | 4 | 7 |
| 3 | M | 17 | Sociologie | 1 | 2 |
| 4 | M | 22 | Linguistique | 7 | 2 |
| 5 | NB | 18 | Biologie | 2 | 1 |
| 6 | F | 22 | Droit | 1 | 6 |
| 7 | F | 17 | Gestion | 6 | 1 |
| 8 | F | 22 | Gestion | 1 | 2 |
| 9 | M | 18 | Sciences économiques | 1 | 7 |
| 10 | MtF | 20 | Études françaises | 4 | 1 |
| 11 | M | 17 | Musique | 5 | 1 |

Figure 1. Une image d'un fichier de données en rectangle. Il s'agit d'un tableur avec une rangée pour chacune des personnes dans le jeu de données et une colonne pour chacune des caractéristiques.

Puisque nous discutons de structure des données, voici quelques règles de base pour bien organiser les données en rectangle, de type tableur, afin d'en faciliter la gestion :

- Organisez les données dans un seul rectangle, avec les sujets/cas dans chacune des rangées et les variables/caractéristiques dans chacune des colonnes; ajoutez une rangée en haut pour l'en-tête avec des noms brefs qui décrivent ce que représente chacune des colonnes;
- Inscrivez un seul élément par cellule et ne jumelez pas les cellules. Chacune des cellules devrait comporter une seule information qui correspond à une rangée et une colonne (un cas et une variable);
- Créez un dictionnaire des données – un document distinct qui explique le contenu de vos rangées et colonnes;
- N'ajoutez pas de calculs ou de fonctions dans les fichiers de données originales;
- N'utilisez pas de polices colorées ou de surlignage en tant que données.

La figure ci-dessus illustre à quoi ressembleront des données ainsi structurées. Les données organisées dans ce type de format peuvent être lues et utilisées par tout logiciel de tableur ou progiciel statistique.

## Qu'est-ce que la gestion des données de recherche?

La **gestion des données de recherche** est un terme général qui décrit ce que font les chercheuses et chercheurs pour structurer, organiser et entretenir les données avant, pendant et après leur travail de recherche. En ce sens, toute personne qui recueille ou utilise des données avec l'intention de mener une recherche fait de la gestion des données de recherche. Créer un fichier de données, décider où il sera sauvegardé, lui attribuer un nouveau nom ou le déplacer dans un nouvel emplacement représentent toutes des activités de gestion des données de recherche. La gestion des données de recherche (GDR) est également un domaine d'étude émergent. Cette nouvelle discipline se préoccupe d'étudier et de développer des moyens plus efficaces de gérer des données de recherche. L'idée qui sous-tend la gestion des données est l'utilisation d'un ensemble de techniques pour structurer, organiser et documenter les informations qui serviront de base à la recherche et de le faire de façon à ce que d'autres puissent comprendre et reproduire votre recherche, ainsi qu'utiliser les données qui ont servi à votre recherche.

Le **cycle de vie des données de recherche** est souvent utilisé pour illustrer la nature cyclique d'une recherche. Les chercheuses et chercheurs commencent par planifier leur recherche. Ensuite, les données sont recueillies, traitées et nettoyées avant d'être analysées pour permettre aux chercheuses et chercheurs de formuler des conclusions. Finalement, des mesures sont prises pour préserver les données à long terme et pour les rendre disponibles à d'autres qui les utiliseront pour leur étude. En pratique, le cycle est plus complexe avec plusieurs étapes qui se chevauchent. Par exemple, la préservation des données originales doit commencer dès la collecte des données pour éviter tout risque de perte, et les chercheuses et chercheurs peuvent souvent traiter, analyser et traiter à nouveau leurs données tout au long du processus. Il s'agit d'une perspective très axée autour des données, puisque le cycle de recherche comprend également plusieurs autres étapes, comme la soumission de demandes de financement, ainsi que la rédaction et la publication des résultats.

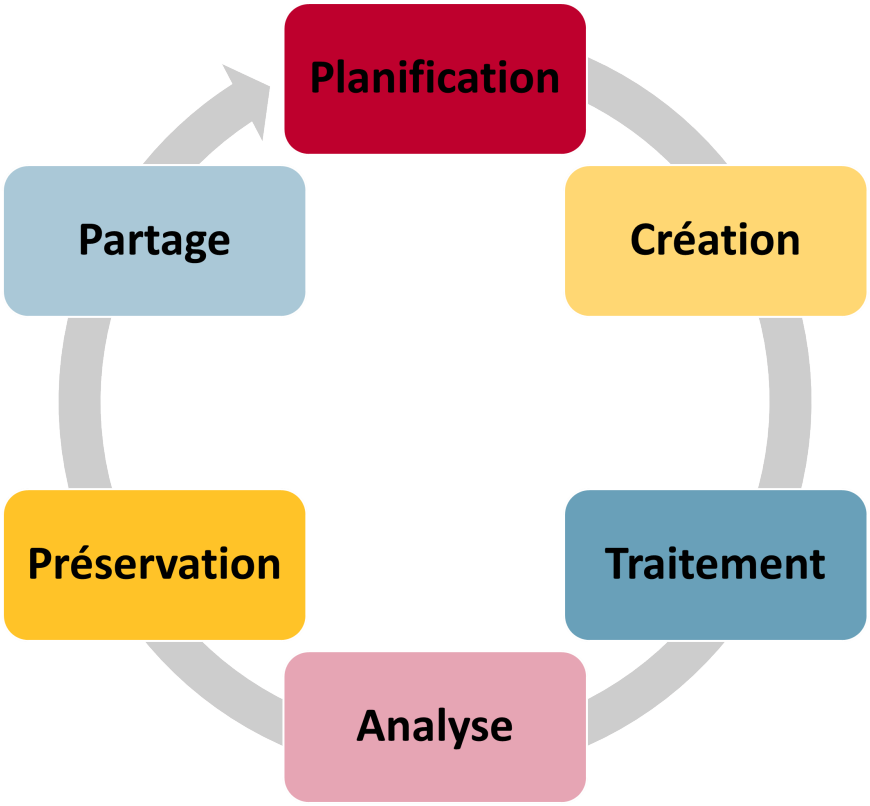



Figure 2: Le cycle de vie des données de recherche

## Reproductibilité, réplicabilité et traçabilité

La reproductibilité, la réplicabilité et la traçabilité sont trois concepts à la fois reliés mais distincts, qui sont essentiels pour bien comprendre l'importance d'une bonne GDR. Pour qu'une recherche soit **reproductible**, il faut que des chercheuses ou chercheurs qui ne faisaient pas partie de l'équipe de recherche originale puissent reprendre la recherche en utilisant les mêmes données, méthodes et codes et aboutir aux mêmes résultats. Concrètement, cela implique que les chercheuses ou chercheurs externes doivent avoir accès aux données, au code et à une documentation détaillée.

Pour qu'une recherche soit **réplicable**, les chercheuses ou chercheurs qui ne faisaient pas partie de l'équipe de recherche originale doivent être en mesure de reproduire la recherche originale avec des données différentes ou nouvellement recueillies et arriver aux mêmes résultats ou à des résultats semblables. Pour ce faire, les méthodes de l'équipe de recherche originale doivent avoir été documentées et publiées, mais les données originales ne doivent pas nécessairement être disponibles.

Pour qu'une recherche soit **traçable**, les chercheuses ou chercheurs qui ne faisaient pas partie de l'équipe de recherche originale doivent être en mesure de reproduire le jeu de données analysé à partir du jeu de données original, tel qu'il a été collecté ou acquis. Si les données sont traçables, il est possible de conclure avec

confiance qu'aucune modification non documentée n'a été faite au jeu de données. Les chercheuses ou chercheurs externes devraient aussi pouvoir comprendre le raisonnement derrière chaque modification apportée aux données, qui a apporté ces modifications et le processus décisionnel derrière chacune d'elles. Les données de recherche constituent des preuves – pour les adeptes de séries d'enquêtes policières, c'est comme la chaîne de possession qui assure que les preuves d'une enquête criminelle n'ont pas été contaminées.

Vous vous rappelez des règles de structures des données mentionnées plus tôt dans le chapitre? Des formats et des structures simples, uniformisés et couramment utilisés sont importants pour la reproductibilité, la réplicabilité et la traçabilité.

Le fait de rendre obligatoires certaines normes particulières pour la gestion des données ne vise pas à créer des contraintes arbitraires pour compliquer la façon de mener une recherche. Les normes aident à préserver l'intégrité de la recherche en incitant les chercheuses et chercheurs à manipuler leurs données de façon à ce que leur démarche et leur travail soient compréhensibles. Ainsi, les données peuvent être reproduites et répliquées. Les conclusions de recherche qui ne peuvent être reprises ou reproduites perdent en crédibilité. Une application réglementée de la GDR augmente aussi la possibilité du partage des données, pas seulement pour que la recherche puisse être reproduite directement, mais aussi pour que les données puissent être réutilisées dans d'autres projets, créant ainsi plus d'occasions de recherche en limitant les coûts. La Politique des trois organismes de 2021 sur la gestion des données de recherches inclut trois exigences qui visent à réaliser cet objectif.

### La crise de la réplicabilité

La crise de la réplicabilité est un problème récurrent dans les sciences physiques et sociales qui remet en question la crédibilité de ces sciences. Vers 2010, des psychologues qui ont voulu reprendre certaines études antérieures pour tenter de reproduire leurs résultats ont été incapables de le faire de façon systématique. Lors d'une importante initiative (https://psyarxiv.com/9654g/) (document en anglais uniquement) qui visait à reproduire 28 études, près de la moitié d'entre elles ne pouvait être reproduite et 32% ont démontré des résultats contraires aux résultats originaux (Klein *et al.*, 2018). Cela implique que certains individus qui dépendent de ces recherches peuvent avoir enseigné, mené des recherches supplémentaires et modifié des pratiques en se basant sur des résultats potentiellement erronés. Des problèmes semblables ont été rapportés dans d'autres domaines, tels que la biologie, la médecine et les sciences économiques. Les études originales peuvent avoir utilisé des données erronées, de mauvaises méthodes d'analyse ou des échantillons

atypiques, parmi les nombreuses causes potentielles des erreurs. Quand les données originales ne sont pas disponibles ou traçables, difficile de le savoir.

## Les trois exigences de la Politique des trois organismes

Les trois exigences comme établies par la Politique des trois organismes sur la gestion des données de recherche (Gouvernement du Canada, 2021a) sont :

1. Les stratégies institutionnelles. Les établissements (généralement les établissements d'enseignement postsecondaire et les hôpitaux) admissibles à administrer des fonds des trois organismes doivent élaborer des stratégies formelles de GDR et les communiquer aux organismes subventionnaires selon une échéance établie. Ces stratégies doivent décrire la façon dont ils prévoient d'appuyer leurs chercheuses et chercheurs dans l'amélioration de leurs pratiques de GDR et dans l'application des deux autres exigences. Les liens vers les stratégies soumises aux organismes subventionnaires sont disponibles sur la page des stratégies institutionnelles (https://science.gc.ca/site/science/fr/financement-interorganismes-recherche/politiques-lignes-directrices/gestion-donnees-recherche/strategies-institutionnelles-gestion-donnees-recherche-publiees).
2. Les plans de gestion des données. Les organismes subventionnaires commenceront à exiger que les chercheuses et chercheurs soumettent des plans qui décrivent la façon dont leurs données seront gérées, du moins pour certaines opportunités de financement. Ces plans seront pris en compte lorsque les organismes subventionnaires auront déterminé la façon dont les fonds seront accordés .
3. Le dépôt des données. Lorsque les bénéficiaires de financement publient un article ou tout autre résultat découlant de la recherche financée par les organismes subventionnaires, les données et le code qui appuient les résultats de la recherche doivent être déposés dans un dépôt numérique. Il s'agit d'une exigence assez limitée. Une chercheuse peut recueillir une douzaine de variables, mais rédiger un article qui n'utilise directement qu'une partie d'entre elles. C'est ce sous-ensemble qui doit être déposé. Il est également important de noter que le dépôt n'est pas synonyme de partage. Les données confidentielles ou qui ne devraient pas être partagées doivent être déposées dans un endroit privé et sécurisé.

# Les plans de gestion des données (PGD)

Un **plan de gestion des données** (PGD) est une description formelle de tout le processus de la chercheuse ou du chercheur, de la collecte des données jusqu'à leur élimination ou suppression. Les PGD ont existé sous différentes formes depuis les années 1960 (Smale *et al*., 2020), mais leur adoption a été lente et reste toujours peu répandue dans certaines disciplines. À l'international, les PGD sont souvent exigés par les organismes subventionnaires, notamment au Royaume-Uni et aux États-Unis. Des outils ou gabarits ont été développés pour aider les chercheuses et chercheurs à élaborer des plans qui pourront répondre aux exigences des organismes subventionnaires. L'outil principal utilisé au Canada est l'**Assistant PGD**. Il s'agit d'un outil en ligne (https://assistant.portagenetwork.ca/) qui pose aux personnes qui l'utilisent une série de questions sur leurs données et plans de recherche en offrant des conseils et une aide contextuelle qui aident à répondre aux questions.

L'objectif des PGD est d'aider les chercheuses et chercheurs dans la gestion de leurs données au cours de toutes les étapes du cycle des données de recherche, de la collecte jusqu'au partage. Ils sont souvent décrits comme des documents vivants ou évolutifs qui doivent être mis à jour au fil des besoins identifiés par les chercheuses ou chercheurs pendant leur travail avec les données. Ils peuvent comporter une variété d'éléments – Williams *et al*. (2017) ont identifié 43 sujets pouvant constituer des éléments nécessaires à un PGD – et les éléments exigés ou utiles peuvent varier d'une discipline ou d'un type de données à l'autre. Les éléments d'un PGD visent à inciter les chercheuses et chercheurs à tenir compte de la façon dont leurs données seront manipulées et des ressources nécessaires avant le début de leur recherche. La Politique des trois organismes demande aux chercheuses et chercheurs de soumettre un plan qui aborde les éléments suivants :

- comment les données seront recueillies, documentées, formatées, protégées et préservées;
- comment seront utilisés les jeux de données existants et quelles nouvelles données seront créées au cours du projet de recherche;
- est-ce que les données seront partagées et si oui, comment;
- l'endroit où les données seront déposées.

Les agences qui financent la recherche au Canada et à l'international veulent que les chercheuses et chercheurs utilisent des PGD pour démontrer que leurs données seront recueillies, stockées et conservées de façon à faciliter la transparence, le partage et la réutilisation des données ainsi que la reproductibilité des résultats. Les chercheuses et chercheurs qui en font usage jouissent d'un avantage lors du dépôt d'une demande de financement pour la collecte ou l'utilisation des données. Les PGD représentent aussi des avantages pour les chercheuses et chercheurs, leur permettant de mieux planifier et de travailler plus efficacement avec leurs données. Les exigences pour les PGD constituent, en fait, une forme d'ingénierie sociale qui vise à inciter les chercheuses et chercheurs à améliorer leur recherche.

Ces bienfaits ne sont généralement pas prouvés. En théorie, la prise en compte minutieuse de tous les éléments d'un PGD devrait entraîner une amélioration de la recherche. Toutefois, la théorie ne répond pas toujours à la pratique. En effet, un examen de toute la littérature montre qu'il existe très peu de preuves publiées et systématiques des bienfaits réels des PGD pour les chercheuses et chercheurs, établissements et organismes de financement (Smale *et al.*, 2020). Puisque les PGD ont été conçus pour améliorer les activités de recherche, il est regrettable que si peu d'attention ait été accordée à étudier s'ils réussissent à répondre à cet objectif ou s'ils peuvent être modifiés et améliorés.

Nous ferons un survol rapide des sujets qui font régulièrement partie des PGD.

## La collecte des données

Les chercheuses et chercheurs doivent faire la liste des types de données qui seront probablement recueillies ou acquises, et identifier les formats de fichiers dans lesquels ces données seront sauvegardées. Dès le début, les chercheuses et chercheurs devraient envisager l'utilisation de formats qui permettent la préservation, le partage et la réutilisation des données; de bons formats sont ceux qui peuvent être utilisés par des progiciels facilement accessibles. Les formats ouverts sont encore mieux; ils ont des normes publiées de sorte que toute personne ayant la formation nécessaire peut écrire un logiciel pour les lire. Les formats ouverts sont à l'épreuve du temps.

Le fait de tenir compte des conventions pour le nommage des fichiers avant même de commencer la collecte des données peut être étonnamment important. Les chercheuses et chercheurs qui n'établissent pas d'avance leur système peuvent se retrouver avec une variété de fichiers avec des noms de type « donnees.csv, » « donnees2.csv, » « donneesfinales.csv, » « donneesnettoyees.csv, » etc. Un exemple d'un bon système pour nommer et pour faire le suivi des différentes versions d'une collection de données peut être « nomdescriptif-changementfait-date.ext. » L'inclusion du changement et de la date dans le nom du fichier constitue une forme rudimentaire de **contrôle des versions**; cette question sera abordée de façon plus détaillée dans le chapitre 10, « Soutenir la recherche reproductible avec la curation active de données. » Le contrôle des versions devrait également comprendre la mise en place d'autres systèmes pour améliorer la traçabilité des données, tels que de noter toute information liée aux changements apportés aux données dans un fichier principal de documentation ou d'effectuer tous les changements aux données en utilisant des codes qui sont mis à jour et sauvegardés après chaque changement.

## Documentation et métadonnées

La documentation est essentielle, tant pour la préservation que pour la traçabilité. Si un fichier est sauvegardé sur disque en tant que séquence de 0 et de 1, mais que personne ne sait ce que représentent ces chiffres, le

fichier n'a donc pas vraiment été préservé. La documentation doit comprendre des éléments tels qu'un document maître indiquant l'origine des données et de quelle façon elles ont été recueillies, des tableurs dont les noms de colonnes sont faciles à comprendre et l'enregistrement d'informations détaillées sur tous les changements apportés aux fichiers de données.

La documentation peut aussi inclure l'attribution de noms aux fichiers et aux dossiers qui sont directement lisibles par une personne ainsi que la création d'une structure raisonnée pour les dossiers et sous-dossiers. Une forme courante de documentation supplémentaire est le fichier **LISEZ-MOI**. Il s'agit tout simplement d'un fichier qui accompagne un dossier et qui fait la liste de tous les fichiers dans ce dossier, qui décrit le contenu de chacun des fichiers et qui explique le rapport entre les différents fichiers (p. ex., s'il y a un fichier qui contient le code utilisé pour générer des fichiers de données).

Pour plusieurs types de données, dont les fichiers de santé et de sondages, les guides de codification sont également importants. Les guides de codification décrivent la structure et le contenu des fichiers de données en fonction d'un schéma quelconque. Par exemple, un guide de codification pour un sondage fera la liste de toutes les questions posées (qui seront codées comme variables), décrira les différentes options de réponses potentielles, expliquera la façon dont les échantillons du sondage ont été sélectionnés et toutes les variables supplémentaires créées par les chercheuses ou chercheurs. Idéalement, vous devriez avoir suffisamment de documentation sur vos données déposées pour qu'une personne qui possède les connaissances dans votre domaine soit en mesure de :

- comprendre et suivre les étapes que vous avez effectuées pour recueillir vos données et les décisions que vous avez ensuite prises en cours de route;
- prendre votre fichier de données originales et reproduire les changements que vous avez apportés qui ont menés à la forme finale des données;
- exécuter les analyses qui ont produit vos résultats finaux publiés.

La section pour la documentation dans un PGD devrait également inclure les informations qui expliquent la façon dont les chercheuses et chercheurs s'assureront de suivre et d'enregistrer chaque modification apportée au fichier de données. Si plusieurs personnes travaillent avec les données, il est particulièrement important d'établir un système.

**Les fichiers de code**

Les programmes statistiques, tels que SPSS, Stata et R, ainsi que les langages de programmation à usage général, tels que Python, vous permettent de modifier et d'analyser les données en inscrivant des commandes dans un fichier de code et de les exécuter. Certains programmes, tels que SPSS, vous permettent aussi de générer des commandes par le biais d'options dans le menu. Si des changements ont été apportés à vos données en utilisant des fichiers de code, vous serez toujours en mesure d'y retourner pour bien comprendre la nature des changements apportés à vos données.

## Le stockage et les sauvegardes

Dans la section sur le stockage et les copies de sauvegarde, les chercheuses et chercheurs peuvent expliquer où les données seront stockées et de quelles façons elles seront sécurisées. Le stockage d'une seule copie des données – sur un disque dur personnel qui peut ne pas fonctionner ou sur une clé USB qui peut être endommagée – est étonnamment courant (Cheung *et al*., 2022). Comme plusieurs l'ont découvert, c'est aussi une très mauvaise idée. Une bonne idée est la mise en place d'un système qui assure la sauvegarde régulière des données. La règle du 3-2-1 pour la sauvegarde est largement utilisée : il devrait y avoir 3 copies de chaque fichier, les copies devraient se retrouver sur deux médias différents et une des copies devrait se retrouver dans un emplacement externe. Si les données sont stockées là où il y a un système de sauvegarde automatisé (tel qu'un serveur départemental ou un service infonuagique), le besoin de créer des copies de sauvegardes supplémentaires est réduit puisqu'une copie se trouve déjà dans le système de sauvegarde.

## La préservation et le partage

La transparence d'une recherche ainsi que la préservation et le partage des données de recherche constituent les objectifs principaux de la GDR; il est donc essentiel d'en parler dans un PGD. Le modèle d'excellence pour le partage des données est de rendre accessible un jeu de données complet et bien documenté dans une archive en ligne afin qu'il puisse être téléchargé. Le jeu de données devrait être accompagné d'une licence ouverte ou Creative Commons, ce qui permet sa réutilisation de façon explicite. Certaines licences incluent une stipulation comme quoi les données utilisées pour des recherches éventuelles doivent être citées de façon

appropriée (même si, sans stipulation, les bonnes pratiques et la courtoisie professionnelle encouragent à le faire).

Si les données sont partagées, l'étape la plus importante est d'identifier le dépôt approprié. Il existe plusieurs dépôts appropriés. Plusieurs établissements (universités, collèges, hôpitaux, etc.) ont des dépôts de données institutionnels dotés de fonctions permettant d'ingérer les données dans des formats conçus pour la préservation. Ces établissements s'engagent à préserver et à sauvegarder les données. Certaines publications savantes individuelles peuvent aussi héberger des archives qui donnent accès aux données liées aux articles qu'elles publient. Il existe également des dépôts disciplinaires qui hébergent des types particuliers de données, telles que des données génomiques ou géospatiales.

Toutefois, le partage ouvert dans un dépôt n'est pas toujours recommandé, et pour certains types de données (dont les données médicales), le partage peut être contraire à l'éthique. Des questions de confidentialité, d'engagements pris auprès de sujets de recherche, de souveraineté des données autochtones, de propriété des données et de propriété intellectuelle peuvent toutes représenter des situations où le partage ouvert de données n'est pas une option. Dans ces cas-là, les chercheuses et chercheurs doivent trouver des moyens alternatifs de partage. Une solution de rechange est de partager une documentation sur les données dans un dépôt et d'inviter les personnes intéressées à communiquer avec l'équipe de recherche pour obtenir un accès aux données. Parfois, certaines parties d'une collection de données peuvent être partagées tandis que d'autres sont considérées comme trop sensibles. Les personnes intéressées peuvent avoir à s'engager à respecter certaines normes éthiques ou d'autres conditions qui s'appliquent. Dans ces situations, les données devront être préservées autrement, dans une archive sécurisée ou sur un réseau privé. Consultez le chapitre 13 sur les données sensibles pour plus d'informations.

Dans le PGD, la section qui traite de préservation et de partage doit expliquer la façon précise dont les données seront préservées à long terme. Elle doit aussi énoncer les dispositions pour le partage des données, y compris le dépôt où elles seront stockées, les parties de données qui seront partagées et, le cas échéant, les conditions d'accès. Si les données ne peuvent être partagées, le PGD doit en expliquer les raisons.

## Conclusion

La gestion des données de recherche est un terme général qui s'applique au travail des chercheuses et chercheurs en lien avec la façon dont leurs données sont organisées et maintenues pendant et après la tenue de leur recherche. Il s'agit d'un domaine en plein essor qui incite les bibliothécaires, les spécialistes des données et les chercheuses et chercheurs à se poser des questions sur les meilleurs moyens de gérer les données tout en intégrant la transparence de la recherche, la préservation ainsi que le partage des données pour qu'elles

puissent être critiquées, étudiées et utilisées par d'autres chercheuses et chercheurs ainsi que par le public intéressé par la recherche. Ultimement, la GDR vise à améliorer la recherche.

### Questions de réflexion

1. Choisissez un domaine d'étude et décrivez quelques exemples de données de recherche qui pourraient être utilisées par des chercheuses ou chercheurs dans ce domaine. Quels types de défis pourraient être liés à la gestion de ces données?
2. Consultez la Politique des trois organismes sur la gestion des données de recherche.
3. Trouvez la stratégie de GDR de votre établissement (ou d'un établissement local). Qu'est-ce qu'elle vous dit sur la façon dont l'établissement perçoit la GDR?
4. Consultez l'Assistant PGD (https://assistant.portagenetwork.ca/) ou utilisez le gabarit de l'Annexe 1 et créez un PGD pour un projet de recherche fictif.

### Éléments clés à retenir

- La gestion des données de recherche (GDR) est un terme général qui se rapporte aux activités entreprises par des chercheuses et chercheurs dans leur travail avec les données. En tant que domaine d'étude, la GDR incite à examiner des questions fondamentales sur les meilleures façons de mener des recherches.
- Les trois agences fédérales de financement de la recherche du Canada ont établi une politique sur la gestion des données de recherche pour encourager les chercheuses et chercheurs à rendre leur recherche plus transparente, à préserver et à partager leurs données.
- Les plans de gestion des données (PGD) sont des documents préparés par les chercheuses et chercheurs pour décrire la façon dont leurs données seront gérées. Ces documents abordent plusieurs aspects du travail avec les données, dont la collecte des données, la documentation, le stockage, le partage et la préservation.

# Bibliographie

Cheung, M., Cooper, A., Dearborn, D., Hill, E., Johnson, E., Mitchell, M. et Thompson, K. (2022). Les pratiques avant les politiques : comportements en matière de gestion des données de recherche au Canada. *Partnership: Revue canadienne de la pratique et de la recherche en bibliothéconomie et sciences de l'information, 17*(1), juillet 2022, 1-80. https://doi.org/10.21083/partnership.v17i1.6779.

Gouvernement du Canada. (2021a). *Politique des trois organismes sur la gestion des données de recherche*. https://science.gc.ca/site/science/fr/financement-interorganismes-recherche/politiques-lignes-directrices/gestion-donnees-recherche/politique-trois-organismes-gestion-donnees-recherche (https://science.gc.ca/site/science/fr/financement-interorganismes-recherche/politiques-lignes-directrices/gestion-donnees-recherche/politique-trois-organismes-gestion-donnees-recherche)

Gouvernement du Canada. (2021b). *Politique des trois organismes sur la gestion des données de recherche – Foire aux questions*. https://science.gc.ca/site/science/fr/financement-interorganismes-recherche/politiques-lignes-directrices/gestion-donnees-recherche/politique-trois-organismes-gestion-donnees-recherche-foire-aux-questions#1b (https://science.gc.ca/site/science/fr/financement-interorganismes-recherche/politiques-lignes-directrices/gestion-donnees-recherche/politique-trois-organismes-gestion-donnees-recherche-foire-aux-questions#1b)

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams Jr., R. B., Alper, S., Aveyard, M., Axt J. R., Babalola, M. T., Bahník, Š., Batra, R., Berkics, M., Bernstein, M. J., Berry D. R., Bialobrzeska, O., Binan E. D., Bocian, K., Brandt, M. J., Busching, R., ... Nosek, B. A. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science, 1*(4), 443-490. https://doi.org/10.1177/2515245918810225 (https://doi.org/10.1177/2515245918810225)

Smale, N. A., Unsworth, K., Denyer, G., Magatova, E. et Barr, D. (2020). A review of the history, advocacy and efficacy of data management plans. *International Journal of Digital Curation, 15*(1), 1-29. https://doi.org/10.2218/ijdc.v15i1.525 (https://doi.org/10.2218/ijdc.v15i1.525)

Williams, M., Bagwell, J. et Zozus, M. N. (2017). Data management plans: The missing perspective. *Journal of Biomedical Informatics, 71*, 130-142. https://doi.org/10.1016/j.jbi.2017.05.004 (https://doi.org/10.1016/j.jbi.2017.05.004)

## À propos de l'auteur

Kristi Thompson

Kristi Thompson est bibliothécaire en gestion des données de recherche à l'Université Western. Elle a précédemment occupé les postes de bibliothécaire des données à l'Université de Windsor et de spécialiste des données à l'Université Princeton. Elle détient un baccalauréat en informatique de l'Université Queen's et une maîtrise en science de l'information de l'Université Western. Kristi soutient des projets de recherche, administre des logiciels d'archivage de données, travaille avec les comités d'éthique de la recherche de l'Université Western et participe au niveau national au développement de l'infrastructure des données de recherche. Elle a coédité le livre *Databrarianship : the Academic Data Librarian in Theory and Practice* et a publié sur des sujets allant des algorithmes d'anonymisation des données à la psychologie intergénérationnelle. kthom67@uwo.ca (mailto:kthom67@uwo.ca) | ORCID 0000-0002-4152-0075 (https://orcid.org/0000-0002-4152-0075)

2.

# LES PRINCIPES FAIR ET LA GESTION DES DONNÉES DE RECHERCHE

Minglu Wang et Dany Savard

**Objectifs d'apprentissage**

À la fin de ce chapitre, vous pourrez :

1. Expliquer l'historique des principes FAIR.
2. Comprendre les principales significations et exigences, en plus des mécanismes qui sous-tendent les principes FAIR.
3. Connaître les outils et les cadres disponibles pour améliorer la conformité des données aux principes FAIR.
4. Comprendre la façon dont les principes FAIR sont inclus et cités dans les politiques de recherche et de disponibilité des données.
5. Évaluer comment les dépôts de données soutiennent les principes FAIR.
6. Trouver les communautés ou les initiatives qui utilisent les principes FAIR dans leur écosystème de gestion des données de recherche.

## Introduction

L'écosystème de données actuel est complexe et en pleine croissance. Les **principes FAIR** (Facile à trouver, Accessible, Interopérable, Réutilisable) sont des principes directeurs qui encouragent les **personnes qui s'occupent de l'intendance des données** à améliorer la repérabilité et la réutilisation des **données de recherche** par les systèmes informatiques. Dans ce chapitre, nous explorerons l'étendue des principes et des

outils utilisés pour évaluer et améliorer la conformité aux principes FAIR d'un jeu de données. Nous discuterons également de l'impact de ces principes et nous examinerons de quelle façon ils ont été adoptés.

# Un petit historique des principes FAIR

## Pourquoi avons-nous besoin de principes directeurs pour les données de recherche?

La nécessité d'établir des exigences pour la **gestion des données de recherche** (GDR) a d'abord été proposée par quelques organismes nationaux de financement de la recherche en Europe, en raison de l'essor de la science fondée sur l'utilisation intensive de données. Depuis, les exigences en matière de **plans de gestion des données**, de citations de données et de disponibilité des données sont toutes devenues essentielles à la conduite responsable de la recherche et elles ont introduit de nouvelles conditions auxquelles doivent adhérer les chercheuses et chercheurs qui souhaitent publier ou recevoir du financement public (Hrynaszkiewicz *et al*., 2020). Les personnes qui s'occupent de l'intendance des données ont aidé les chercheuses et chercheurs à répondre aux exigences associées à la GDR en encourageant la préservation des données, en fournissant des formations sur la préparation des données et en développant des infrastructures pour le stockage sécuritaire des données. Alors que les progrès dans les infrastructures technologiques ont rendu possible l'analyse computationnelle d'importantes quantités de données, l'augmentation correspondante du nombre de dépôts de données et de normes créées pour diffuser les données dans différentes disciplines et différents secteurs a contribué à la création de silos. Cette situation empêche le regroupement des données en vue de produire des recherches utiles. Conséquemment, l'élaboration de principes plus larges pour encadrer le partage responsable de données est devenue de plus en plus nécessaire pour les différents membres de la communauté élargie de données de recherche.

## Les origines des principes directeurs FAIR

C'est en 2014, lors d'une « non-conférence » tenue aux Pays-Bas intitulée *Jointly Designing a Data FAIRport* que les principes fondamentaux pour les données de recherche **interopérables** ont été abordés pour la première fois (Data FAIRport, 2014). L'année suivante, une ébauche de guide a été rédigée et publiée par le FAIR Data Publishing Group de la coalition FORCE11 afin de recevoir des commentaires et l'approbation de la communauté. (FORCE11, 2014a). En 2016, Barend Mons avec un groupe de contributrices et contributeurs a rédigé un article dans *Scientific Data* pour discuter de la nécessité d'établir des principes directeurs FAIR pour les actifs numériques (Wilkinson *et al*., 2016). Ces principes sont conçus pour aider les

humains et les machines à surmonter les obstacles liés à la découvrabilité, l'accessibilité, la réutilisation et la citation des données de recherche.

Depuis cette première publication, une version des principes FAIR a été mise à jour par GO FAIR (https://www.go-fair.org/fair-principles/) (en anglais uniquement). Avec le temps, ces principes ont influencé non seulement les chercheuses et chercheurs qui souhaitent préparer leurs données au partage, mais aussi les dépôts de données qui souhaitent évaluer et améliorer leur infrastructure, en plus d'autres parties prenantes qui souhaitent évaluer et améliorer leurs politiques pour soutenir un écosystème de données FAIR.

# Que sont les principes directeurs FAIR?

## Les principes directeurs FAIR

L'objectif premier de ces principes est d'assurer que les machines et les humains peuvent facilement trouver, accéder, interopérer avec et réutiliser de façon appropriée la grande quantité d'informations disponibles à des fins scientifiques. Ils se veulent des principes de haut niveau et indépendants du domaine, ce qui veut dire qu'ils ont une large portée et peuvent être appliqués à différents types de données à travers une multitude de disciplines. En évitant l'attribution de spécifications techniques, les principes directeurs FAIR permettent différentes mises en œuvre pour les normes et caractéristiques de gestion des données qu'ils proposent.

Les principes FAIR énumérés ci-dessous ne sont qu'un aperçu de la liste plus complète des principes et sous-catégories, disponible ici (en anglais uniquement) : https://www.go-fair.org/fair-principles/ (https://www.go-fair.org/fair-principles/):

**Faciles à trouver**

Les humains et les systèmes informatiques doivent pouvoir trouver les données et les **métadonnées**. Les **métadonnées lisibles par machine** sont essentielles à la découverte automatique de jeux de données et de services.

**F1.** Les (méta)données sont assorties d'un **identifiant unique pérenne** à l'échelle internationale (IUP).

**F2.** Les données sont décrites au moyen de métadonnées riches (tel que défini par R1 ci-dessous).

**F3.** Les métadonnées incluent clairement et de façon explicite l'identifiant des données qu'elles décrivent.

**F4.** Les (méta)données sont enregistrées et indexées dans une ressource recherchable.

**Accessibles**

Une fois que la personne utilisatrice a trouvé les données, elle doit savoir comment y accéder, ce qui peut nécessiter des détails relatifs à l'authentification ou l'autorisation.

**A1.** Les (méta)données sont récupérables par leur identifiant au moyen d'un protocole de communication normalisé.

**A1.1** Le protocole est ouvert, gratuit et il est possible de l'implémenter de manière universelle.

**A1.2** Le protocole permet une procédure d'authentification et d'autorisation lorsque requis.

**A2.** Les métadonnées sont accessibles, même quand les données ne le sont plus.

**Interopérables**

Les données ont généralement besoin d'être intégrées à d'autres données et doivent interopérer avec des applications ou des processus de travail pour permettre l'analyse, le stockage et le traitement.

**I1.** Les (méta)données utilisent un langage formel, accessible, partagé et applicable globalement à des fins de représentation de la connaissance.

**I2.** Les (méta)données utilisent des vocabulaires qui adhèrent aux principes FAIR.

**I3.** Les (méta)données comprennent des références qualifiées aux autres (méta)données.

**Réutilisables**

L'objectif principal de FAIR est de maximiser la réutilisation des données, donc les données et métadonnées doivent être bien décrites pour qu'elles puissent être reproduites et/ou combinées dans différents contextes.

**R1.** Les (méta)données ont une pluralité d'attributs précis et pertinents.

**R1.1.** Les (méta)données sont diffusées selon une licence d'utilisation claire et accessible.

**R1.2.** Les (méta)données sont associées à une provenance détaillée.

**R1.3.** Les (méta)données se conforment aux normes de leurs communautés respectives.

Dans le chapitre 10, « Soutenir la recherche reproductible avec la curation active de données, » vous en saurez plus sur les mesures à appliquer pour rendre les données interopérables et réutilisables dans le cadre de la curation active des données.

## Les mécanismes clés des principes directeurs FAIR: les métadonnées, les identifiants pérennes et les licences

L'utilisation appropriée des métadonnées (des informations sur les données) est fondamentale aux principes FAIR. Au même titre que le matériel de recherche traditionnel (comme les livres et articles avec des informations bibliographiques), les données de recherche doivent être décrites de façon structurée en utilisant des **vocabulaires contrôlés** qui peuvent être lisibles pour les humains et les machines, permettant ainsi aux données d'être repérées et réutilisées. Ainsi, les métadonnées représentent une partie intégrante des données de recherche parce qu'elles fournissent aux utilisatrices et utilisateurs d'importantes informations au sujet d'un jeu de données telles que sa documentation, ses identifiants, ses licences et autres éléments applicables. Les métadonnées qui décrivent les données de recherche originales devraient être riches et assez précises pour permettre aux humains et aux machines de comprendre le contexte et les limites d'un jeu de données, mais elles devraient aussi être offertes par le biais de descriptions normalisées pour que l'interprétation des données de recherche puisse se faire à travers différents domaines. Pour réussir cet équilibre, des chercheuses et chercheurs d'une variété de disciplines ont adopté des normes bien élaborées pour les métadonnées, telles que celles énumérées par le Research Data Alliance (RDA) (https://rdamsc.bath.ac.uk) (en anglais uniquement).

Les autres mécanismes qui garantissent la repérabilité et la réutilisation des données sont les identifiants uniques pérennes (IUP ou *Persistent IDentifier*, PID) et les licences qui encadrent la façon dont les données peuvent être utilisées. Un IUP enregistré fournit à chaque jeu de données et ses métadonnées un moyen d'identification stable et unique qui permet de suivre tout changement ou mouvement en ligne. Les chercheuses et chercheurs qui partagent des données sur leur propre site Web ne peuvent généralement pas attribuer de tels identifiants et sont plutôt encouragés à déposer leurs données dans des dépôts de données pour avoir accès à du soutien en matière d'utilisation d'identifiants uniques pérennes, dont les **identifiants numériques d'objets** (DOI) (p. ex., https://doi.org/10.1000/182 (https://doi.org/10.1000/182)).

Plusieurs chercheuses et chercheurs s'inquiètent que leurs données soient mal utilisées et hésitent à les partager (Wiley *et al*., 2019, p.5). Les personnes qui utilisent des données, quant à elles, n'arrivent souvent pas à réutiliser et repartager en toute confiance les données secondaires issues d'un jeu de données de recherche original en raison d'un manque de clarté lié aux permissions de réutilisation des données. Pour remédier à ce problème, des licences de données standard, telles que les licences Creative Commons (http://www.creativecommons.org/), les licences Open Data Commons (http://opendatacommons.org/), ou des ententes

personnalisées d'utilisation des données peuvent encourager la réutilisation des données tout en protégeant les droits des gens qui les ont créées en matière de crédit et d'attribution. La licence fournit des informations sur l'utilisation légale et éthique des données permettant ainsi de définir les modalités de la relation entre les personnes qui créent, éditent et utilisent un même jeu de données. Vous en saurez davantage sur les licences dans le chapitre 12, « Planification de la gestion des données pour les processus de travail en science ouverte. »

## Les principes FAIR et l'ouverture des données

Les efforts mis en place pour rendre les données FAIR n'impliquent pas forcément le partage ouvert et sans restriction des données. Par exemple, les jeux de données peuvent comporter des identifiants pérennes et des métadonnées FAIR sans qu'ils puissent être réutilisés en raison des conditions de leur licence. Le *Fair Principles Working Detailed Document* propose quatre niveaux de conformité aux principes FAIR pour des jeux de données à l'intérieur d'un dépôt de données. Ces niveaux décrivent les différents degrés potentiels d'accès aux données :

1. Chaque jeu de données comporte un identifiant unique pérenne et offre des métadonnées FAIR.
2. Chaque jeu de données comporte des métadonnées définies pour les gens qui utilisent les données afin de fournir des informations riches sur la provenance.
3. Les éléments de données à l'intérieur des jeux de données sont FAIR mais ne sont pas en **libre accès** et comportent des restrictions précises en matière de réutilisation.
4. Les jeux de données et éléments de données sont FAIR, ouverts au public et comportent des licences bien définies (FORCE11, 2014b).

Les principes directeurs FAIR permettent aux responsables de l'intendance des données de participer aux importantes décisions en matière de publication et offrent la possibilité de recourir à d'autres principes. Par exemple, les principes CARE (avantages Collectifs, Autorité pour contrôler, Responsabilité et Éthique) pour la gouvernance des données autochtones, publiés en 2019 par la Global Indigenous Data Alliance, reconnaissent l'importance de la souveraineté des données autochtones et de centraliser les droits et intérêts des peuples autochtones dans le traitement des données autochtones. À plusieurs égards, les principes CARE et FAIR se complètent et incitent les chercheuses et chercheurs à tenir compte de la variété des personnes participantes et des objectifs associés aux données de recherche. La **souveraineté des données autochtones** est abordée plus en détail dans le chapitre 3.

# Comment rendre vos données FAIR: outils et conseils

## Les principes directeurs FAIR et les plans de gestion des données

Certaines opportunités de financement exigent la rédaction de plans de gestion des données (PGD), conformément à la **Politique des trois organismes sur la gestion des données de recherche** (Gouvernement du Canada, 2021). Dans ces PGD, les chercheuses et chercheurs doivent décrire leurs méthodologies et stratégies en tenant compte des principes directeurs FAIR. Par exemple, les chercheuses et chercheurs devraient documenter les données de façon efficace dès les premières étapes d'un projet pour que des métadonnées complètes et de grande qualité puissent être générées pour la diffusion. En outre, pour déposer et préserver leurs données dans des dépôts qui adhèrent aux principes directeurs FAIR, les chercheuses et chercheurs devraient négocier des licences pour le partage de données avec les différentes parties prenantes et obtenir tôt dans le processus de collecte la permission des personnes qui participent à la recherche de partager les données.

## Des outils destinés à la communauté de recherche pour évaluer et améliorer la conformité aux principes FAIR

Une variété d'outils a été développée pour aider les chercheuses et chercheurs à comprendre les principes FAIR et à mettre en œuvre des pratiques qui s'alignent sur ces principes. Ces outils vont de simples listes de contrôle à des ressources personnalisées conçues en fonction des pratiques des chercheuses et chercheurs. Vous trouverez ci-dessous une liste d'outils d'évaluation disponibles ou en développement dotés de différentes caractéristiques selon qui les utilise. Nous recommandons d'utiliser ces outils lorsque vous préparez vos données pour les rendre FAIR.

1. La liste de contrôle *How FAIR Are Your Data* (https://zenodo.org/record/5111307#.Yj3Vi5rMI-Q)? (Jones et Grootveld, 2017)

Développée par un réseau de service de données européen, il s'agit simplement d'une liste d'une page basée sur les principes directeurs FAIR avec de légères modifications qui rendent les concepts et la terminologie plus faciles d'accès pour les chercheuses et chercheurs. Cette liste de contrôle constitue un bon outil d'initiation pour les chercheuses et chercheurs qui travaillent depuis peu dans le domaine de la GDR.

2. Le *FAIR Data Self Assessment Tool (https://ardc.edu.au/resources/aboutdata/fair-data/fair-self-assessment-tool/)* (Australian Research Data Commons, 2022)

Cet outil a été développé par le Australian Research Data Commons. En répondant à des questions associées aux principes directeurs FAIR, les chercheuses et chercheurs peuvent évaluer la conformité de leurs pratiques par rapport à chacun des sous-principes FAIR ainsi que de déterminer leur conformité dans le contexte plus large des quatre principes FAIR. L'outil permet de comparer leurs méthodes de traitement des données avec les meilleures pratiques leur permettant ainsi d'identifier ce qui peut être amélioré.

3. L'outil *FAIR Aware (https://doranum.fr/appli-fair-aware-pleine-page-vf/)* (Data Archiving and Networked Services, 2021) traduit et rendu disponible en français par DoRANum

Élaboré par le Data Archiving and Networked Services des Pays-Bas, l'outil *FAIR Aware* permet une évaluation plus détaillée pour aider les chercheuses et chercheurs à comprendre et à mieux mettre en œuvre les principes FAIR. La chercheuse ou le chercheur doit identifier son domaine de recherche, son rôle et l'organisme associé, mais le contenu de l'outil d'évaluation est le même, peu importe qui l'utilise. Les chercheuses et chercheurs doivent répondre à 10 questions en lien avec chacun des principes directeurs FAIR puis évaluer leur volonté de se conformer aux pratiques recommandées. Une fois les réponses soumises, un rapport est fourni et celui-ci donne non seulement un aperçu du niveau de connaissance des principes FAIR de la chercheuse ou du chercheur, mais aussi des conseils et des ressources pour les aspects à améliorer.

4. Le *F-UJI Automated FAIR Data Assessment Tool (https://www.f-uji.net/)* (Devaraju et Huber, 2020)

L'outil F-UJI (*FAIRsFAIR Research Data Object Assessment Service*) est conçu pour évaluer la conformité aux principes FAIR des jeux de données de recherche à partir de paramètres mesurables complets et détaillés établis par le projet FAIRsFAIR (Devaraju *et al*., 2020).

## Des conseils supplémentaires pour rendre vos données FAIR

Outre les outils d'évaluation, des services de données de recherche au niveau international et national ont élaboré des lignes directrices pour rendre les données FAIR, tant de façon générale que pour des disciplines particulières. En voici quelques exemples :

- OpenAIRE (un organisme qui soutient le développement de la science ouverte en Europe) a créé des guides pour les chercheuses et chercheurs, dont *How to make your data FAIR (https://www.openaire.eu/how-to-make-your-data-fair)* (OpenAIRE, s.d.);

- *How to FAIR* (https://www.howtofair.dk/) (Danish National Forum for Research Data Management, s.d.) élaboré par le biais d'entrevues avec un groupe composé d'une variété de chercheuses, chercheurs et de bibliothécaires;
- *Top 10 FAIR Data & Software Things* (https://librarycarpentry.org/Top-10-FAIR/) (Library Carpentry, s.d.) offre de petits guides autonomes sur une variété de sujets et de disciplines qui peuvent être utilisés par les membres de la communauté de recherche (p. ex., l'astronomie, l'imagerie médicale, la musique, etc.);
- *Sustainable and FAIR Data Sharing in the Humanities* (https://allea.org/portfolio-item/sustainable-and-fair-data-sharing-in-the-humanities/) (ALLEA, 2020) fournit des conseils pratiques pour les chercheuses et chercheurs qui souhaitent rendre plus FAIR leurs données numériques en sciences humaines.

Au Canada, les chercheuses et chercheurs de l'Institut de cardiologie de l'Université d'Ottawa et de l'Institut de recherche de l'Hôpital d'Ottawa ont développé une série de cours sur le traitement des données, dont *FAIR Principles (https://journalologytraining.ca/courses/fair-principles/)* (Centre for Journalology, s.d.). Le contexte canadien n'offre pas beaucoup plus en matière de conseils sur les principes FAIR. Pour les chercheuses, chercheurs et bibliothécaires intéressés par ce domaine, il est possible de consulter le guide *How to Be FAIR with Your Data : A Teaching and Training Handbook for Higher Education Institutions* (Engelhardt *et al.*, 2022) pour des exemples de formations en lien avec les principes FAIR offertes par différents établissements européens d'enseignement supérieur.

## Les impacts politiques des principes FAIR

Les principes FAIR ont été utilisés par des agences gouvernementales, des établissements universitaires, des organismes de financement de la recherche, des sociétés savantes, des maisons d'édition et de nombreuses autres parties prenantes pour mettre en valeur l'importance de l'intendance des données de recherche tant au niveau culturel, économique et social. Conséquemment, les principes sont devenus fondamentaux pour les structures organisationnelles qui cherchent à influencer les chercheuses et chercheurs dans leurs méthodes de gestion et de partage des données. Certains exemples d'impacts politiques comprennent la Commission européenne qui cite les principes FAIR comme ayant directement influencé le développement du nuage européen pour la science ouverte (Hill, 2019, p. 284), en plus du National Institutes of Health des États-Unis qui cite l'application des principes FAIR pour les données dans leur *Data Management and Sharing Policy* (National Institutes of Health, 2020).

Au Canada, une des principales recommandations du gouvernement dans sa *Feuille de route pour la science ouverte* (Bureau du conseiller scientifique en chef du Canada, 2020) est à la mise en œuvre des principes FAIR pour les agences et départements fédéraux. Ce plan vise à assurer la mise en œuvre complète, à partir de janvier

2025, de l'interopérabilité des données scientifiques et de recherches ainsi que des normes de métadonnées de toutes les données liées aux agences et départements gouvernementaux. En matière de subventions de recherche, la Politique des trois organismes pour la gestion des données de recherche stipule que les trois agences fédérales de financement de la recherche du Canada – les Instituts de recherche en santé du Canada (IRSC), le Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG) et le Conseil de recherches en sciences humaines du Canada (CRSH) – appuient les principes FAIR et s'attendent à ce que les chercheuses et chercheurs partagent leurs données conformément à ces principes et aux normes propres à leur discipline, sans enfreindre d'exigences éthique, culturelle, juridique ou commerciale (2021). De plus, les maisons d'édition canadiennes de publications savantes, comme les Éditions Sciences Canada (s.d.), s'inspirent des efforts d'autres maisons d'édition spécialisées qui utilisent les principes FAIR pour circonscrire le contenu de leur politique d'accès aux données. La conformité à ces politiques peut impliquer l'utilisation d'outils pour les chercheuses et chercheurs – comme ceux énumérés plus tôt – afin d'assurer que les données soient aussi alignées que possible sur les principes FAIR avant d'être diffusées. Outre la préparation des données, ces exigences cherchent également à influencer l'opinion de la communauté de recherche quant au choix d'un dépôt de données de recherche et la façon dont ce choix pourra maintenir la conformité aux principes FAIR au-delà de la publication initiale des données.

## Les principes FAIR et les dépôts

Les principes FAIR permettent de reconnaître la valeur actuelle et potentielle des dépôts de données. Wilkinson *et al*. (2016) font la promotion de cette idée en discutant des avantages et des limites des dépôts de données de recherche. Ils soutiennent aussi que les dépôts devraient évoluer en fonction des besoins des chercheuses et chercheurs au niveau de la découvrabilité et de la réutilisation (p. 2-4). La chercheuse ou le chercheur doit déterminer si un dépôt de données répond aux besoins spécifiques en GDR de sa discipline, si celui-ci lui permet de se conformer aux exigences éthiques et juridiques pertinentes et si les fonctionnalités offertes reflètent les principes FAIR.

Les dépôts de données de recherche sont des contenants de données spécialement conçus pour stocker des données de recherche, des fichiers associés et des métadonnées afin de permettre un accès stable et à long terme aux données (Boyd, 2021, p. 25-26). Les dépôts constituent des éléments cruciaux de l'infrastructure numérique; ils sont établis pour encourager la découvrabilité des données de recherche et pour aider les chercheuses et chercheurs à publier et à diffuser leurs données. Choisir un dépôt plutôt qu'un autre dépendra souvent de facteurs tels que les normes disciplinaires, les exigences des maisons d'édition ou des organismes subventionnaires ou les lignes directrices en matière de partage des données. Le choix peut aussi se faire selon d'autres éléments comme la simplicité ou l'aspect pratique du processus de dépôt des données, le type de fichiers acceptés, le soutien offert pour la curation des données ou les schémas de métadonnées et les vocabulaires contrôlés que le dépôt utilise pour décrire les jeux de données de recherche qui sont stockés. La

prise en compte de tous ces éléments incitera la chercheuse ou le chercheur à choisir soit un dépôt propre à sa discipline, soit à sa communauté ou un dépôt plus généraliste. La chercheuse ou le chercheur pourra ensuite déterminer si le dépôt de son choix met en pratique les principes FAIR en évaluant la présence ou l'absence de certaines fonctionnalités particulières.

Dans leur article sur l'amélioration de l'interopérabilité entre différents types de dépôts, Hahnel et Valen (2020) soutiennent que pour fonctionner conformément aux principes FAIR, un dépôt doit appliquer les directives suivantes :

- attribuer des identifiants uniques pérennes (DOI, ORCID et GRID) à ses données et au matériel connexe;
- fournir avec ses données des **interfaces de programmation d'applications** (API) documentés;
- soutenir des options robustes pour la curation des données et souscrire à des lignes directrices en matière d'accessibilité Web;
- offrir des licences bien définies qui appuient la réutilisation des données;
- décrire sa démarche vers la durabilité en documentant les processus de travail en matière de préservation et de reprise après sinistre (p. 195-197).

Ces conseils à propos des caractéristiques optimales pour les dépôts s'inspirent de recommandations semblables émises par OpenAIRE et par l'initiative FAIR Sharing (Cannon *et al*., 2021). Certains de ces éléments se retrouvent également dans les principes TRUST pour les dépôts numériques, publiés par Lin *et al*. (2020).

Pour évaluer la façon dont certains dépôts de données importants au Canada et à l'international ont documenté leur engagement envers les principes FAIR, consultez les exemples suivants :

- Dépôt fédéré de données de recherche: https://www.frdr-dfdr.ca/docs/fr/principes_fair/ (https://www.frdr-dfdr.ca/docs/fr/principes_fair/);
- Zenodo (en anglais uniquement): https://about.zenodo.org/principles/ (https://about.zenodo.org/principles/);
- Figshare (en anglais uniquement): https://knowledge.figshare.com/publisher/fair-figshare (https://knowledge.figshare.com/publisher/fair-figshare).

Il est aussi possible de trouver des dépôts appropriés en consultant le répertoire re3data (https://www.re3data.org), un outil multidisciplinaire qui liste plus de 2800 entrées de dépôts de données. Il est possible d'effectuer des recherches dans le répertoire re3data selon des critères particuliers tels que le type d'API et les normes de métadonnées. Un autre répertoire intéressant est celui de FAIRsharing (https://fairsharing.org/databases/), appuyé par le Research Data Alliance. Ce dernier fournit une plateforme multidisciplinaire où les

chercheuses et chercheurs peuvent consulter les entrées pour trouver des dépôts, des normes de données et des politiques de données. Ces deux outils sont d'excellentes options pour trouver des dépôts en fonction des disciplines particulières.

Certains dépôts plus importants à vocation commerciale, communautaire ou éditoriale peuvent offrir des options plus flexibles et spécialisées qui s'alignent sur l'orientation FAIR. Toutefois, en choisissant un dépôt, il faut évaluer s'il permet d'adhérer à certaines normes disciplinaires, s'il donne accès au soutien nécessaire pour la conformité aux exigences éthiques ou juridiques et s'il aide à assumer certaines responsabilités envers des communautés qui ont des attentes en matière d'accès à leurs données. Le choix d'un dépôt plutôt qu'un autre devrait se faire en équilibrant la conformité aux principes FAIR contre ces autres exigences toutes aussi importantes.

## Pour s'impliquer

Pour les personnes qui souhaitent appuyer la mise en œuvre plus large des principes FAIR, l'initiative GO FAIR rassemble des individus, des établissements et des organismes pour collaborer à l'élaboration de politiques, au développement de compétences et à l'élaboration de normes techniques et de technologies. Cette collaboration est réalisée par le biais des réseaux de mise en œuvre GO FAIR qui rassemblent des partenaires qui travaillent à soutenir la création de produits uniques. Pour en savoir plus sur les réseaux de mise en œuvre ou sur comment en faire partie, consultez https://www.go-fair.org/implementation-networks/ (https://www.go-fair.org/implementation-networks/) (en anglais uniquement).

## Conclusion

Les principes FAIR ont clarifié les mesures à prendre pour réaliser certains objectifs du mouvement de GDR. Ces principes, combinés à d'autres principes directeurs, ont été adoptés par les organismes subventionnaires, les maisons d'édition et par une variété de communautés de recherche. Ils ont contribué à rassembler et à harmoniser les efforts pour soutenir l'accès et la réutilisation des données. Les chercheuses et chercheurs devraient continuer à suivre l'évolution des principes FAIR en ce qui a trait à leur influence non seulement sur l'écosystème des données de recherche au niveau national et international, mais aussi sur la réutilisation des données dans leur propre discipline.

## Questions de réflexion

1. Utilisez l'outil *FAIR Aware* pour évaluer vos connaissances et vos compétences pour rendre les données FAIR.
2. En utilisant les principes FAIR comme référence, évaluez la conformité aux principes FAIR des jeux de données suivants et faites des suggestions pour améliorer leur conformité :

    1. *Don Valley Historical Mapping Project*: https://doi.org/10.5683/SP2/PONAP6 (https://doi.org/10.5683/SP2/PONAP6)
    2. *Soil and Plant Phytoliths from the Acacia-Commiphora Mosaics at Olduvai Gorge (Tanzania)*: https://doi.org/10.20383/101.0122 (https://doi.org/10.20383/101.0122)
    3. *CLOUD: Canadian Longterm Outdoor UAV Dataset*: https://www.dynsyslab.org/cloud-dataset (https://www.dynsyslab.org/cloud-dataset)

*Un élément interactif H5P a été exclu de cette version du texte. Vous pouvez le consulter en ligne ici : https://ecampusontario.pressbooks.pub/gdrcanada/?p=27#h5p-5 (https://ecampusontario.pressbooks.pub/gdrcanada/?p=27#h5p-5)*

## Éléments clés à retenir

- Les principes directeurs FAIR représentent des objectifs de haut niveau qui orientent l'optimisation continue des données de recherche, des métadonnées et des environnements de publication des données pour faciliter l'accès et la réutilisation des données à travers

différents domaines grâce aux identifiants uniques pérennes, aux métadonnées riches et normalisées ainsi qu'aux licences.
- Les chercheuses et chercheurs peuvent suivre des orientations ou utiliser des outils pour en savoir plus sur les principes FAIR, pour évaluer leurs pratiques actuelles de GDR et pour planifier des stratégies pour rendre plus FAIR leurs données de recherche et leurs activités de publication.
- Les principes FAIR ont influencé les politiques liées à la disponibilité des données tant au niveau des gouvernements, des organismes de financement en recherche et des maisons d'édition.
- Les chercheuses et chercheurs peuvent rendre leur gestion des données et leurs activités de partage plus conformes aux principes FAIR en s'assurant de choisir des dépôts de données qui offrent des caractéristiques alignées sur ces principes.
- Les registres des dépôts de données de recherche constituent des outils importants pour identifier les dépôts qui offrent des options conformes aux principes FAIR et aux normes de certaines disciplines ou autres exigences de nature juridique, éthique et/ou communautaire.

# Lectures et ressources supplémentaires

## Les principes FAIR et CARE

The Global Indigenous Data Alliance. (2019). *CARE principles for Indigenous data governance*. https://www.gida-global.org/care (https://www.gida-global.org/care)

GO FAIR. (s.d.). *FAIR principles*. https://www.go-fair.org/fair-principles/ (https://www.go-fair.org/fair-principles/)

Research Data Alliance. (s.d.). *Metadata standards catalogue*. https://rdamsc.bath.ac.uk/ (https://rdamsc.bath.ac.uk/)

## Les principes et les dépôts FAIR

Le répertoire re3data. https://www.re3data.org (https://www.re3data.org)

Le répertoire FAIRsharing. https://fairsharing.org/databases/ (https://fairsharing.org/databases/)

## Pour s'impliquer

La mise en œuvre GO FAIR. https://www.go-fair.org/implementation-networks/ (https://www.go-fair.org/implementation-networks/)

# Bibliographie

ALLEA. (2020). *Sustainable and FAIR data sharing in the humanities: Recommendations of the ALLEA working group e-humanities*. Digial Repository of Ireland. https://doi.org/10.7486/DRI.TQ582C863 (https://doi.org/10.7486/DRI.TQ582C863)

Australian Research Data Commons. (2022). *FAIR data self assessment tool*. https://ardc.edu.au/resources/aboutdata/fair-data/fair-self-assessment-tool/ (https://ardc.edu.au/resources/aboutdata/fair-data/fair-self-assessment-tool/)

Boyd, C. (2021). Understanding research data repositories as infrastructures. *Proceedings of the Association for Information Science and Technology, 58*(1), 25–35. https://doi.org/10.1002/pra2.433 (https://doi.org/10.1002/pra2.433)

Bureau du conseiller scientifique en chef du Canada. (2020). *Feuille de route pour la science ouverte*. Gouvernement du Canada. https://science.gc.ca/site/science/fr/bureau-conseillere-scientifique-chef/science-ouverte/feuille-route-pour-science-ouverte (https://science.gc.ca/site/science/fr/bureau-conseillere-scientifique-chef/science-ouverte/feuille-route-pour-science-ouverte)

Cannon, M., Graf, C., McNeice, K., Chan, W. M., Callaghan, S., Carnevale, I., Cranston, I., Edmunds, S. C., Everitt, N., Ganley, E., Hrynaszkiewicz, I., Khodiyar, V. K., Leary, A., Lemberger, T., MacCallum, C. J., Murray, H., Sharples, K., Soares E Silva, M., Wright, G., ... (Moderator) Sansone, S-A. (2021). *Repository features to help researchers: An invitation to a dialogue*. Zenodo. https://doi.org/10.5281/zenodo.4683794 (https://doi.org/10.5281/zenodo.4683794)

Centre for Journalology. (s.d.). *FAIR principles*. https://journalologytraining.ca/courses/fair-principles/ (https://journalologytraining.ca/courses/fair-principles/)

Danish National Forum for Research Data Management. (s.d.). *How to FAIR*. https://www.howtofair.dk/ (https://www.howtofair.dk/)

Data Archiving and Networked Services. (2021). *FAIR Aware*. https://fairaware.dans.knaw.nl/ (https://fairaware.dans.knaw.nl/)

Data FAIRport. (2014). *Data FAIRport conference: Jointly designing a data FAIRport*. https://www.datafairport.org/component/content/article/8_news/9_item1/index.html (https://www.datafairport.org/component/content/article/8_news/9_item1/index.html)

Devaraju, A. et Huber, R. (2020). *F-UJI – An Automated FAIR Data Assessment Tool (v1.0.0)*. Zenodo. https://doi.org/10.5281/zenodo.4063720 (https://doi.org/10.5281/zenodo.4063720)

Devaraju, A., Huber, R., Mokrane, M., Herterich, P., Cepinskas, L., de Vries, J., L'Hours, H., Davidson, J. et Angus W. (2020). *FAIRsFAIR data object assessment metrics (0.5)*. Zenodo. https://doi.org/10.5281/zenodo.6461229 (https://doi.org/10.5281/zenodo.6461229)

Éditions Science Canada (s.d.). *Principes et politique relatifs à la disponibilité des données*. https://cdnsciencepub.com/pb-assets/resources/csp/About_Data_Principles_and_Policy_F-1601768276523.pdf (https://cdnsciencepub.com/pb-assets/resources/csp/About_Data_Principles_and_Policy_F-1601768276523.pdf)

Engelhardt, C., Biernacka, K., Coffey, A., Cornet, R., Danciu, A., Demchenko, Y., Downes, S., Erdmann, C., Garbuglia, F., Germer, K., Helbig, K., Hellström, M., Hettne, K., Hibbert, D., Jetten, M., Karimova, Y., Kryger Hansen, K., Kuusniemi, M. E., Letizia, V., McCutcheon, V., ... Zhou, B. (2022). *D7.4 How to be FAIR with your data. A teaching and training handbook for higher education institutions*. Zenodo. https://doi.org/10.5281/ZENODO.5665492 (https://doi.org/10.5281/ZENODO.5665492)

FORCE11. (2014a, 1 septembre). *FAIR data publishing group*. Archived groups. https://force11.org/group/fair-data-publishing-group/ (https://force11.org/group/fair-data-publishing-group/)

FORCE11. (2014b, 10 septembre). *Guiding principles for findable, accessible, interoperable and re-usable data publishing version b1. 0*. https://force11.org/info/guiding-principles-for-findable-accessible-interoperable-and-re-usable-data-publishing-version-b1-0/ (https://force11.org/info/guiding-principles-for-findable-accessible-interoperable-and-re-usable-data-publishing-version-b1-0/)

Gouvernement du Canada. (2021). *Politique des trois organismes sur la gestion des données de recherche.* https://science.gc.ca/site/science/fr/financement-interorganismes-recherche/politiques-lignes-directrices/gestion-donnees-recherche/politique-trois-organismes-gestion-donnees-recherche (https://science.gc.ca/site/science/fr/financement-interorganismes-recherche/politiques-lignes-directrices/gestion-donnees-recherche/politique-trois-organismes-gestion-donnees-recherche)

Hahnel, M. et Valen, D. (2020). How to (easily) extend the FAIRness of existing repositories. *Data Intelligence, 2*(1–2), 192–198. https://doi.org/10.1162/dint_a_00041 (https://doi.org/10.1162/dint_a_00041)

Hill, T. (2019). Turning FAIR into reality: Review. *Learned Publishing, 32*(3), 283–286. https://doi.org/10.1002/leap.1234 (https://doi.org/10.1002/leap.1234)

Hrynaszkiewicz, I., Simons, N., Hussain, A., Grant, R. et Goudie, S., 2020. Developing a research data policy framework for all journals and publishers. *Data Science Journal, 19*(1), 1-15. http://doi.org/10.5334/dsj-2020-005 (http://doi.org/10.5334/dsj-2020-005)

Jones, S. et Grootveld, M. (2017). *How FAIR are your data?* Zenodo. https://doi.org/10.5281/ZENODO.1065990 (https://doi.org/10.5281/ZENODO.1065990)

Library Carpentry. (s.d.). *Top 10 FAIR data & software things*. Zenodo. https://doi.org/10.5281/zenodo.2555498 (https://doi.org/10.5281/zenodo.2555498)

Lin, D., Crabtree, J., Dillo, I. Downs R. R., Edmunds R., Giaretta, D., De Giusti, M., L'Hours, H., Hugo, W., Jenkyns, R., Khodiyar, V., Martone, M. E., Mokrane, M., Navele, V., Petters, J., Sierman, B., Sokolova, D. V., Stockhause, M. et Westbrook, J. (2020). The TRUST principles for digital repositories. *Scientific Data, 7*, 144. https://doi.org/10.1038/s41597-020-0486-7 (https://doi.org/10.1038/s41597-020-0486-7)

National Institutes of Health. (2020). *Final NIH policy for data management and sharing*. https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html (https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html)

OpenAIRE. (s.d.). *How to make your data FAIR*. https://www.openaire.eu/how-to-make-your-data-fair (https://www.openaire.eu/how-to-make-your-data-fair)

Wiley, C. A. et Burnette, M. H., (2019). Assessing data management support needs of bioengineering and biomedical research faculty. *Journal of eScience Librarianship, 8*(1), 1-19. https://doi.org/10.7191/jeslib.2019.1132 (https://doi.org/10.7191/jeslib.2019.1132)

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., ... Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data, 3*(1), 160018. https://doi.org/10.1038/sdata.2016.18 (https://doi.org/10.1038/sdata.2016.18)

## À propos des auteurs

### Minglu Wang

Minglu Wang est une bibliothécaire en gestion des données de recherche (GDR) à l'Université York. Elle a publié des articles de recherche, des chapitres de livre, des documents de travail et des communications lors de colloques au sujet des bibliothèques universitaires et des services de GDR. Mme Wang est une membre active de l'Association of College & Research Libraries (ACRL), une division de l'American Library Association. Pendant plusieurs années, elle a rédigé des articles et des livres blancs pour les publications *Top Trends* et *Environmental Scan* de l'ACRL. Elle fait partie du Groupe d'experts sur la recherche et l'intelligence de l'équipe de GDR de l'Alliance de recherche numérique du Canada. Elle a participé à la conception et à la rédaction du rapport sur le sondage sur les capacités en GDR des établissements canadiens. Courriel : mingluwa@yorku.ca (denied:about:blank) | ORCID : 0000-0002-0021-5605 (https://orcid.org/0000-0002-0021-5605)

## Dany Savard

Dany Savard est bibliothécaire associé pour les collections et les services de recherche à la bibliothèque de l'Université de Toronto à Mississauga. Il a contribué à des articles de recherche sur les thèmes de la découverte des données de recherche et des dépôts de données. Il est membre du Groupe d'experts sur la découverte et les métadonnées du réseau d'expertes et experts de l'Alliance de recherche numérique du Canada et préside actuellement le groupe de travail sur le paysage des dépôts de données canadiens. Il est titulaire d'un MLIS de l'Université Western et d'une maîtrise en politique publique et en administration de l'Université métropolitaine de Toronto. Courriel : dany.savard@utoronto.ca (mailto:dany.savard@utoronto.ca) | ORCID : 0000-0001-7472-7390 (https://orcid.org/0000-0001-7472-7390)

3.

# SOUVERAINETÉ DES DONNÉES AUTOCHTONES : EN MARCHE VERS L'AUTODÉTERMINATION ET DE BONNES DONNÉES

Mikayla Redden et Dani Kwan-Lafond

**Objectifs d'apprentissage**

À la fin de ce chapitre, vous pourrez :

1. Exprimer clairement l'importance de la souveraineté des données autochtones et son rôle dans l'autodétermination des peuples autochtones.
2. Identifier les données qui se concentrent sur les déficits et expliquer en quoi ces données sont préjudiciables.
3. Définir les différences entre les prémisses de la culture de recherche occidentale/dominante et celles de la culture de recherche autochtone et comprendre leur impact sur la prise de décision fondée sur les données lors du processus de la recherche.

## Introduction

« Peuples autochtones » est probablement l'expression la plus galvaudée au sein de la population mondiale contemporaine pour décrire les peuples colonisés ou anciennement colonisés qui, de nos jours, sont unis sur le plan politique en raison d'un passé commun de dépossession et d'exploitation territoriale et culturelle à la suite de la colonisation. Ce chapitre se concentre sur l'histoire et les itérations contemporaines de **vol et d'exploitation du savoir** des communautés autochtones (dans ce contexte particulier, il s'agit de collecte de

connaissances autochtones sans demander la permission des partenaires au sein de la communauté ou sans les consulter). Il y est également question de la souveraineté de ces communautés en ce qui a trait à leurs propres données. L'exploration en profondeur du savoir et la souveraineté des données se croisent, car les données numériques constituent la manière la plus courante de stocker et d'archiver les connaissances qu'utilisent les membres des communautés et les chercheuses et chercheurs.

Pour commencer, nous présenterons un bref historique de la communauté politique mondiale des peuples autochtones. Nous mettrons l'accent sur l'impact de la Déclaration des Nations Unies sur les droits des peuples autochtones (DNUDPA) au Canada. Dans le contexte canadien, « Autochtone » fait référence à trois groupes distincts sur le plan ethnique et culturel : Premières Nations, Métis et Inuit.

## Les Nations Unies et l'autodétermination des peuples autochtones

La DNUDPA a été adoptée en 2007 par tous les États membres des Nations Unies, à l'exception de quatre États-nations coloniaux : le Canada, la Nouvelle-Zélande, l'Australie et les États-Unis (Nations Unies, 2007). Ils ont signé la Déclaration en 2012, après que les peuples autochtones de ces États membres et leurs alliés aient déployé des efforts considérables. La DNUDPA est un prolongement du système des Droits de la personne, lequel est défini le plus clairement dans la Déclaration universelle des droits de la personne de 1960. La DNUDPA ne contient pas de droits de la personne, mais elle définit et affirme les droits dont sont souvent privés les peuples autochtones (Erueti, 2022).

L'accès aux droits des peuples autochtones, ou le fait de les en priver, varient beaucoup à l'échelle internationale. Si les États colonisateurs se définissent en partie par leur domination majoritaire sur les communautés autochtones à l'intérieur de leurs frontières, il y a encore des aspects locaux, historiques et culturels particuliers à ces contextes à prendre en considération.  Les contextes locaux auront une incidence sur la manière de négocier et de prendre des décisions sur l'accès et le contrôle des données, de l'information et du savoir; toutefois, ils sont trop nombreux pour que nous nous y attardions maintenant. Nous soulignons plutôt la nécessité de comprendre la politique et le rôle des cadres stratégiques lorsqu'il est question de favoriser la souveraineté des données, de l'assurer, tout en prévenant le vol et l'exploitation du savoir.

La DNUDPA constitue le cadre de référence des droits de la personne le plus connu et le plus récent qui tente de restituer et de maintenir les droits des peuples autochtones. Il est important de mentionner la Convention 169 de l'Organisation internationale du Travail (OIT) (1989), que la plupart des États-nations d'Amérique centrale et d'Amérique du Sud ont signée et adoptée avant que la DNUDPA soit élaborée. L'OIT est une agence de l'ONU. Elle met l'accent sur les travailleuses et travailleurs et les conditions de travail des États-nations membres. La Convention 169 est une révision de la Convention 107 relative aux populations

aborigènes et tribales (1957), en plus d'un changement de titre. Elle a vu le jour dans le sillage de la Deuxième Guerre mondiale en raison de l'oppression et de la discrimination dont étaient victimes les peuples autochtones.

La Convention 107 et la Convention 169 révisée sont des lois dans les États-nations qui les ont adoptées (Hanson, s.d.-a; s.d.-b). La Convention 169 est composée de 44 articles qui définissent des normes minimales en matière de soins de santé, d'éducation et d'emploi. Elle reconnaît également le droit à l'**autodétermination** et en appelle aux États-nations de protéger les peuples autochtones afin qu'ils ne soient pas déplacés (Hanson, s.d.-a; s.d.-b). Alors que les cadres stratégiques précédents en matière de droits de la personne reposaient sur les droits individuels, la DNUDPA élargit ces droits aux groupes collectifs des peuples autochtones, y compris ceux qui vivent en tant que groupes minoritaires au sein d'États-nations plus grands (comme c'est le cas au Canada). Ce cadre stratégique mondial met l'accent non seulement sur les droits et les identités collectives, mais également sur l'autodétermination et le droit au consentement préalable, donné librement et en connaissance de cause (CLPCC). De plus, il fait référence aux préjudices historiques et actuels et présente également des mesures de réparation (Erueti, 2022). L'adoption de la DNUDPA était ambitieuse en ceci qu'elle dépend de l'adoption, par chaque État membre, d'une législation en vertu de la DNUDPA (contrairement à la Convention 169).

## Une histoire de peuples autochtones et de mauvaises données

Les relations entre les peuples autochtones et le gouvernement dans leurs pays anglo-saxons tourne autour des politiques administratives et des programmes qui en découlent. C'est notamment le cas au Canada où le mandat de Services aux Autochtones Canada (SAC) se lit, en partie, comme suit : « Notre vision est d'appuyer et d'habiliter les Autochtones afin qu'ils puissent offrir de façon indépendante des services et aborder les différentes conditions socio-économiques au sein de leurs communautés. » (Services aux Autochtones Canada, 2022). SAC se concentre sur les désavantages et les disparités sociales des peuples autochtones et sur la manière dont l'État-nation colonial peut les aider. Le discours est semblable dans les mandats du United States Department of the Interior Bureau of Indian Affairs (2023) et du cadre de travail *Closing the Gap* de la National Indigenous Australian Agency (2022) (pour accéder à une analyse détaillée de ces politiques, consultez Walter *et al.,* 2021). Chacune de ces organisations coloniales fonde les décisions relatives à leurs politiques sur des données.

Dans tous ces pays, les données dépeignent les peuples autochtones comme ayant une mauvaise santé, des niveaux d'éducation plus faibles et un statut socioéconomique inférieur, ce qui se traduit souvent par des taux d'incarcération, de victimisation et de suicide incroyablement élevés – des taux justifiés par les chiffres colligés. Toutes ces nations disposent de politiques actives pour assimiler les peuples autochtones au sein de la société

anglo-saxonne en retirant de force les enfants de leurs familles et de leurs communautés. Les Autochtones ne contestent pas ces données, mais ils remettent en question les prémisses de nature sociale, raciale et culturelle faites par les personnes qui recueillent les données (Walter et Andersen, 2013). De telles prémisses ne nous donnent qu'un aperçu étroit, colonisé des réalités autochtones (Walter et Suina, 2019). Par conséquent, les politiques et programmes élaborés sur la base de ces données ne reflètent pas les besoins des peuples autochtones. Toutes les données recueillies auprès de ces derniers devraient être contrôlées, interprétées et gérées par eux, en leur y donnant accès.

Ce chapitre présente l'idée de la **souveraineté des données autochtones**, soit le droit des peuples autochtones de collecter, d'analyser, d'interpréter, de gérer, de distribuer et de réutiliser les données auxquelles ils ont accès et qui sont dérivées de leurs communautés ou en lien avec elles. Il y sera également question des cadres de travail et des stratégies qui affirment la souveraineté des données autochtones dans la culture de recherche dominante.

## Données autochtones : de quoi s'agit-il? En quoi la situation serait-elle différente dans le cadre de l'autodétermination autochtone?

« *Données autochtones* » est une expression générique qui fait référence à l'information et au savoir au sujet des personnes, des groupes, des organisations, des manières de savoir et de vivre, des langues, des cultures, de la terre et des ressources naturelles. Elles existent sous divers formats, notamment le **savoir traditionnel**, soit l'information transmise de génération en génération. Il comprend les langues, les récits, les cérémonies, les danses, les chants, les arts, la chasse, le piégeage, la cueillette, la préparation ainsi que le stockage des aliments et des médicaments, la spiritualité, les croyances et les visions du monde. Les données autochtones comprennent également les données numériques et numérisées recueillies par des chercheuses et chercheurs, des gouvernements et des établissements non gouvernementaux (Walter, 2018; First Nations Information Governance Centre, 2016; Walter *et al*., 2021; Walter et Suina, 2019).

À l'échelle des nations colonisées, les données autochtones recueillies par les chercheuses ou chercheurs gouvernementaux et non gouvernementaux mettent l'accent sur les différences, les écarts, les désavantages, le dysfonctionnement et la privation des peuples autochtones, ou ce que Walter (2016, 2018) appelle les 5 D (*differences, disparities, disadvantages, dysfunction, deprivation*). Ces données manquent de contexte social et culturel, car la collecte est effectuée par des chercheuses ou chercheurs et des responsables issus de points de vue non autochtones qui comparent les données à leurs réalités coloniales. Peu importe les analyses réalisées, les statistiques qui orientent les politiques ne sont pas valides, car elles proviennent de données 5 D; par conséquent, l'accent est entièrement mis sur les déficits (Walter et Suina, 2019).

Les données doivent varier largement parmi les communautés autochtones; toutefois, tous s'entendent pour dire que les données autochtones doivent rendre compte des réalités sociale, politique, culturelle et historique des vies autochtones afin de soutenir les besoins autodéterminés des peuples autochtones (Walter, 2018; Walter et Suina, 2019). Ces données se trouvent au cœur du mouvement de souveraineté des données autochtones et sont affirmées en vertu de la DNUDPA.

Le mouvement de souveraineté des données autochtones préconise l'autonomie gouvernementale, c'est-à-dire que les peuples autochtones contrôleraient tous les aspects du processus de recherche, de la conception de l'idée à l'utilisation des données obtenues. Sans la souveraineté des données autochtones, il est impossible de vérifier qu'elles rendent compte de la riche diversité des visions du monde, des manières de savoir, des priorités, des cultures et des valeurs autochtones (Walter et Suina, 2019).

**Organisations d'autonomie gouvernementale en matière de données autochtones dans les nations anglo-saxonnes:**

- Canada : Le Centre de gouvernance de l'information des Premières Nations (https://fnigc.ca/fr/)
- Australie : Maiam nayri Wingara (https://www.maiamnayriwingara.org/)
- Nouvelle-Zélande : Te Mana Raraunga (https://www.temanararaunga.maori.nz/)
- États-Unis : United States Data Sovereignty Network

# Interagir avec le savoir autochtone

Avant d'aborder les pratiques exemplaires relatives à l'utilisation des données autochtones, vous devriez connaître les prémisses suivantes qui distinguent les pratiques de recherche autochtones des pratiques eurocentriques.

**Tableau 1. Différences entre les pratiques de recherche eurocentriques et autochtones.**

| Prémisses eurocentriques | Prémisses autochtones |
|---|---|
| Les chercheuses et chercheurs demeurent objectifs et impartiaux. | **La recherche N'EST PAS objective et impartiale**. C'est impossible. Les chercheuses et chercheurs sont en relation avec tout ce qui vit, y compris les sujets humains et non humains de leur recherche. Les émotions sont reliées à la connaissance. Lorsque nous pensons, nous utilisons la raison, qui est liée à nos |

| | |
|---|---|
| | émotions, ce qui rend la recherche subjective. |
| La recherche est planifiée et dirigée par les chercheuses et chercheurs. | **La recherche repose sur la communauté**. Les membres de la communauté façonnent toujours la question de la recherche. Quel que soit le sujet, la recherche nous permet de rassembler des connaissances qui concourent à un objectif commun : créer une action sociale. La connaissance combinée à l'action mène au changement social. |
| Le processus de recherche n'affecte pas la chercheuse ou le chercheur. | **La croissance personnelle** de la chercheuse ou du chercheur constitue un résultat important (parce que la recherche est subjective). |
| Aucun élément de l'échantillon, matériel ou humain, n'est plus précieux que les autres (hormis une étude de cas). | **Les membres les plus âgés de la communauté sont probablement ceux qui détiennent les connaissances les plus précieuses**. Omettre de les faire participer au processus de recherche signifie qu'il ne repose pas sur le savoir traditionnel. Une mise en garde : tous les membres âgés de la communauté ne sont pas nécessairement des « aînées et aînés ». Il y a des membres plus jeunes de la communauté qui détiennent un savoir ou un vocabulaire traditionnel. Par conséquent, les expressions « Enseignante traditionnelle et Enseignant traditionnel », « Gardienne et Gardien du savoir » ou « Gardienne et Gardien du vocabulaire » s'avèrent davantage descriptives. |

En plus de ces prémisses, les peuples autochtones prennent en considération les quatre « R » suivants dans tout le cycle de recherche, y compris la publication : relationnalité, respect, réciprocité et responsabilité.

- **La relationnalité** se trouve au cœur de toutes les visions du monde et des systèmes de connaissances autochtones (Wilson, 2008). Les relations alimentent toutes nos expériences, pour citer Littletree *et al*. (2020); elles sont « centrales à la définition d'Autochtone[1] » [traduction]. Dans nos interactions avec le monde, nos relations font en sorte que nous sommes redevables dans chaque situation. Nos relations comprennent celles avec la terre, nos ancêtres et les générations à venir. Nous sommes nos relations et nous sommes composés de relations avec quatre mondes : les parties intellectuelle, spirituelle, émotionnelle et physique de nous-mêmes (Archibald, 2008). Il est essentiel que les chercheuses et chercheurs qui s'intéressent aux modes de savoir autochtones comprennent que toutes les données, les livres, les articles, les récits, l'art et les autres produits sont nés de relations (Meyer, 2008). Fruit du peuple, c'est ce qu'on appelle le savoir commun. Celui-ci peut prendre diverses formes : récits et langues partagés, cérémonies, célébrations et cycles de la vie (Holm *et al*., 2003). Dans les modes de savoir autochtones, ces relations se transforment en actions : poser des questions, observer des danses, écouter des proches, rêver, raconter des histoires, vivre des événements de la vie, fabriquer de l'art et participer à des activités intergénérationnelles comme planter des graines et en prendre soin pour qu'elles germent et poussent afin d'en faire la récolte à l'automne. L'expression de ces moyens de savoir prend une forme

1. "*at the heart of what it means to be Indigenous*"

tangible : des documents, des chansons, des outils, des robes traditionnelles, des contes écrits ou oraux, des livres, de la nourriture, des peintures, des gravures et de la poterie. Souvent, des organisations du savoir les détiennent, comme des bibliothèques, des musées, des écoles, des organisations sacrées et des nations autochtones (Kidwell, 1993). Le caractère relationnel au centre de ces éléments échappe souvent aux chercheuses et chercheurs ainsi qu'aux organisations du savoir non autochtones. Pour comprendre davantage la relationnalité, consultez le modèle conceptuel de Littletree (2018) ainsi que les travaux de Holm *et al.* (2003); Archibald (2008); Wilson (2008); Meyer (2008); Kidwell (1993), qui l'ont orienté.

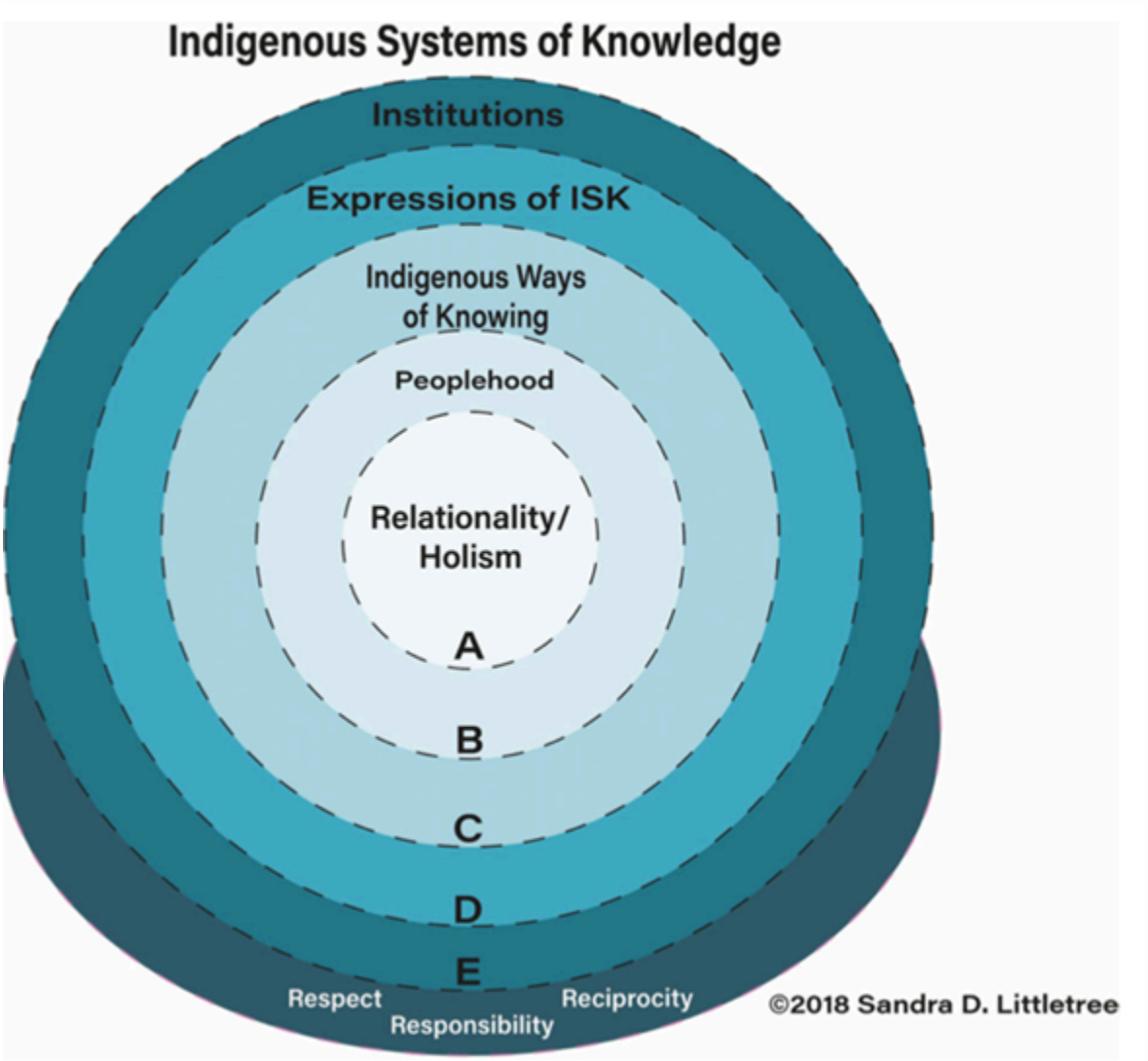



*Figure 1.* Indigenous systems of knowledge conceptual model.

Figure 1. Systèmes de savoirs autochtones selon Sandra D. Littletree. Tous droits réservés. Utilisé avec permission. Traduction: Systèmes de savoirs autochtones (SSA), Établissements, Expressions des SSA, Modes de savoirs autochtones, Peuple, Relationnalité/Holisme, Respect, Responsabilité, Réciprocité.

Le respect, la réciprocité et la responsabilité soutiennent la relationnalité.

- **Respect** de la terre, des protocoles culturels, de l'histoire, de la langue et de la santé intellectuelle, spirituelle, émotionnelle et physique. Évitez les suppositions à propos du savoir avec lequel vous travaillez. Adoptez une approche éclairée; gardez l'esprit ouvert. Le savoir que vous interrogez peut être associé à des événements historiques douloureux et susciter un traumatisme important.
- **Réciprocité** à l'égard de l'information que vous recevez. Faites preuve d'ouverture à donner et à recevoir de l'information. Il y a une longue histoire d'exploitation du savoir par les colonisateurs à l'égard des communautés autochtones. La réciprocité ne fait pas seulement référence à la compensation pécuniaire, bien qu'il soit important de rémunérer les personnes et les communautés pour leur temps et les renseignements communiqués. Elle se rapporte également à l'appui des communautés qui récupèrent les traditions et les expressions culturelles importantes.
- **Responsabilité** d'obtenir un consentement éclairé et de nourrir toute relation créée pour la vie, bien au-delà de la durée du projet de recherche. Les visions du monde autochtones nous disent que le temps n'est pas linéaire : il est circulaire. La communauté avec laquelle vous collaborez doit guider le processus et prendre des décisions au sujet de son savoir et de ses renseignements, et ce, à toutes les étapes du projet.

Pour accéder à un examen approfondi de ces prémisses et considérations, consultez *Research Is Ceremony: Indigenous Research Methods* (2008) de Shawn Wilson et *Centering Relationality: A Conceptual Model to Advance Indigenous Knowledge Organization Practices* (2020) de Sandra Littletree, Miranda Belarde-Lewis et Marisa Elena Duarte.

## Autogouvernance des données des Premières Nations au Canada

En 1994, le gouvernement fédéral a exclu les peuples des Premières Nations qui vivent au sein des communautés autochtones des enquêtes nationales sur la population (Centre de gouvernance de l'information des Premières Nations, 2022a; 2022b). Préoccupés par le manque de données qui en a résulté, les personnes qui défendent les Premières Nations et les universitaires se sont réunis et c'est ainsi que le Centre de gouvernance de l'information des Premières Nations a vu le jour. Deux ans plus tard, l'Assemblée des Premières Nations a mis sur pied le comité directeur national (CDN) dans le but d'élaborer la première enquête longitudinale régionale sur la santé des Premières Nations et des Inuits, une première itération de l'enquête régionale sur la santé des Premières Nations (ERS). Le premier rapport a été publié en 1997 et constitue la seule enquête nationale sur la santé gérée par les Premières Nations et fondée sur les conceptions occidentales et traditionnelles de la santé et du bien-être (First Nations Centre, 1997). Plus tard, un groupe de l'Université de Harvard a passé en revue le rapport en soulignant que

> [le sondage] était unique puisque les Premières Nations s'en approprient le processus de recherche, qu'il intègre explicitement les valeurs des Premières Nations dans la conception de la recherche et qu'il comprend un engagement collaboratif intensif des peuples des Premières Nations et de leurs représentants à chaque étape du processus de recherche[2] [traduction] (Harvard Project on American Indian Economic Development, 2006).

En 1998, le CDN a défini un ensemble de principes intitulé Principes de PCAP® pour faire en sorte que les peuples des Premières Nations possèdent leurs propres données, tout comme ils possèdent leurs propres terres. Plus tard, le CDN est devenu le Comité de gouvernance de l'information des Premières Nations, puis le Centre de gouvernance de l'information des Premières Nations (CGIPN), un organisme autonome à but non lucratif.

**PCAP®** est un acronyme qui signifie propriété, contrôle, accès et possession. Ces quatre principes gouvernent la manière dont les données et l'information relatives aux Premières Nations devraient être collectées, protégées, utilisées et partagées. Les PCAP® ont été créés pour combler une lacune dans les lois occidentales qui ne reconnaissent pas les droits des communautés et des peuples autochtones à contrôler leur information. Les principes traduisent les visions du monde autochtone en matière de gouvernance et de droits collectifs. D'un point de vue historique, les peuples autochtones n'ont pas été consultés quant aux renseignements recueillis à leur sujet, aux personnes qui les ont recueillis, à la manière de les stocker ou qui y aura accès. Ce manque de gouvernance a mené à une cueillette de données qui manque de pertinence à l'égard des priorités et des préoccupations des peuples autochtones.

Ces principes affirment les droits et l'autodétermination des communautés autochtones à détenir, à contrôler, à accéder et à posséder l'information relative à leurs peuples, et exigent que les chercheuses et chercheurs qui souhaitent mener des recherches avec une communauté autochtone prennent connaissance de ces principes avant d'entamer leurs travaux. Les principes peuvent aider toute personne qui travaille (ou qui veut travailler) avec une recherche, des données, de l'information ou des connaissances culturelles autochtones et ils soutiennent le chemin vers la souveraineté des données des peuples autochtones (Centre de gouvernance de l'information des Premières Nations, 2022b). Le CGIPN et le Collège Algonquin ont élaboré un cours en ligne (https://fnigc.ca/fr/les-principes-de-pcap-des-premieres-nations/suivre-le-cours/) afin de former les chercheuses et chercheurs au sujet des principes de PCAP® et de leur histoire.

2. "*[the Survey] was unique in First Nations ownership of the research process, its explicit incorporation of First Nations values into the research design and in the intensive collaborative engagement of First Nations people and their representatives at each stage of the research process*"

Regardez la bande-annonce du cours ici : https://youtu.be/qwwJCo5_eKI (https://youtu.be/qwwJCo5_eKI)

## Les principes de PCAP® des Premières Nations

1. **Propriété** : les communautés ou groupes détiennent leurs propres connaissances, données et renseignements au même titre que tout individu à l'égard de ses renseignements personnels.
2. **Contrôle** : les collectivités ont le contrôle de toutes les étapes du projet de recherche – de la collecte de données au stockage, et tout ce qui se situe entre les deux. Les collectivités détiennent le contrôle et le pouvoir décisionnel de tous les aspects de la recherche et de l'information qui les touchent.
3. **Accès** : les collectivités devraient pouvoir accéder à leur information et à leurs données collectives, peu importe où elles se trouvent. Les collectivités devraient pouvoir gérer l'accès et le contrôle de leur information et prendre des décisions à cet égard.
4. **Possession** : principe plus concret que celui de la propriété. Il s'agit du contrôle physique des données, du mécanisme qui permet de faire valoir et de protéger la propriété.

## Une liste non exhaustive de stratégies pour mener des recherches en respectant les principes de PCAP®

(Adaptée de l'œuvre de Schnarch (2005), National Aboriginal Health Organization (2005), et du First Nations Information Governance Centre (2016))

- Prévoyez qu'il vous faudra plus de temps. Vous devrez obtenir des permissions auprès des entités décisionnelles communautaires comme la ou le Chef ainsi que le Conseil, les comités consultatifs et les Gardiennes et Gardiens du savoir, en plus de votre propre conseil d'éthique, des personnes participantes, des organismes de financement, entre autres. Le consentement communautaire est tout aussi important que le consentement éclairé des personnes participantes. Il faut cesser la recherche si la communauté n'y consent pas.
- Négociez la relation de recherche et rédigez un accord qui affirme vos droits et responsabilités ainsi que ceux de la communauté et de tous les autres partenaires du processus de recherche. Faites en sorte que

toutes les parties comprennent le document, l'acceptent et en reçoivent une copie.

- Cherchez des sources de financement dont les politiques affirment l'autodétermination et la souveraineté autochtones.
- Offrez des explications et donnez de la rétroaction sur tous les aspects du projet, notamment votre but, les avantages prévus et les risques du projet; les méthodes que vous prévoyez d'utiliser; la manière dont vous souhaitez recruter des personnes participantes; la manière de rapporter vos résultats et ce que vous comptez faire des données obtenues.
- Respectez les protocoles de confidentialité en ce qui a trait à la culture et à la communauté, le bien-être, les droits individuels et collectifs des peuples autochtones. Suivez des lignes directrices rigoureuses en matière d'éthique. Veillez à prendre en compte le fait que chaque communauté peut avoir des interprétations et des niveaux de confort différents en ce qui concerne les principes de PCAP® et d'autres cadres d'autodétermination.
- Soutenez les intérêts de la communauté et maximisez les avantages des travaux. Il s'agit notamment de vous appuyer sur les initiatives autochtones qui ont connu du succès et d'offrir des possibilités de renforcement des capacités.
- Présentez toutes vos communications, tous vos résumés et rapports de recherche à la communauté dans la langue appropriée avant de les publier.
- Veillez à ce que les collectivités autochtones aient accès à leurs données, et non seulement aux rapports ou aux publications qui en résultent.

## Analyse critique des principes de PCAP®

Des critiques pourraient affirmer que le développement des capacités est un préalable nécessaire aux données et aux recherches contrôlées par les peuples autochtones. Ils pourraient argumenter que la communauté ne dispose pas de l'expertise requise, ce qui pourrait entraîner des risques et des conséquences. Certains pourraient encourager les membres des Premières Nations, les Inuits et les Métis qui détiennent des diplômes d'enseignement supérieur à participer aux recherches et celles et ceux qui n'en ont pas à en obtenir en lien avec la recherche.

Ces deux solutions favorisent les individus, mais qu'en est-il du renforcement d'une nation et d'une communauté en plus du développement professionnel? Pour obtenir des diplômes d'études supérieures, il faut souvent quitter sa communauté, ce qui peut entraîner une aliénation. Encore récemment, les peuples des Premières Nations, les Inuits et les Métis devaient renoncer à leur statut d'Autochtone et s'assimiler à la société coloniale blanche. Les vrais bénéficiaires de cette situation sont les établissements où ces personnes étudient ou pour lesquels elles travaillent.

Les occasions de travailler en tant que chercheuse ou chercheur à temps plein au sein des communautés sont

très rares. La capacité de fonctionner dans deux mondes (c'est-à-dire, trouver un équilibre entre les valeurs autochtones et gérer les responsabilités communautaires tout en faisant progresser sa carrière universitaire) constitue un défi – un défi qui force parfois les gens à prendre des décisions déchirantes. En outre, suggérer que les communautés ne peuvent pas mener elles-mêmes des recherches éthiques et bénéfiques est, au mieux, préjudiciable. Pour être utiles, il n'est pas nécessaire que les recherches soient spécialisées, qu'elles emploient des méthodologies complexes ou qu'elles regorgent de jargon scientifique.

## Se tourner vers l'avenir grâce aux principes de PCAP®

À ce jour, les principes de PCAP® constituent l'outil le plus puissant des peuples des Premières Nations au Canada et de leurs alliés en ce qui a trait à l'affirmation de la souveraineté sur leurs données. Les principes peuvent remettre en question les mauvaises données ainsi que les pratiques de recherche erronées et en favoriser de meilleures. Néanmoins, il faut encore relever des défis pour aller de l'avant de manière significative.

- Pour les comités d'éthique en recherche : évaluez toutes les demandes de recherche futures à la lumière des principes de PCAP®, ou d'un autre cadre adéquat, afin que la recherche s'y conforme. Qu'en est-il de l'évaluation des recherches en cours ou historiques? N'oublions pas que les communautés autochtones ont subi un préjudice important du fait de pratiques de recherche abusives. La vérité et la réconciliation ne demeurent-elles pas un engagement déclaré de plusieurs établissements d'enseignement et organismes gouvernementaux?
- Pour les groupes qui rédigent des politiques : abordez les principes de propriété, de contrôle, d'accès et de possession des données et de la recherche pour toutes les politiques et révisez les politiques antérieures. Les politiques institutionnelles en matière d'incendie et de tabagisme, celles relatives au stockage et à la diffusion des données et celles portant sur la propriété intellectuelle constituent autant d'exemples de politiques existantes ayant une incidence sur la recherche communautaire.
- Pour les chercheuses et chercheurs : faites preuve de souplesse, acceptez le compromis, remettez en question vos propres idées préconçues en ce qui a trait à la propriété, au contrôle, à l'accès et à la possession des travaux que vous pourriez envisager comme « vôtres ». Souvenez-vous qu'une véritable recherche communautaire a pour but de créer une action sociale positive et déterminée par la communauté autochtone.
- Pour les professionnelles et professionnels en gestion des données : pensez à la communauté qui fait l'objet du projet de recherche avant d'élaborer un **plan de gestion des données**. Remettez-vous-en à la communauté pour évaluer sa compréhension et ses niveaux de confort avec l'autodétermination des données, et ce, peu importe l'ensemble de principes à partir duquel vous travaillez. Abordez toujours les projets en ayant à l'esprit la relationnalité, le respect, la réciprocité et la responsabilité.

**Autres cadres de travail pour l'autodétermination autochtone et de bonnes pratiques de recherche**

- Canada : *National Inuit Strategy On Research (https://www.itk.ca/national-strategy-on-research-launched/)* (2018), Inuit Tapiriit Kanatami
    - Remarque : à la connaissance des autrices, il n'existe pas de cadre pour les nations Métis.
- Australie : *Communique (https://static1.squarespace.com/static/5b3043afb40b9d20411f3512/t/5b6c0f9a0e2e725e9cabf4a6/1533808545167/Communique%2B-%2BIndigenous%2BData%2BSovereignty%2BSummit.pdf)* (2018), Maiam nayri Wingara
- Nouvelle-Zélande : *Principles of Maori Data Sovereignty (https://static1.squarespace.com/static/58e9b10f9de4bb8d1fb5ebbc/t/5bda208b4ae237cd89ee16e9/1541021836126/TMR+Ma%CC%84ori+Data+Sovereignty+Principles+Oct+2018.pdf)* (2018), Te Mana Raraunga
- Mondial : *CARE Principles for Indigenous Data Governance (https://www.gida-global.org/care#)* (2019), Global Indigenous Data Alliance

# Conclusion

Ce chapitre met l'accent sur l'histoire et les itérations contemporaines de la dépossession et de l'appropriation des connaissances et de savoir-faire des autochtones, y compris un historique de la communauté politique mondiale des peuples autochtones, particulièrement l'impact de la Déclaration des Nations Unies sur les droits des peuples autochtones (DNUDPA) au Canada. Nous avons mis en contexte l'importance de la souveraineté des données autochtones d'un point de vue historique. Nous avons partagé des pratiques exemplaires sur la manière de travailler avec des données autochtones afin de remettre en question des pratiques historiquement préjudiciables. Parmi ces pratiques exemplaires, il a été question des principes de PCAP®, l'outil le plus puissant dont disposent les peuples des Premières Nations au Canada et leurs alliés dans l'affirmation de la souveraineté sur leurs données.

## Questions de réflexion

1. Quelle prémisse ou quelle considération en lien avec la redéfinition de vos pratiques de recherche afin d'intégrer les modes de connaissance autochtones vous a semblé la plus difficile? Pourquoi?
2. Pouvez-vous déterminer une stratégie que vous pourriez mettre en place dans votre propre recherche afin de respecter les principes de PCAP® ou d'autres principes d'autodétermination?
3. Pensez à demander à votre établissement de fournir l'occasion aux chercheuses et chercheurs de suivre le cours de formation en ligne Les Fondamentaux des principes de PCAP® ou d'organiser un atelier offert par le CGIPN avant de présenter votre prochaine demande à votre comité d'éthique en recherche. Si le financement nécessaire n'est pas disponible, consultez les ressources suivantes:

    - Regardez cette courte vidéo : *Comprendre les principes de PACP® des Premières Nations (https://youtu.be/qwwJCo5_eKI)* du Centre de gouvernance de l'information des Premières Nations (2014)
    - Regardez cette présentation : *La souveraineté en matière de données des Premières Nations et les vingt-cinq ans de PCAP® avec Aaron Franks (https://youtu.be/46wqFGvbRxU)* (en anglais avec sous-titre en français), faite lors du Canada Open Data Summit en 2022
    - Découvrez cette page Web : *Les principes de PCAP® des Premières Nations (https://fnigc.ca/fr/les-principes-de-pcap-des-premieres-nations/)*
    - Imprimez ce document à mettre à disposition sur votre lieu de travail : *Les principes de PCAP® des Premières Nations (https://fnigc.ca/wp-content/uploads/2022/10/OCAP_Brochure_20220927_web.pdf)* du Centre de gouvernance de l'information des Premières Nations (2022)
    - Lisez ce document : *Exploration of the impact of Canada's information management regime on First Nations data sovereignty (https://fnigc.ca/wp-content/uploads/2022/09/FNIGC_Discussion_Paper_IM_Regime_Data_Sovereignty_EN.pdf)* du Centre de gouvernance de l'information des Premières Nations (2022)
    - Lisez ce document : *Ownership, Control, Access, and Possession (OCAP®): The path to First Nations information governance (https://fnigc.ca/wp-content/uploads/2020/09/5776c4ee9387f966e6771aa93a04f389_ocap_path_to_fn_information_governance_en_fi*

*nal.pdf)*, CGIPN (2014)
- Imprimez cette infographie : *Indigenous Peoples' rights in data (https://www.gida-global.org/data-rights)* de la Global Indigenous Data Alliance (GIDA) (2022)
- Parcourez cette présentation : *Indigenous data sovereignty and governance (https://static1.squarespace.com/static/5d3799de845604000199cd24/t/637acfbec86a122d68b0f317/1668992965093/Final_Attribution_NonCommercial_NoDerivatives_4_International.pdf)* de la GIDA (2022)

**Éléments clés à retenir**

- Les données collectées par les États-nations coloniaux mènent à « l'altérisation » des peuples autochtones en les comparant à une réalité de colonisation anglo-saxonne dépourvue de contexte social et culturel et axée sur les désavantages et disparités sociales. Cette approche invalide toute politique qui en découle. La souveraineté des données autochtones est essentielle si l'objectif est d'établir des politiques et des programmes valides et utiles.
- Les chercheuses et chercheurs informés des systèmes de connaissances autochtones émettent des hypothèses différentes que celles et ceux qui s'en tiennent aux modes occidentaux. Les chercheuses et chercheurs autochtones font également en sorte que leurs travaux reposent sur la communauté en mettant l'accent sur la relationnalité, le respect, la réciprocité et la responsabilité.
- Des données autochtones adéquates sont autodéterminées; c'est-à-dire que les peuples autochtones en sont propriétaires, les contrôlent, déterminent qui peut y accéder et en supervisent le stockage.

## Lectures et ressources supplémentaires

Centre de gouvernance de l'information des Premières Nations. (2014, 12 septembre). *Comprendre les principes de PACP® des Premières Nations* [Vidéo]. Youtube. https://youtu.be/qwwJCo5_eKI (https://youtu.be/qwwJCo5_eKI)

Lovett, R. Lee, V., Kukutai, T., Cormack, D., Rainie, S. C. et Walker, J. (2019). Good data practices for Indigenous data sovereignty and governance. Dans A. Daly, S. K. Devitt, et M. Mann (dir.), *Good data* (pp. 26-36). Institute of Network Cultures.

Toombs, E., Drawson, A. S., Chambers, L., Bobinski, T. L. R., Dixon, J. et Mushquash, C. J. (2019). Moving towards an Indigenous research process: A reflexive approach to empirical work with First Nations communities in Canada. *The International Indigenous Policy Journal, 10*(1). https://doi.org/10.18584/iipj.2019.10.1.6 (https://doi.org/10.18584/iipj.2019.10.1.6)

Tuhiwai Smith, L. (2012). *Decolonizing methodologies: Research and Indigenous peoples*. University of Otago Press.

# Bibliographie

Archibald, J.-A. (2008). *Indigenous storywork: Educating the heart, mind, body, and spirit*. UBC Press.

Centre de gouvernance de l'information des Premières Nations. (2022a). *Notre histoire*. https://fnigc.ca/fr/a-propos-de-nous/notre-histoire/#slide-1 (https://fnigc.ca/fr/a-propos-de-nous/notre-histoire/#slide-1)

Centre de gouvernance de l'information des Premières Nations. (2022b). *Les principes de PCAP® des Premières Nations*. https://fnigc.ca/fr/les-principes-de-pcap-des-premieres-nations/ (https://fnigc.ca/fr/les-principes-de-pcap-des-premieres-nations/)

Erueti, A. (2022). *The UN declaration on the rights of Indigenous Peoples: A new interpretative approach*. Oxford University Press.

First Nations Centre. (1997). *First Nations and Inuit Regional Health Surveys, 1997*. https://fnigc.ca/wp-content/uploads/2020/09/71d4e0eb1219747e7762df4f6a133a3d_rhs_1997_synthesis_report.pdf (https://fnigc.ca/wp-content/uploads/2020/09/71d4e0eb1219747e7762df4f6a133a3d_rhs_1997_synthesis_report.pdf)

First Nations Information Governance Centre. (2016). Pathways to First Nations' data and information sovereignty. Dans T. Kukutai, et J. Taylor (dir.), *Indigenous data sovereignty: Toward an agenda*, (pp. 139-156). ANU Press.

Hanson, E. (s.d.-a.). *ILO convention 107*. UBC Indigenous Foundations. https://indigenousfoundations.arts.ubc.ca/ilo_convention_107/ (https://indigenousfoundations.arts.ubc.ca/ilo_convention_107/)

Hanson, E. (s.d.-b.). *ILO convention 169*. UBC Indigenous Foundations. https://indigenousfoundations.arts.ubc.ca/ilo_convention_169/ (https://indigenousfoundations.arts.ubc.ca/ilo_convention_169/)

Harvard Project on American Indian Economic Development. (2006). *Review of the First Nations regional longitudinal health survey (RHS) 2002/2003*. https://fnigc.ca/wp-content/uploads/2020/09/67736a68b4f311bfbf07b0a4906c069a_rhs_harvard_independent_review.pdf (https://fnigc.ca/wp-content/uploads/2020/09/67736a68b4f311bfbf07b0a4906c069a_rhs_harvard_independent_review.pdf)

Holm, T., Pearson, J. D. et Chavis, B. (2003). Peoplehood: A model for the extension of sovereignty in American Indian studies. *Wicazo Sa Review, 18*(1), 7–24. https://doi.org/10.1353/wic.2003.0004 (https://doi.org/10.1353/wic.2003.0004)

Kidwell, C. S. (1993). Systems of Knowledge. Dans A. M. Josephy et F. E. Hoxie (dir.), *America in 1492: The world of the Indian peoples before the arrival of Columbus*, (pp. 369–403). Vintage Books.

Littletree, S., Belarde-Lewis, M. et Duarte, M. (2020). Centering relationality: A conceptual model to advance Indigenous knowledge organization practices. *Knowledge Organization, 47*(5), 410-426. https://doi.org/10.5771/0943-7444-2020-5-410 (https://doi.org/10.5771/0943-7444-2020-5-410)

Meyer, M. A. (2008). Indigenous and authentic: Hawaiian epistemology and the triangulation of meaning. Dans N. K. Denzin, Y. S. Lincoln, et L. T. Smith (dir.), *Handbook of critical and indigenous methodologies*, (pp. 217–232). Sage.

Nations Unies. (2007). *Déclaration des Nations Unies sur les droits des peuples autochtones*. https://social.desa.un.org/sites/default/files/migrated/19/2018/11/UNDRIP_F_web.pdf (https://social.desa.un.org/sites/default/files/migrated/19/2018/11/UNDRIP_F_web.pdf)

National Aboriginal Health Organization. (2005). *Ownership, control, access, and possession (OCAP) or self-determination applied to research: A critical analysis of contemporary First Nations research and some options for First Nations communities*. https://ruor.uottawa.ca/bitstream/10393/30539/1/OCAP_Critical_Analysis_2005.pdf (https://ruor.uottawa.ca/bitstream/10393/30539/1/OCAP_Critical_Analysis_2005.pdf)

National Indigenous Australians Agency (2022). *Closing the gap*. Australian Government. https://www.niaa.gov.au/Indigenous-affairs/closing-gap (https://www.niaa.gov.au/Indigenous-affairs/closing-gap)

Schnarch, B. (2005). Ownership, Control, Access, and Possession (OCAP) or Self-Determination Applied to Research: A critical analysis of contemporary First Nations research and some options for First Nations

communities. *Journal of Aboriginal Health, 1*(1), 80-95. https://jps.library.utoronto.ca/index.php/ijih/article/view/28934 (https://jps.library.utoronto.ca/index.php/ijih/article/view/28934)

Services aux Autochtones Canada. (2022). *Mandat*. Gouvernement du Canada. https://www.sac-isc.gc.ca/fra/1539284416739/1539284508506 (https://www.sac-isc.gc.ca/fra/1539284416739/1539284508506)

United States Department of the Interior Bureau of Indian Affairs. (2023). *Mission statement*. https://www.bia.gov/bia (https://www.bia.gov/bia)

Walter, M. (2016). Data politics and Indigenous representation in Australian statistics. Dans T. Kukutai et J. Taylor (dir.), *Indigenous data sovereignty: Toward an agenda*, (pp. 79-87). ANU Press.

Walter, M. (2018). The voice of indigenous data: Beyond the markers of disadvantage. *Griffith Review*, (60), 256–263.

Walter, M. et Andersen, C. (2013). *Indigenous statistics: A quantitative research methodology*. Routledge.

Walter, M., Kukutai, T., Russo Carroll, S. et Rodriguez-Lonebear, D. (2021). *Indigenous data sovereignty and policy*. Taylor & Francis.

Walter, M. et Suina, M. (2019). Indigenous data, Indigenous methodologies, and Indigenous data sovereignty. *International Journal of Social Research Methodology, 22*(3), 233-243. https://doi.org/10.1080/13645579.2018.1531228 (https://doi.org/10.1080/13645579.2018.1531228)

Wilson, S. (2008). *Research is ceremony: Indigenous research methods*. Fernwood Publishing.

## À propos des auteurs

### Mikayla Redden

Je suis une femme multiraciale, Anishinaabe et d'origine anglocoloniale. Je suis petite-fille, fille, sœur, tante, aidante et apprenante. Je vis et travaille dans la région de Tkaronto Purchase, mais je suis née et j'ai grandi dans la région du Traité 20. Bien que je sois membre de la Première Nation de Curve Lake, je n'ai pas été élevée dans la communauté. Mon arrière-grand-père est John « Jack » Jacobs. Jack était marié à mon arrière-grand-mère, Edith Marsden, de la Première nation de Scugog. Jack s'est émancipé et a émancipé ses enfants en vertu de l'article 214 de la Loi sur les Indiens en mars 1935. Cela signifie qu'ils ont renoncé à leur identité indienne et se sont assimilés à la société des colons blancs. Notre famille s'est installée dans la ville voisine de Burleigh Falls, en Ontario, où elle a trouvé une communauté avec une population Métis locale. La branche de la famille dont je suis issue s'est ensuite installée à Keene, en Ontario. J'ai le privilège de marcher dans deux mondes :

j'apprends de mes relations dans les réserves et hors des réserves, à la fois urbaines et rurales, traditionnelles et contemporaines, et je suis en mesure d'appliquer des éléments de cette connaissance à ma vie professionnelle en tant que bibliothécaire universitaire.

## Dani Kwan-Lafond

Je suis une femme multiraciale, née sur le territoire du Traité 4, et je suis membre de nombreuses communautés par le biais de ma famille et de mes proches, notamment les communautés asiatique, française, juive et anishnaabe. Je donne des cours axés sur l'inégalité sociale, la race et les relations entre autochtones et colons. Je ne m'identifie pas comme autochtone et mon travail se concentre sur les politiques coloniales et les idéologies qui maintiennent l'inégalité, ainsi que sur l'apprentissage basé sur la terre, l'autochtonisation et l'apprentissage expérientiel. Je vis, je travaille et je forme une communauté sur les terres historiques et actuelles de la nation Anishnabek, qui abrite également la Confédération Haudenosaunee et d'autres peuples autochtones, ainsi que de nombreuses populations nouvellement arrivées au Canada.

PARTIE II

# CONTEXTE CANADIEN POUR LA GESTION DES DONNÉES DE RECHERCHE

4.

# HISTORIQUE ET PAYSAGE CANADIEN DE LA GESTION DES DONNÉES DE RECHERCHE

Eugene Barsky; Elizabeth Hill; Tatiana Zaraiskaya; Minglu Wang; et Lucia Costanzo

**Objectifs d'apprentissage**

À la fin de ce chapitre, vous pourrez :

1. Décrire l'histoire et le contexte de la gestion des données de recherche au Canada.
2. Identifier les différents groupes qui participent à la gestion des données de recherche au Canada.
3. Comprendre les progrès régionaux en matière de gestion des données de recherche.
4. Comprendre les outils technologiques et les dépôts de données utilisés en collaboration par les chercheuses et chercheurs canadiens.

## Introduction

Le Canada et plusieurs autres pays développés mettent en place des exigences en matière de **gestion des données de recherche** (GDR) pour un éventail de disciplines universitaires. Les obstacles à la gestion, à la préservation et au partage de données, dont il sera question dans d'autres chapitres, sont surmontés à l'aide de recommandations et de recours aux normes établies par les différentes communautés, comme les **métadonnées** reconnues, la documentation des données et les dépôts disciplinaires.

Comme vous l'avez appris, les Instituts de recherche en santé du Canada (IRSC), le Conseil de recherche en sciences naturelles et en génie du Canada (CRSNG) et le Conseil de recherche en sciences humaines (CRSH)

sont les **trois organismes** de financement de recherche fédéraux. En mars 2021, ils ont publié la Politique des trois organismes sur la gestion des données de recherche (https://science.gc.ca/site/science/fr/financement-interorganismes-recherche/politiques-lignes-directrices/gestion-donnees-recherche/politique-trois-organismes-gestion-donnees-recherche) afin de commencer à introduire graduellement des exigences relatives aux **plans de gestion des données** (PGD) pour certains programmes de subventions. Par le biais de ces programmes, les organismes subventionnaires encouragent fortement les établissements de recherche à fournir à leurs chercheuses et chercheurs un environnement qui permet des pratiques robustes de gestion et de curation des données et à soutenir la gestion et le versement des **données de recherche** dans des dépôts sécuritaires, accessibles, et qui assurent la curation de leur contenu. Même avant la publication de cette politique, des têtes dirigeantes et des organismes visionnaires, particulièrement des bibliothèques canadiennes, ont promu des initiatives et déployé des efforts pour sensibiliser localement les gens à la gestion de données.

Au cours de la dernière décennie, les bibliothèques universitaires canadiennes ont travaillé à la prestation de soutien en GDR auprès de leurs communautés (Steeleworthy, 2014; Liss, 2018). Les collaborations entre les bibliothèques universitaires et la communauté de recherche abordent les principaux défis en ce qui a trait à l'infrastructure, aux services et à la formation par le biais d'initiatives comme le réseau Portage (Portage) et Données de recherche Canada (DRC). Ces deux entités font désormais partie de l'Alliance de recherche numérique du Canada (https://alliancecan.ca/fr) (Alliance).

Dans ce chapitre, nous donnons un aperçu de la GDR canadienne, laquelle a commencé par des initiatives sur le terrain avant de passer à des efforts nationaux à plus grande échelle. Ce chapitre actualise et développe des travaux réalisés il y a quelques années de cela (Barsky *et al*., 2017).

## Bref historique de la gestion des données de recherche au Canada

Depuis la fin du 20$^{e}$ siècle, les bibliothèques universitaires discutent et plaident en faveur de la centralisation des services d'archivage et de découverte de données ainsi que pour un meilleur accès aux données de recherche au Canada. Toutefois, dans un pays avec une population relativement peu nombreuse et dispersée géographiquement, la centralisation constitue un défi de taille. Lorsque les initiatives en GDR ont commencé, les bibliothèques universitaires canadiennes ont réussi à consolider les collections de données en **sciences sociales**, surtout les collections gouvernementales mises à la disposition des chercheuses et chercheurs aux fins d'**analyse secondaire**. Les bibliothèques universitaires ont également participé à la création d'une communauté de pratique nationale en matière de GDR. En tirant profit des liens étroits entre les chercheuses, les chercheurs, les bibliothécaires et les spécialistes des données, le réseau des **responsables de l'intendance de données** a pu non seulement collaborer à l'élaboration d'outils et d'infrastructures de GDR, mais il a

également pu offrir de nouvelles ressources à leurs communautés de recherche grâce à de la formation sur les données, une consultation et des services de dépôt de données.

## Donner accès aux données de Statistiques Canada

L'Initiative de démocratisation des données (https://www.statcan.gc.ca/fr/microdonnees/idd) (IDD), un service par abonnement qui donne accès aux données de Statistiques Canada, illustre parfaitement comment la collaboration en gestion des données peut aider à bâtir et à entretenir une infrastructure de diffusion des données et former des expertes et experts des données. Le programme de l'IDD a vu le jour en 1996 à la suite de consultations entre Statistiques Canada, l'Association des bibliothèques de recherche du Canada (ABRC) et la Fédération des sciences humaines (Boyko et Watkins, 2011). L'IDD est une réaction à la fois aux coûts élevés pour les fichiers de microdonnées à usage public de Statistiques Canada et à l'absence d'infrastructure dans les universités canadiennes pour donner accès à ces données (Humphrey, 2005). En raison des coupes budgétaires dans les années 1980, les prix établis pour l'accès aux fichiers de microdonnées à usage public visaient un recouvrement total des coûts; par conséquent, même les recherches les mieux financées ne pouvaient pas se les procurer.

La collection de l'IDD comprend des milliers de fichiers de données pour des centaines d'ensembles d'enquêtes. Sa taille et la demande des chercheuses et chercheurs ont influencé dans les bibliothèques la croissance de l'infrastructure de données nécessaire pour gérer et préserver l'accès à ces données. Au départ, peu de bibliothèques disposaient d'une expertise pour prendre en charge les services de données. Toutefois, Statistiques Canada exigeait un point de contact au sein des bibliothèques qui diffusaient les données. Les bibliothèques ont donc dû développer l'expertise de leur personnel par le biais d'activités de formation de l'IDD (Humphrey, 2005). Ces programmes de formation, en plus de permettre de développer la compétence du personnel dans les bibliothèques, ont créé une communauté universitaire nationale de données. La nécessité de prendre en charge le programme de l'IDD a également mené à la mise sur pied d'initiatives visant à offrir ou améliorer la transmission des données aux spécialistes en données et aux personnes utilisatrices. Ces systèmes de diffusion des données comprennent Odesi et Abacus en Ontario et en Colombie-Britannique ainsi que d'autres systèmes au Québec et dans les provinces de l'ouest (Hill et Gray, 2016). Vous pouvez approfondir vos connaissances en consultant les sources citées pour ce chapitre.

## Stratégie de données de recherche nationale au début des années 2000

Dans les années 2000, Hackett (2001) a défini une vaste gamme d'enjeux en lien avec l'acquisition, la préservation des données de recherche canadiennes et leur accès. Localiser les données canadiennes

préalablement recueillies et y accéder étaient les principaux problèmes soulevés. Cette difficulté était due aux coûts élevés, à l'absence d'un répertoire de ressources ou d'un service de dépôt central et à l'absence d'un organisme national pour définir les normes et offrir une direction, du financement ainsi qu'une infrastructure (Hackett, 2001). Il y avait quelques exceptions. Les disciplines des sciences physiques et de la génétique disposaient déjà d'une culture internationale de partage des données dans des dépôts disciplinaires. L'importance du partage des données de recherche pour la pratique scientifique dans ces disciplines a mené à la création de dépôts canadiens pour lesquels aucune politique n'était nécessaire. C'est notamment le cas du Polar Data Catalogue (https://www.polardata.ca/) (un projet du Canadian Cryospheric Information Network), du Centre canadien de données astronomiques (https://www.cadc-ccda.hia-iha.nrc-cnrc.gc.ca/fr/) (une initiative du Canadian Advanced Network for Astronomical Research), et de CBRAIN (https://mcin.ca/technology/cbrain/) (une initiative du McGill Centre for Integrative Neuroscience, MCIN). Toutefois, le manque de normes interdisciplinaires coordonnées en matière de curation des données et de métadonnées demeure problématique.

Au cours des 20 dernières années, le gouvernement fédéral a consulté diverses communautés de recherche ainsi que la Bibliothèque nationale du Canada et les Archives nationales du Canada (désormais Bibliothèque et Archives Canada) au sujet des avantages et des défis liés à la GDR. En 2005, le gouvernement canadien a publié le rapport *National Consultation on Access to Scientific Research Data* (NCASR). Il s'agit du résultat d'un groupe de travail d'expertes et experts comptant plus de 70 têtes dirigeantes du Canada, entre autres dans les domaines de la recherche, de l'administration et des bibliothèques (Strong et Leach, 2005). Le rapport comprenait une recommandation pour la mise sur pied d'un comité directeur national qui verrait à la création d'une archive de données nationale et à la coordination de la gestion des données. Il incluait également une recommandation pour le financement de projets dans tous les domaines au Canada. Toutefois, cette approche a échoué à obtenir du soutien politique (Humphrey, 2012a).

Sans comité directeur national et sans ressources issues du gouvernement fédéral, les bibliothèques universitaires ont dû trouver une autre voie. Elles ont créé des dépôts institutionnels et interinstitutionnels pour diffuser et archiver les données, particulièrement celles des données dites à longue traîne (*long-tail*), soit un grand nombre de jeux de données relativement petits produits par plusieurs disciplines (Heidorn, 2008). Ces derniers sont diversifiés et souvent difficiles à gérer (Cooper *et al*., 2021). Les bibliothèques possédaient une expertise en archivage et en préservation des résultats de recherche et, par leur implication dans le projet de l'IDD, elles se sont investies dans des solutions d'accès et de diffusion des données sous licence. Elles ont été reconnues comme étant bien placées pour relever le défi de la gestion des jeux de données à longue traîne.

## Une approche locale à l'infrastructure canadienne de GDR voit le jour au début des années 2010

En 2008, le Conseil national de recherches du Canada (CNRC) a créé un groupe de travail sur la stratégie de données de recherche pour mettre en œuvre les recommandations du NCASR. Le groupe comptait plus de 70 leaders en recherche scientifique au Canada. En même temps, l'ABRC, un groupe qui représente les plus grandes bibliothèques universitaires au Canada et deux institutions fédérales, a commencé à participer à diverses conversations au sujet de l'avenir de l'infrastructure canadienne de recherche numérique. Graduellement, l'ABRC a argumenté que la GDR, le calcul de haute performance (représentée par Calcul Canada) et le réseau de recherche à haute vitesse (représenté par le Réseau canadien pour l'avancement de la recherche, de l'industrie et de l'enseignement (CANARIE)) devraient être perçus comme des piliers d'une importance équivalente dans une infrastructure de recherche (Humphrey, 2012b).

En 2011, l'ABRC et le groupe de travail sur la stratégie des données de recherche ont organisé un sommet sur les données de recherche qui a mené à la création de Données de recherche Canada (DRC) en 2012. Depuis 2014, le projet est appuyé par CANARIE, un organisme sans but lucratif ayant pour mission de faire fonctionner la structure nationale du réseau de recherche et d'éducation au Canada. DRC a aidé à la mise sur pied de comités et au lancement de projets techniques; il a collaboré avec des organisations internationales pour favoriser l'infrastructure et l'expertise en matière de données de recherche. DRC a coordonné le Sommet sur l'encadrement des services de données nationaux (ESDN), tenu pour la première fois en 2017, puis annuellement de 2019 à 2022. Cet événement réunissait des groupes et des spécialistes en GDR, notamment des organismes de financement, des responsables de la curation de dépôts de données disciplinaires et des bibliothécaires de données de partout au pays. Les discussions ont souligné l'importance d'accorder la priorité à une infrastructure et à des services de GDR coordonnés à l'échelle nationale pour l'avenir de l'infrastructure de recherche numérique canadienne (Attendees of the NDSF Summit, 2019).

Dans le cadre des efforts de l'ABRC pour améliorer le niveau de préparation des bibliothèques aux services de soutien aux données de recherche, un cours en GDR a été offert au début de 2013. Dans le sillage de ce cours est née la Communauté de pratique canadienne en gestion des données de recherche, un forum qui permet de discuter continuellement des activités en GDR au Canada.

Les directrices et directeurs de l'ABRC ont établi des relations concrètes avec les organisations qui offrent l'infrastructure informatique de recherche aux bibliothèques canadiennes, notamment CANARIE, Calcul Canada (calcul à haute performance), Canadian University Council of Chief Information Officers (CUCCIO) et la Bibliothèque scientifique nationale. Le projet pilote ARC, qui devait durer un an, a été mis sur pied en 2014. Il avait pour objectif de favoriser la création d'une communauté de pratique en données de recherche au Canada. Résultat : création d'un réseau d'expertes et experts comprenant des bibliothécaires universitaires, des gens qui développent des systèmes et du code ainsi que des prestataires de services de

données. Le projet ARC s'est avéré un succès et est devenu le réseau Portage en 2015. Sa mission : offrir une direction aux chercheuses et chercheurs du Canada par le biais d'un réseau d'expertes et experts partout au pays. Le 1er avril 2021, Portage a intégré l'Alliance; RDC a fait de même au printemps 2022. À l'heure actuelle, l'Alliance offre une infrastructure ainsi qu'un service de recherche numérique intégrée à l'ensemble des chercheuses et chercheurs universitaires au Canada.

## Politique de GDR nationale de la fin des années 2010 jusqu'au début des années 2020

En 2016, dans le sillon d'étapes entreprises par d'autres pays, les organismes fédéraux de financement de la recherche ont commencé à mettre sur pied une politique en GDR avec la publication d'une Déclaration de principes sur la gestion des données de recherche (https://science.gc.ca/site/science/fr/financement-interorganismes-recherche/politiques-lignes-directrices/gestion-donnees-recherche/declaration-principes-trois-organismes-gestion-donnees-numeriques). Ce document propose aux chercheuses et chercheurs, aux communautés de recherche, aux établissements ainsi qu'aux organismes de financement de collaborer à la création d'un environnement ouvert et solide pour les données de recherche canadiennes.

En 2018, les organismes subventionnaires ont annoncé avoir rédigé une ébauche de politique de GDR et ont entamé une consultation publique. Elles ont reçu plus d'une centaine de commentaires provenant de diverses parties prenantes à propos de la recherche autochtone, la surveillance, la conformité ainsi qu'au sujet des trois piliers de mise en œuvre abordés par la politique : la stratégie de GDR, les PGD et le dépôt de données. En mars 2021, les organismes subventionnaires ont annoncé la création officielle de la **Politique des trois organismes sur la gestion des données de recherche** qui vise l'excellence dans la GDR au Canada. La politique reconnaît toutefois le contexte diversifié des demandes scientifiques disciplinaires, les limites légales et éthiques, les capacités des établissements ainsi que l'**autodétermination** et la mobilisation des communautés autochtones. À la suite de cette annonce tant attendue, la politique a créé un mandat de soutien à la GDR au sein des établissements de recherche.

Ainsi, chaque établissement canadien doit présenter une stratégie de GDR afin que les organismes de financement puissent évaluer le degré de préparation à l'échelle de tous les établissements. Le fait d'élaborer une stratégie de GDR permet aux responsables des établissements de réfléchir aux lacunes locales et de mettre en place des solutions en plus de favoriser la collaboration avec d'autres établissements. La publication de la Politique des trois organismes sur la gestion des données de recherche coïncide avec la création de l'Alliance, un organisme national sans but lucratif ayant pour objectif d'harmoniser et d'améliorer l'accès aux outils et aux services numériques à l'intention de la communauté de recherche canadienne. L'Alliance se concentre principalement sur la création d'un réseau de services de GDR national dans trois domaines : calcul informatique de pointe, gestion des données de recherche et logiciels de recherche.

# Collaboration nationale : du réseau Portage à l'Alliance

## Origine et organisation actuelle

Portage a été mis sur pied par l'ABRC en 2015 en réaction au Plan d'action du Canada pour un gouvernement ouvert; il était le précurseur de l'Alliance. À l'origine, il s'agissait d'un réseau national de services et de soutien en GDR supporté par la communauté  qui tirait profit des réseaux existants, nationaux et régionaux, des bibliothèques universitaires canadiennes. Portage est le fruit de la vision de personnes à l'avant-garde qui se portaient à la défense de la cause de la GDR (Humphrey, 2012b). Le concept initial du réseau a été abordé lors d'une rencontre informelle au congrès de l'ABRC en 2013.

En 2014, l'ABRC a lancé un projet pilote d'un an d'une communauté de pratique appelée ARC. Forte de cette réussite, elle a créé un réseau d'expertes et experts (REE) en GDR basé sur les bibliothèques. Des modèles fonctionnels et une gouvernance ont été définis au cours des deux années suivantes (de septembre 2015 à août 2017) (Humphrey *et al*., 2016). Depuis, le REE a développé et rendu disponible du matériel de formation, des lignes directrices et des modèles en lien avec la GDR, conformément aux exigences des organismes subventionnaires du Canada afin d'appuyer la communauté de recherche et les personnes responsables de l'intendance des données. Le REE a consolidé les liens existants entre les infrastructures de dépôt de données régionales qui se servent du logiciel Dataverse  ce qui a conduit à des partenariats formels et au lancement de Borealis, le dépôt Dataverse canadien (https://borealisdata.ca/fr/). Le REE a également coordonné l'élaboration de l'application Web d'assistance au plan de gestion des données (l'**Assistant PGD**), la création d'un dépôt connu en tant que Dépôt fédéré de données de recherche (https://www.frdr-dfdr.ca/repo/?locale=fr) (DFDR) et la mise en ligne de Lunaris (https://www.lunaris.ca/fr), une plateforme de découverte de données.

Après avoir rejoint les rangs de l'Alliance en avril 2021, la communauté d'expertes et experts de Portage est devenue membre de l'équipe de GDR de l'Alliance. La gouvernance et les activités futures du REE font encore l'objet de discussions. Le réseau compte désormais plus de 140 expertes et experts provenant de 60 établissements partout au Canada. Il collabore avec une vaste gamme de parties prenantes et de partenaires à l'échelle locale, nationale et internationale afin de développer des services et une infrastructure pour que les chercheuses et chercheurs universitaires puissent accéder au soutien dont ils ont besoin en GDR (Humphrey, 2020). Au moment de rédiger ce document, le REE comprend neuf groupes actifs :

1. Groupe d'experts sur la curation (GEC)
2. Groupe d'experts sur la planification de la gestion des données (GEPGD)
3. Groupe d'experts sur les dépôts de données (GEDD)

4. Groupe d'experts sur Dataverse Nord (Dataverse Nord)
5. Groupe d'experts sur la découverte et les métadonnées (GEDM)
6. Groupe d'experts sur la préservation (GEP)
7. Groupe d'experts sur la recherche et l'intelligence (GERI)
8. Groupe d'experts sur les données sensibles (GEDS)
9. Groupe d'experts national sur la formation (GENF)

Les efforts de la communauté d'expertes et experts en GDR ne cessent de progresser grâce aux travaux de l'équipe de GDR de l'Alliance pour élaborer des ressources communes, une expertise et des documents de formation. Les résultats et publications de chaque groupe peuvent être consultés sur le site Web de l'Alliance. Voici quelques faits saillants des réalisations de la communauté.

## Infrastructures, services et outils

Le réseau canadien actuel de collaborations locales et régionales permet d'augmenter l'efficience et de favoriser une infrastructure, des services et des outils de gestion des données. Les spécialistes des données et les bibliothécaires des établissements universitaires canadiens, ainsi que le personnel en GDR de l'Alliance, ont participé au développement et au soutien des infrastructures et des outils de GDR dont il est question dans ce chapitre. Par exemple, le groupe de travail de Dataverse Nord a été mis sur pied pour rassembler les prestataires de dépôts Dataverse et les bibliothécaires du Canada afin de coordonner au niveau local et national et de discuter au sujet de la formation, des services de soutien, des stratégies de sensibilisation, de la promotion, du développement de l'infrastructure et des besoins.

Une infrastructure de gestion des données multifonctionnelle encore plus importante, le DFDR, a été créée avec l'équipe de GDR de l'Alliance comme prestataire de services et Calcul Canada pour l'hébergement du matériel et de l'infrastructure. Le DFDR a également bénéficié du soutien de plusieurs groupes d'expertes et experts, notamment le GEDM, le GEP et le GEC. Aujourd'hui, le DFDR offre une vaste gamme de services en GDR aux établissements, aux organisations et aux chercheuses et chercheurs du Canada, y compris le stockage, la préservation et la curation des données. Tous les membres de la communauté de recherche canadienne peuvent déposer des données ouvertes dans le DFDR et obtenir un **identifiant numérique d'objet** (DOI) afin d'identifier de manière unique leur jeu de données et générer une adresse Web permanente. Le DFDR est également en mesure de gérer les dépôts de données massives et offre un soutien dédié à la curation des données.

Au départ, le DFDR comprenait une fonction d'indexation de données provenant d'autres dépôts de données canadiens, permettant ainsi leur découverte. Toutefois, en 2022, il a été entendu d'approfondir cette capacité en tant que service distinct appelé Lunaris. Il s'agit d'une plateforme bilingue de type guichet unique pour

chercher des données provenant du DFDR et d'autres sources. Lunaris n'héberge pas les données, il offre plutôt des liens vers des dépôts externes où les données peuvent être téléchargées.

La préservation des données de recherche est essentielle pour s'assurer qu'elles demeurent accessibles et utilisables à long terme. Le Canada ne dispose toujours pas d'un plan ou d'une stratégie solide pour la préservation des données de recherche. Le GEP a vu le jour afin d'améliorer la capacité de Portage à développer une infrastructure et des pratiques exemplaires pour préserver les données de recherche et les métadonnées. Ceci comprend une collaboration avec les parties prenantes pertinentes à des projets de développement de logiciels qui ajoutent des plateformes et des services de préservation à l'infrastructure de GDR au Canada. Le GEP a travaillé avec d'autres groupes d'expertes et experts afin de mettre davantage de l'avant les enjeux relatifs à la préservation, a collaboré avec le DFDR et les dépôts Borealis en ce qui a trait à la fonction de préservation des dépôts et a participé avec le DFDR, SciNet, Scholars Portal et les bibliothèques de l'Université de Toronto sur un projet de pipeline de préservation afin de simplifier l'accès pour les chercheuses et chercheurs à des environnements robustes de **préservation numérique** à long terme.

Au départ, l'outil en ligne Assistant PGD était hébergé et administré par l'Université de l'Alberta; plus tard, la section GDR de l'Alliance en est devenue responsable. La Politique des trois organismes sur la gestion des données de recherche met en évidence l'importance des PGD dans le processus de recherche et les définit comme un des trois piliers principaux. Les trois organismes fédéraux de financement de la recherche ont également annoncé que le PGD deviendrait bientôt une exigence – et non plus une recommandation – pour l'ensemble des chercheuses et chercheurs du Canada qui demandent un financement public. Avant cette annonce, le recours aux PGD constituait déjà une exigence standard des demandes de financement américaines et européennes. L'assistant PGD, créé en partenariat avec les trois organismes subventionnaires, offre des conseils détaillés sur la création d'un plan de gestion des données. De plus, les REE ont élaboré plusieurs documents bilingues, notamment des guides pour :

- créer un plan de gestion des données efficace (https://zenodo.org/records/4012728);
- personnaliser le contenu et l'apparence du PGD (https://doi.org/10.5281/zenodo.3966254) (ressource en anglais uniquement).

Il existe plusieurs modèles et exemples de PGD (https://alliancecan.ca/fr/services/gestion-des-donnees-de-recherche/apprentissage-et-ressources/ressources-de-formation) par discipline, qui mettent en évidence les meilleures pratiques en matière de PGD, dans la section des ressources de formation du site Web de l'Alliance.

## Pratiques exemplaires, normes et orientations

En tant que réseau collaboratif national d'expertes et experts, l'Alliance a favorisé la mise en place et la coordination d'une offre d'infrastructure et d'outils en ligne existants et dispersés : l'Assistant PGD, le

Dataverse de Scholars Portal (renommé Borealis, le dépôt Dataverse canadien, en 2022), le DFDR et Lunaris. Il a également préparé des lignes directrices et des recommandations quant aux pratiques exemplaires en GDR dans le cadre d'une collaboration étroite avec les trois organismes fédéraux de financement de la recherche. Les lignes directrices et la documentation élaborées par les groupes de travail en GDR de l'Alliance sont disponibles sur Zenodo (https://zenodo.org/communities/alliancecan?page=1&size=20) et incluent :

- *Guide pour la curation dans Dataverse* (https://doi.org/10.5281/zenodo.5579827), fruit du groupe de travail sur le guide de curation dans Dataverse. Les pratiques exemplaires pour préparer des jeux de données à leur publication dans le dépôt Dataverse sont décrites;
- *Guide des pratiques exemplaires sur les métadonnées de Dataverse Nord* (https://doi.org/10.5281/zenodo.5668962), fruit du groupe de travail sur Dataverse Nord, il est constamment mis à jour. Ce guide donne un aperçu des pratiques exemplaires en matière de métadonnées et donne des exemples issus de plusieurs disciplines, notamment les données géospatiales;
- *Guide d'évaluation pour la préservation des données de recherche* (https://doi.org/10.5281/zenodo.6283886), fruit du groupe de travail sur l'évaluation à des fins de préservation. Ce guide aborde les besoins des personnes qui créent et préservent les données afin d'évaluer et sélectionner les données de recherche qui bénéficieront d'un accès à long terme;
- Boîte à outils pour les données sensibles à l'intention des chercheuses et chercheurs, publiée en 2020 et continuellement mise à jour par le Groupe d'experts sur les **données sensibles**. Ce guide en trois parties comprend un glossaire, une matrice de risques pour les données et un échantillon de vocabulaire de consentement. Nous avons listé et fourni un lien vers chaque partie du guide dans l'encadré qui suit. Il a été adopté à grande échelle par les établissements canadiens.

*Boîte à outils pour les données sensibles – destinée aux  chercheurs Partie 1* (https://doi.org/10.5281/zenodo.4088986): Glossaire terminologique sur l'utilisation des données sensibles à des fins de recherche

*Boîte à outils pour les données sensibles – destinée aux* (https://zenodo.org/records/4107119)*chercheurs Partie 2 (https://doi.org/10.5281/zenodo.4107119)*: Matrice de risque lié aux données de recherche avec des êtres humains

*Boîte à outils pour les données sensibles – destinée aux chercheurs Partie 3* (https://doi.org/10.5281/zenodo.4107186): Langage en matière de gestion de données de recherche pour le consentement éclairé

## Création de réseaux et de communautés

En plus d'offrir une infrastructure et des pratiques exemplaires en GDR, l'Alliance souhaitait lever les obstacles sociaux, culturels et technologiques associés à l'écosystème de GDR (Humphrey, 2012b). Dans les faits, elle a encouragé la création de communautés et de réseaux variés au cours des dernières années.

Les membres du Groupe d'experts sur les dépôts de données (GEDD) de la GDR de l'Alliance ont participé à la mise sur pied du Consortium DataCite Canada (https://www.crkn-rcdr.ca/fr/consortium-datacite-canada), lancé en janvier 2020. Financé par l'Alliance, le consortium est dirigé par l'équipe de GDR de l'Alliance alors que le Réseau canadien de documentation pour la recherche s'occupe de l'administration. Plus de 50 établissements du consortium ont collaboré à l'élaboration d'une structure de gouvernance et de financement et pour offrir à leurs membres des services de production de DOI et d'enregistrement des métadonnées par le biais de DataCite. Ce consortium constitue une réalisation importante pour les établissements canadiens. Il permet de gérer en collaboration le bassin national de DOI pour une gamme de dépôts de recherche et d'autres actifs numériques tout en disposant d'un scénario de tarification stable, commun et collaboratif pour divers paliers d'établissements de recherche au Canada. Grâce à lui, une communauté de pratique peut également résoudre des problèmes techniques et lancer des projets de DOI novateurs au Canada.

Pour aider les dépôts de données de recherche canadiens à harmoniser leurs pratiques aux normes mondiales, le GEDD a offert un financement en GDR de l'Alliance pour une cohorte de candidatures à la certification CoreTrustSeal (CTS). La première cohorte comptait 12 dépôts, dont plusieurs dépôts institutionnels Borealis qui cherchaient à améliorer leurs pratiques. L'obtention de la certification CTS est précédée d'un long processus. Le GEDD a organisé et supervisé des groupes de rédaction et de lecture au bénéfice du personnel associé aux dépôts qui ont postulé et il a procuré du support au cours du processus d'examen par les pairs.

Le Groupe d'experts sur la curation (GEC) se concentre sur la définition, l'évaluation et la promotion de pratiques exemplaires en curation des données. Ceci comprend des techniques, des méthodes et des outils qui peuvent mieux préparer les données et métadonnées, améliorer la qualité des données et, en fin de compte, faciliter la diffusion et la réutilisation des données. Il comble également un besoin en matière de formation et de soutien pour une nouvelle génération de responsables de la curation des données. La création d'une communauté et d'un réseau est un aspect essentiel des approches adoptées par le groupe d'expertes et experts. En 2019, le GEC a tenu le premier forum canadien sur la curation des données en partenariat avec l'Université McMaster et grâce à un financement du CRSH. Ce forum avait pour but principal de mettre sur pied une communauté de pratique nationale qui rassemble des responsables de l'intendance des données, des bibliothécaires, des prestataires de services de données et des gens qui développent des systèmes. Le programme du forum comprenait toute une gamme de conférences, des discussions et des ateliers qui visaient à faciliter la communication et la collaboration en matière de pratiques et de normes de curation des données.

Le forum visait aussi à développer des compétences ainsi que des ressources de formation. Il a connu un immense succès et a atteint son objectif : des responsables de la curation des données ont participé régulièrement à des réunions avec le GEC, depuis, pour discuter et mettre leurs connaissances à jour en matière de curation des données, d'enjeux actuels et de perfectionnement.

## Recherche et formation

Pour garder le rythme avec un environnement en constante évolution, l'équipe GDR de l'Alliance a créé un groupe d'expertes et experts sur la recherche et l'intelligence (GERI) ainsi qu'une équipe de formation pour surveiller les lacunes en matière de GDR et offrir une formation en temps opportun à la communauté et à ses groupes élargis.

Le GERI met l'accent sur la surveillance permanente des sujets et des mandats en lien avec la GDR. Il guide l'élaboration de pratiques exemplaires en GDR au Canada et informe les communautés concernées des enjeux actuels ou à venir en lien avec les politiques et pratiques. Il garde à jour une *Feuille de route des priorités de recherche (https://doi.org/10.5281/zenodo.3963015)* en GDR pour cerner les lacunes en ce qui a trait aux connaissances, aux compétences, aux services et aux politiques en GDR. Le GERI mène également des études indépendantes et effectue des sondages puis en analyse les résultats pour émettre des recommandations à l'équipe GDR de l'Alliance fondées sur des données probantes. En 2016, il a mis sur pied le consortium canadien de sondage en GDR et a élaboré un outil de sondage commun. En tout, 15 universités l'ont utilisé pour sonder les chercheuses et chercheurs de leurs établissements afin de comprendre leurs pratiques et attitudes en matière de GDR. En 2019, le GERI a mené deux sondages auprès d'établissements canadiens pour mesurer leur capacité en GDR et l'état de l'élaboration de leur stratégie; ceci s'est déroulé avant l'annonce de la Politique des trois organismes en GDR. Les résultats ont démontré qu'il existait des initiatives et des services de GDR et le GERI s'est fait la voix des établissements au sujet de leurs priorités ainsi que leurs besoins de soutien supplémentaires en GDR.

Alors que le paysage de la GDR poursuit son évolution, il est essentiel que les chercheuses et chercheurs, spécialistes des données et autres personnes impliquées en GDR possèdent l'information et reçoivent la formation nécessaire pour être au courant des derniers développements ainsi que des pratiques exemplaires. La création de ressources de formation en GDR a constitué une des principales activités de l'équipe GDR de l'Alliance. Depuis 2017, le Groupe d'experts national sur la formation (GENF) a préparé des documents de formation sur le sujet. Les membres supervisent plusieurs projets collaboratifs précis qui permettent d'élaborer et d'offrir des ressources dans le but de soutenir le perfectionnement de compétences en GDR à l'échelle du Canada. Tout de suite après l'annonce de la Politique des trois organismes en GDR, le GENF a coordonné une série d'ateliers populaires sur les aspects les plus importants de la politique. Ces ateliers ont permis aux chercheuses, chercheurs et aux parties prenantes de comprendre les exigences de la politique et de

se sensibiliser aux ressources ainsi qu'aux outils existants qui peuvent les aider à développer un PGD et des stratégies institutionnelles de GDR.

## Services de dépôt de données dans les bibliothèques canadiennes

Tout comme les réseaux d'expertes et experts, la formation et le soutien mis en place à l'échelle nationale, diverses bibliothèques universitaires ont développé ensemble un service de dépôt de données canadien en utilisant Dataverse (https://support.dataverse.harvard.edu/), un **logiciel ouvert** développé par l'Institute for Quantitative Social Science de Harvard. Il a pour objectifs de stocker, partager, citer, préserver, découvrir et d'analyser les données de recherche. Sa nature ouverte permet aux établissements d'héberger leurs propres installations du logiciel Dataverse et d'offrir une solution adaptée aux besoins de leur propre communauté.

Les installations locales et régionales du logiciel Dataverse au Canada, notamment le Dataverse de Scholars Portal et ceux d'autres établissements et régions, sont passées à un service national appelé Borealis. Scholars Portal a commencé à offrir le service Dataverse en dehors du Ontario Council of University Libraries (OCUL) en 2019 et un service national officiel a été offert en 2020 (https://ocul.on.ca/sites/default/files/Scholars%20Portal%20Annual%20Report%202020-2021.pdf) (document en anglais uniquement) par le biais d'accords avec les quatre consortiums régionaux de bibliothèques universitaires. La nouvelle marque Borealis a été lancée en 2022. L'installation nationale partagée offre la possibilité de créer une marque locale et de fournir des ressources communes de formation aux gens qui utilisent le service. Durant la transition, le Groupe d'experts sur Dataverse Nord a commencé à créer des ressources de formation, à offrir du soutien, faire de la sensibilisation et mettre sur pied des stratégies de promotion. La sensibilisation et la promotion sont importantes, car les universités canadiennes préfèrent souvent stocker les données sur des serveurs hébergés localement.

Les données peuvent être téléversées dans des collections Dataverse en tant qu'éléments d'un réseau plus vaste. Une collection Dataverse contient des jeux de données (données de recherche, code, documentation, métadonnées) et peut être définie pour une chercheuse ou un chercheur, une faculté, une revue savante ou une organisation. Par exemple, un chercheur peut téléverser des données dans la collection de son établissement qui, avec l'ensemble des autres collections institutionnelles, forme Borealis, le dépôt national. Chercheuses et chercheurs, équipes et établissements peuvent créer leur propre compte et téléverser leurs données dans une collection institutionnelle (définie par leur affiliation) ou une collection liée à un projet de recherche, si disponible. Les bibliothécaires et personnes responsables de l'intendance des données peuvent également effectuer la curation des jeux de données et gérer la soumission des données au nom des chercheuses et chercheurs. Le logiciel Dataverse est très souple à ce chapitre. Il est possible d'appliquer une marque locale aux collections et aux sous-collections Dataverse.

Le logiciel Dataverse offre aussi une fonction d'analyse de données dans le navigateur; par conséquent, les personnes qui l'utilisent n'ont pas nécessairement besoin de télécharger les fichiers de données pour les consulter. Les fichiers de données tabulaires téléversés dans le système peuvent être analysés grâce à l'outil Web d'analyse et de visualisation des données. Il est possible d'intégrer Dataverse à d'autres ressources documentaires afin d'améliorer la découverte des données. Par exemple, depuis que tous les partenaires de UBC Abacus Dataverse (bibliothèques de l'Université de Victoria, de l'Université du nord de la Colombie-Britannique et de l'Université Simon-Fraser) se servent de ProQuest Summon comme outil de recherche, les collections Dataverse de leurs bibliothèques respectives deviennent accessibles par le biais de flux du protocole Open Archives Initiative (OAI). Chacun des flux OAI comprend toutes les données provenant des établissements partenaires ainsi que l'information appropriée à propos des licences. Grâce à un processus de découverte amélioré (particulièrement par l'attribution de DOI pour des jeux de données de recherche), il est plus facile pour les chercheuses et chercheurs d'accéder aux données et de les réutiliser (p. ex., dans ORCID, Google, DataCite, Google Data Search, Crossref, entre autres services). Ceci facilite les citations et améliore les profils d'impacts, et ce, tant pour les individus que pour les établissements.

Le logiciel de dépôt Dataverse s'est avéré une plateforme souple qui peut prendre en charge plusieurs modèles de services en GDR proposés par les bibliothèques au Canada. Il offre une vaste gamme de caractéristiques qui peuvent améliorer la découvrabilité des données et leur accès ainsi qu'une excellente gestion des données pour la préservation. Toutefois, le logiciel Dataverse ne constitue pas un système de préservation numérique complet (bien que Borealis prenne en charge la préservation numérique au niveau du bit — vous trouverez une explication à ce sujet dans le chapitre « Préservation numérique des données de recherche » et dans le plan de préservation de Borealis (https://borealisdata.ca/planpreservation/)). Le dépôt ne tient pas compte du format et accepte tous les types de fichiers, non seulement les données tabulaires.

Le Ontario Council of University Librairies a parrainé le travail d'Artefactual pour développer une intégration technique (https://www.archivematica.org/en/docs/archivematica-1.14/user-manual/transfer/dataverse/) (document en anglais uniquement) entre le logiciel Dataverse et Archivematica, un outil ouvert et robuste pour le traitement des objets numériques à des fins de préservation et d'accès. Il est possible d'utiliser ce système de préservation conjointement avec le service Borealis en place ou toutes autres installations Dataverse (à partir des versions 1.8 Archivematica et 4.8.6 de Dataverse).

Le soutien à la GDR au Canada fait l'objet d'une attention nationale. Historiquement, et à l'heure actuelle, les régions et les communautés sont confrontées à des enjeux relatifs au soutien et à l'infrastructure en fonction de leurs propres réseaux, du financement régional ou provincial et de la participation aux décisions des consortiums par région.

## Souveraineté des données autochtones

Plusieurs initiatives et développements mentionnés dans ce chapitre, tout comme d'autres dont il sera question dans ce manuel, ont eu lieu sans tenir compte des peuples autochtones et de leurs données ou sans aborder les injustices historiques à leur égard. En fait, les communautés autochtones sont depuis longtemps maltraitées et négligées en ce qui a trait à la recherche canadienne. Si la Politique des trois organismes sur la gestion des données de recherche aborde désormais clairement les éléments à prendre en considération quant aux données autochtones et que des expertes et experts en données autochtones font partie du Groupe d'experts sur les données sensibles, nous encourageons l'équipe de GDR de l'Alliance à aborder ces enjeux de manière plus exhaustive à court terme.

Les universitaires et les personnes qui militent pour les droits des Premières Nations ont réagi à ces lacunes. Par exemple, le Centre de gouvernance de l'information des Premières Nations (CGIPN), un organisme à but non lucratif, a vu le jour en 2010. Ses premiers travaux remontent à 1996, lorsque l'Assemblée des Premières Nations a mis sur pied un comité directeur national ayant pour mandat de créer un sondage sur la santé des Premières Nations (l'Enquête régionale sur la santé des Premières Nations) à la suite de la décision du Canada d'exclure les populations vivant sur une réserve de projets de cueillette de données longitudinales d'envergure. En 1998, le comité a défini les Principes de **PCAP®** (pour propriété, contrôle, accès et possession) en tant qu'outil et norme pour recueillir et gérer les données des Premières Nations. Pour obtenir plus d'information sur les Principes, consultez le chapitre « Souveraineté des données autochtones ».

# Efforts régionaux

Partout au Canada, des établissements ont adopté des approches particulières pour développer et élargir les services de GDR en fonction de leur taille, des ressources disponibles (ressources humaines et infrastructure) et de leurs axes de recherche. Les bibliothécaires et spécialistes des collèges et des universités sont des membres importants des groupes de travail et des comités sur la GDR institutionnelle et participent à l'élaboration de politiques et de stratégies en GDR.

Plusieurs établissements au pays ont participé à des sondages sur les pratiques et les besoins en GDR, des sondages qui reposaient sur un outil commun créé par des bibliothécaires de l'Université de Toronto en 2015. Par la suite, il a été modifié par plusieurs établissements. Résultat : une compréhension plus riche des pratiques disciplinaires en GDR ainsi que des besoins locaux et nationaux. En outre, il a aidé les chercheuses et chercheurs à mieux connaître les pratiques exemplaires en GDR (Cheung *et al.*, 2022).

Des cours en GDR ont été offerts dans les écoles de bibliothéconomie à l'échelle du pays. Comme nous le mentionnions plus tôt, des régions ont adapté le logiciel de dépôt Dataverse au niveau local et, dans de

nombreux cas, au niveau national. Toutes les régions sont représentées au sein des comités de GDR de l'Alliance. Certains établissements ont réagi à la nécessité d'offrir un soutien en GDR par la création de postes de bibliothécaires ou de rôles en appui à la GDR dans les bibliothèques. Vous trouverez ci-après des initiatives régionales qui mettent en évidence des services et des domaines d'intérêts particuliers.

## La GDR dans la région atlantique

Le CAAL/CBPA (https://caul-cbua.ca/) (Council of Atlantic Academic Libraries/Conseil des bibliothèques postsecondaires de l'Atlantique, anciennement CAUL-CBUA) est le réseau des bibliothèques d'universités et de collèges publics dans la région de l'Atlantique. Il met l'accent sur la mise sur pied et la coordination d'activités de préservation numérique dans la région. Le Digital Preservation and Stewardship Committee (https://caul-cbua.ca/committee/digital-preservation-and-stewardship-committee) (DPSC) a été créé en 2013. Plus tard, ses travaux ont englobé le développement de services de GDR à grande échelle afin d'harmoniser ses travaux à la vision nationale. La plus récente initiative comprend la Subvention pour l'innovation du CAAL/CBPA 2020. Elle a permis d'offrir des ateliers à l'ensemble des établissements de la région atlantique qui ont pu les visionner. Les membres du DPSC ont dirigé l'organisation et la réalisation des ateliers. Ces événements étaient intitulés *Atlantic RDM days* (Journée de la GDR au Canada atlantique) et se déroulaient en français et en anglais. Ils étaient importants pour les collèges et universités qui ne disposaient pas des ressources pour prendre en charge la GDR au niveau de l'établissement, mais qui devaient tout de même se conformer à la Politique des trois organismes sur la gestion des données de recherche et favoriser les pratiques exemplaires en GDR au sein de leur communauté de recherche.

En 2015, l'Université Dalhousie a été l'un des premiers établissements de la région atlantique à mettre sur pied une équipe de GDR, laquelle comprenait des partenaires de différents services (Bureau de la recherche, services de technologie universitaire et bibliothèques de l'Université Dalhousie). Elle a été le premier établissement canadien à élaborer puis à publier une stratégie de GDR, comme l'exige la Politique des trois organismes sur la gestion des données de recherche. Désormais, l'université offre un cours en GDR (*Managing Research Data*).

Plusieurs établissements de la région atlantique se sont joints au dépôt national Borealis afin d'offrir des services d'archivage de données à leur communauté de recherche locale. D'autres ont accepté de conserver leurs propres instances de dépôts Dataverse sur leurs serveurs institutionnels locaux. Ceci s'explique par la disponibilité des ressources institutionnelles locales à entretenir le dépôt et à le garder à jour. Par exemple, depuis 2018, les bibliothèques de l'Université du Nouveau-Brunswick ont hébergé un dépôt Dataverse local (https://dataverse.lib.unb.ca/). L'hébergement et l'entretien sont réalisés de manière indépendante grâce à la collaboration de l'équipe système des bibliothèques et le comité des services de GDR des bibliothèques.

Comme d'autres établissements canadiens, toutes les universités de recherche dans la région atlantique ont accès à l'infrastructure nationale d'archivage de données DFDR.

## La GDR au Québec

Depuis les années 1960, les bibliothèques universitaires du Québec collaborent au sein du Bureau de la Coopération Interuniversitaire (BCI), anciennement appelé la Conférence des Recteurs et des Principaux des Universités du Québec (CREPUQ). En 1967, le Comité de coordination des bibliothèques a été créé et, quelques années plus tard, il est devenu le Sous-comité des bibliothèques (Roy et Bégin, 1969). En 2023, le Sous-comité est devenu une entité à part entière, le Partenariat des bibliothèques universitaires (PBUQ).

Au sein même du sous-comité des bibliothèques, un comité centré sur la GDR a été lancé en 2015 et ses membres sont à l'origine de différentes initiatives de promotion et de soutien : Semaine des données à cœur (https://libguides.pbuq.ca/friendly.php?s=GDR/semaine-donnees-a-coeur) (version francophone de la *Love Data Week*), LibGuide (https://uquebec.libguides.com/gdr/introduction) de ressources en GDR pour le réseau des universités du Québec grâce à une subvention du CRSH et une grande implication dans la traduction de ressources canadiennes en français. L'accent sera probablement mis sur le fait de garder le rythme des besoins croissants au cours des prochaines années.

La communauté québécoise de GDR Québec a également participé à d'importants projets d'infrastructure, soit l'internationalisation de Dataverse et GeoIndex.

L'internationalisation de Dataverse s'est déroulée en deux étapes : la première a commencé en 2015 et la deuxième, quelques années plus tard (Bilodeau, 2018). Marie-Hélène Vézina, une bibliothécaire à l'Université de Montréal avec de l'expérience en développement de projets numériques, s'est jointe au personnel de Scholars Portal. Avec le soutien de la communauté de Dataverse plus large, notamment l'Institute for Quantitative Social Science de Harvard, l'équipe a internationalisé le logiciel Dataverse. Si une partie de la traduction avait déjà été effectuée, rien n'était en place pour prendre en charge le multilinguisme. Le code développé a été intégré au code de base de Dataverse, ce qui a permis le déploiement d'une installation bilingue (anglais et français) par Scholars Portal. L'Université de Montréal a fourni la traduction en français. Scholars Portal et les établissements du BCI ont finalisé une entente formelle qu'ils ont signée au printemps 2019. Les premières collections de Dataverse institutionnelles ont été rendues accessibles aux chercheuses et chercheurs à l'été 2022 (Vézina, 2022).

Le projet GeoIndex a été lancé en 2006 pour les besoins de diffusion des données géospatiales à l'Université Laval. La même année, le projet a gagné un prix ESRI Canada et en 2012, le prix Innovation des services documentaires du Québec. Un important accord entre le Bureau de coopération interuniversitaire (BCI) et le ministère de l'Énergie et des Ressources naturelles (MERN) a mené à l'ouverture de la plateforme aux autres

universités québécoises en 2019 afin qu'elles déposent et partagent leurs données. GeoIndex est unique par rapport à d'autres plateformes de données géospatiales en raison de son utilisation du répertoire des vedettes-matière et de son mode de recherche basé sur le contour géospatial des données. En 2021, un module s'est ajouté pour les photos aériennes, permettant de rassembler l'ensemble de l'inventaire des universités québécoises à un seul endroit.

## La GDR en Ontario

L'Ontario compte 23 universités publiques et 24 collèges. Depuis les années 1960, les bibliothèques de ces établissements ont collaboré avec succès par l'intermédiaire du Ontario Council of University Libraries (OCUL). À ses débuts, l'OCUL était impliqué dans des services traditionnels de bibliothèque, comme l'octroi de licences en consortium à des publications savantes et la facilitation d'un partage efficace des ressources. Au cours de ces premières années, plusieurs établissements ont développé leur propre système de dépôt de données, y compris l'archive de données en sciences sociales de l'Université Carleton fondée en 1965 dans la faculté de sociologie et d'anthropologie; la bibliothèque des ressources en données de l'Université Western, lancée à la fin des années 1970 et qui collabore avec le laboratoire informatique des sciences sociales pour diffuser et archiver divers projets de recherche de la faculté; la bibliothèque de données et de cartes de l'Université de Toronto, établie en 1988, dont les services comprennent l'acquisition et la préservation de jeux de données produits par les chercheuses et chercheurs de l'établissement.

En 2002, l'OCUL a formé Scholars Portal, une infrastructure technologique partagée qui héberge les collections numériques en nombre croissant de l'OCUL et y donne accès. Alors que les services de données gagnent en importance, les bibliothèques de l'Ontario ont perçu l'occasion de collaborer sous l'égide de l'OCUL pour améliorer les services, réduire le dédoublement des efforts et mieux gérer les ressources limitées. Au cours de la dernière décennie, l'OCUL a entrepris plusieurs projets d'infrastructure de données ayant connu du succès, y compris l'élaboration d'Odesi, un portail collaboratif de données en sciences sociales, et Scholars GeoPortal (https://geo1.scholarsportal.info/), un portail de données géospatiales. Bien que ces deux portails de données contiennent certaines données de recherche, ils sont conçus comme des collections de données publiées provenant de sources officielles faisant autorité telles que les agences statistiques gouvernementales. Ainsi, ils ne sont pas propices à l'inclusion à grande échelle des productions de données de recherche institutionnelle des établissements membres. Ces systèmes se concentrent sur la découverte et l'accès plutôt que sur la préservation à long terme (Moon, 2014).

Pour cette raison, l'Ontario et le Canada avaient besoin d'autres solutions pour répondre à la demande croissante de dépôts de données de recherche en bibliothèque. En 2011, Scholars Portal s'est joint au projet pilote de la bibliothèque de l'Université de Colombie-Britannique, a installé un dépôt Dataverse puis l'a mis à la disposition des membres de l'OCUL. Ce projet avait pour but d'aborder le besoin exprimé par la

communauté d'avoir un service de dépôt situé en Ontario qui permettrait aux chercheuses et chercheurs d'effectuer de façon autonome un dépôt grâce à un service en ligne. Le logiciel Dataverse a été choisi en raison de sa prise en charge des données de recherche ainsi que de l'intégration à la plateforme du schéma de métadonnées du **Data Documentation Initiative** (DDI). Le personnel de Scholars Portal a préparé des documents et du matériel de formation et a formé les membres du personnel des bibliothèques de l'OCUL quant aux avantages d'intégrer le logiciel Dataverse dans leurs services offerts pour la gestion et le dépôt des données de recherche. Résultat : le dépôt Dataverse de Scholars Portal, désormais nommé Borealis, permet à certaines bibliothèques de l'OCUL de lancer des services de GDR sans devoir disposer d'une infrastructure technique ni du personnel pour soutenir les dépôts par elles-mêmes. Les modèles de service varient d'une bibliothèque à l'autre, du dépôt libre-service à la curation prise en charge par la bibliothèque. De nos jours, le service a évolué grandement. Plusieurs autres établissements de partout au pays s'y sont joints ou ont migré leurs données de recherche vers Borealis, ce qui en fait le centre national pour l'archivage des données de recherche. Le soutien à l'utilisation de Borealis est en grande partie assuré localement par le personnel de bibliothèque et est indépendant de l'infrastructure hébergée et supportée par Scholars Portal.

La communauté de données de l'OCUL, dont l'objectif initial était d'aborder l'accès aux données de l'IDD de Statistiques Canada, a évolué pour devenir un forum de soutien à la GDR. Les expertes et experts issus des établissements universitaires de l'Ontario sont devenus des membres clés de la communauté de GDR de l'Alliance et de ses groupes de travail.

## La GDR dans les Prairies

Les établissements des provinces des Prairies ont exercé une grande influence sur les collaborations nationales en GDR au cours de la dernière décennie. Au début de 2015, les bibliothèques de l'Université de l'Alberta ont mis en œuvre la première instance canadienne d'un outil en ligne ouvert pour aider les chercheuses et chercheurs à rédiger des PGD. À l'époque, un code DMPOnline d'origine britannique était utilisé; l'Université de la Colombie-Britannique et l'Université de l'Alberta ont été les premiers établissements à en adapter la version canadienne.

Presque immédiatement, le projet a été adapté par d'autres établissements canadiens au sein du cadre Portage de l'ABRC et il a été appelé DMP Builder. Plus tard, au cours du cycle de vie de l'outil, il est devenu l'Assistant PGD qui comprend des options en anglais et en français afin de mieux servir la communauté universitaire francophone. Plus de 50 établissements canadiens se servent désormais de l'Assistant PGD avec des lignes directrices adaptées à chaque établissement par le REE en GDR de l'Alliance. Voilà plus d'une décennie que les bibliothèques de l'Université de l'Alberta commanditent l'Assistant PGD pour la communauté de GDR canadienne, qui lui en est très reconnaissante.

Depuis la fin de 2015, le service des technologies dédié à la recherche de l'Université de la Saskatchewan

(USask) a mis en œuvre une initiative semblable en partenariat avec le bureau du vice-président à la recherche. Grâce au financement de Calcul Canada, l'équipe de l'USask a été choisie pour créer une interface nationale de découverte de données de recherche au Canada. Elle est la principale responsable du développement et du fonctionnement de la plateforme Lunaris, désormais sous l'égide de l'Alliance. En 2016, elle a adapté le code source de base, celui-ci étant ouvert, des *Open Collections (https://github.com/ubc-library/docs-open-collections-api)* de la bibliothèque de l'Université de la Colombie-Britannique comme interface de découverte principale ainsi que le code de base ouvert de Geodisy (https://researchcommons.library.ubc.ca/geodisy-phase-2/), également le fruit des bibliothèques de l'Université de la Colombie-Britannique, comme interface de découverte de données fondées sur les cartes. À l'aide du logiciel ouvert Archivematica, l'équipe de l'USask a également développé une excellente collaboration avec la plateforme Globus Connect (https://www.globus.org/globus-connect) pour travailler avec des données massives et préserver les données de recherche numériques.

## La GDR en Colombie-Britannique

Depuis longtemps, les établissements de Colombie-Britannique participent à la GDR avec à leur tête l'Université de la Colombie-Britannique (UBC), l'Université Simon-Fraser (SFU) et l'Université de Victoria (UVic). La bibliothèque de l'UBC est l'une des plus grandes bibliothèques universitaires au Canada. Elle réalise des activités de GDR ad hoc depuis le début des années 1970. En 2008, afin d'aider les établissements régionaux de plus petite envergure, l'UBC a conclu une entente pour rendre le dépôt de données Abacus accessibles à d'autres universités de la province. Au moment de la rédaction du présent document, quatre grandes bibliothèques de recherche universitaire en Colombie-Britannique (Université Simon-Fraser, Université de Victoria, Université du nord de la Colombie-Britannique, Université de la Colombie-Britannique) se servent de l'instance du dépôt Dataverse de l'UBC en tant que dépôt de données sous licence.

Les données sont mises à la disposition des personnes utilisatrices de chaque établissement conformément à leurs licences en utilisant la Fédération canadienne d'accès, un organisme qui gère les identités numériques en éducation supérieure et en recherche par le biais d'un cadre de confiance pour le contrôle des accès. L'équipe des données de la bibliothèque de l'UBC offre une formation élémentaire et avancée sur le dépôt Dataverse aux groupes, aux facultés et aux laboratoires sur le campus de l'université et à ses partenaires dans d'autres bibliothèques universitaires et établissements de recherche. Après la formation, ces groupes devraient gérer leurs propres données dans les collections Dataverse qui leur ont été affectées. La UBC Library School (aussi appelée iSchool) a été parmi les premiers établissements canadiens à offrir un cours de premier cycle en gestion des données de recherche.

Les bibliothèques de SFU et d'UVic ont également apporté beaucoup au paysage de la GDR au Canada. Au début des années 2010, la bibliothèque de SFU a développé Radara, son propre dépôt de données de recherche fondé sur Islandora (désormais déprécié et remplacé par le DFDR) et est devenue le chef de file

canadien en cryptage à divulgation nulle de connaissance (https://www.lib.sfu.ca/help/publish/research-data-management/frdr-encryption) (document en anglais uniquement) pour les données sensibles. Les bibliothèques de l'UVic ont également fait des expériences en matière de services de GDR et ont pu répondre aux besoins des équipes de recherche en matière de licences uniques, notamment pour les jeux de données d'Inforoute Santé Canada (https://borealisdata.ca/dataverse/canadahealthinfoway).

## La GDR dans le nord du Canada

Le nord du Canada est composé de trois territoires : les Territoires du Nord-Ouest, le Nunavut et le Yukon. Les deux établissements de recherche situés dans cette région sont l'Université du Yukon et le Collège Aurora. Dans le cadre de la stratégie de GDR institutionnelle (mandatée par la Politique des trois organismes sur la gestion des données de recherche), les bibliothécaires de l'Université du Yukon et les membres du bureau des services à la recherche ont collaboré à la création d'un dépôt institutionnel, Arca, hébergé par la British Columbia Electronic Library Network (BC ELN). Ce réseau est une initiative pour les dépôts numériques en Colombie-Britannique fondés sur le logiciel Islandora qui vise principalement les établissements et des collèges de plus petite envergure. Les résultats de recherche téléversés par les chercheuses et chercheurs de l'Université du Yukon dans BC ELN Arca (https://bceln.ca/category/projects/arca) sont collectés par Lunaris.

En octobre 2022, les bibliothécaires du Collège Aurora à Inuvik ont participé à un panel sur les stratégies institutionnelles organisé par l'Alliance. Ce fut l'occasion pour les bibliothécaires de communiquer leur expérience unique pour traiter les enjeux en matière de GDR dans un petit établissement du nord. Certains établissements du nord du Canada, comme l'Université du Yukon, collaborent avec d'autres universités et collèges en Colombie-Britannique pour élaborer leurs stratégies institutionnelles en GDR, conformément à la Politique des trois organismes sur la gestion des données de recherche. Ils participent à un groupe ad hoc pour créer des plans d'action et partager leurs visions des services en GDR dans les établissements de plus petite envergure.

# Conclusion

Il s'agit d'une période passionnante pour la GDR au Canada et il a fallu des années de travail dévoué et une collaboration provinciale de haut niveau pour y arriver. Les bibliothèques entrevoient les possibilités d'interagir avec leurs communautés et entre elles. Évidemment, ces occasions s'accompagnent de défis, comme une infrastructure numérique dispendieuse qu'il faut constamment gérer. Nous croyons que Portage et la formation de l'Alliance offrent le meilleur potentiel pour répondre à des besoins jusque-là insatisfaits. Toutefois, pour y arriver, il faudra disposer d'un financement durable.

Le développement d'outils ouverts, d'une infrastructure et de services de soutien en matière de GDR est essentiel pour que les universitaires du Canada puissent réussir à intégrer ces activités à leurs processus de travail. Les bibliothèques universitaires disposent d'une histoire de prise en charge de l'accès aux données, de leur diffusion et de leur préservation; elles ont pour mandat de participer à la préservation des résultats de recherche de leur communauté (p. ex., dépôts institutionnels). Les bibliothèques peuvent diriger l'adoption de pratiques exemplaires et de normes ouvertes. Elles peuvent également établir des partenariats avec toute une gamme de parties prenantes pour l'élaboration d'une infrastructure et d'outils. La communauté des bibliothèques canadienne favorise activement le partage des données de recherche depuis les années 1960 et est en bonne posture pour jouer un rôle de leadership à ce chapitre à l'avenir.

**Questions de réflexion**

1. Qu'avez-vous appris au sujet de la communauté des données au Canada?
2. Selon vous, comment  la communauté des données des bibliothèques universitaires canadiennes se compare-t-elle aux autres secteurs de la bibliothéconomie universitaire?
3. À la lumière du mouvement international en matière de science ouverte, quels défis percevez-vous en gestion des données de recherche aujourd'hui?
4. Quels partis ou organismes sont les mieux placés pour soutenir la GDR auprès des chercheuses et chercheurs canadiens?

**Éléments clés à retenir**

- Les services de données, la sensibilisation, les infrastructures, les outils et la culture de GDR en général ont évolué au cours des décennies et ce, tant sur le plan local, régional, que national.
- Les bibliothécaires de données, les spécialistes des données, les consortiums de

bibliothèques, les agences de financement public et les organismes gouvernementaux jouent un rôle clé dans la définition des besoins et dans l'élaboration des services en GDR.
- Pour favoriser les pratiques exemplaires en gestion des données en soutien aux services de GDR, le gouvernement, les établissements, les prestataires de services et la communauté de recherche doivent continuer à collaborer à toutes les étapes du cycle de vie de la recherche.
- Plusieurs outils et infrastructures techniques existent pour soutenir la GDR et ils devront évoluer afin de répondre aux besoins actuels et nouveaux.

## Remerciements

La première version de la section sur la GDR au Québec a été rédigée par Ève Paquette-Bigras, bibliothécaire universitaire à l'Université de Montréal. Les autrices et auteurs remercient Ève d'avoir fourni l'historique et le contexte des réalisations en matière de GDR dans cette province.

## Lectures et ressources supplémentaires

Doiron, J., Neilson, M. et Nicholson, R. (2020). *Data management planning in Canada* [Livre blanc]. NDRIO. https://alliancecan.ca/en/document/261 (https://alliancecan.ca/en/document/261)

Gouvernement du Canada. (2021). *Politique des trois organismes sur la gestion des données de recherche*. https://science.gc.ca/site/science/fr/financement-interorganismes-recherche/politiques-lignes-directrices/gestion-donnees-recherche/politique-trois-organismes-gestion-donnees-recherche (https://science.gc.ca/site/science/fr/financement-interorganismes-recherche/politiques-lignes-directrices/gestion-donnees-recherche/politique-trois-organismes-gestion-donnees-recherche)

Lavoie, B. et Dempsey, L. (2004). Thirteen ways of looking at...digital preservation. *D-Lib Magazine, 10*(7–8). https://doi.org/10.1045/july2004-lavoie (https://doi.org/10.1045/july2004-lavoie)

Moon, J. (2021, 8 novembre). *Mise à jour de Portage et l'Alliance de recherche numérique du Canada*. Alliance de recherche numérique du Canada. https://alliancecan.ca/fr/nouveautes/nouvelles/mise-jour-de-portage-et-lalliance-de-recherche-numerique-du-canada (https://alliancecan.ca/fr/nouveautes/nouvelles/mise-jour-de-portage-et-lalliance-de-recherche-numerique-du-canada)

Read, K., McDonald, G., Mackay, B. et Barsky, E. (2014). A commitment to First Nations data governance: A

primer for health librarians. *Journal of the Canadian Health Libraries Association / Journal de l'Association des bibliothèques de la santé du Canada, 35*(1), 11–15. https://doi.org/10.5596/c14-003 (https://doi.org/10.5596/c14-003)

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J-. W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., ... Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data, 3*, 1–9. https://doi.org/10.1038/sdata.2016.18 (https://doi.org/10.1038/sdata.2016.18)

# Bibliographie

Attendees of the NDSF Summit. (2019). *Kanata declaration (Version 2.0)*. National Data Services Framework Summit 2019 (NDSF 2019), Ottawa, Canada. Zenodo. https://doi.org/10.5281/zenodo.3234815 (https://doi.org/10.5281/zenodo.3234815)

Barsky, E., Laliberté L., Leahey, A. et Trimble, L. (2017). Collaborative research data curation services: A view from Canada. Dans L. R. Johnston (dir.), *Curating research data: Volume one: Practical strategies for your digital repository* (p. 79-101). Association of College and Research Libraries. https://dx.doi.org/10.14288/1.0340778 (https://dx.doi.org/10.14288/1.0340778)

Bilodeau, G. (2018, 3 mai). *Gestion des données de recherche (GDR): Écosystème Canadien – un bref survol* [Présentation]. Semaine de la conduite responsable en recherche, Université Laval, Québec, Canada. https://www.ulaval.ca/sites/default/files/recherche-creation/documents/conduite%20responsable/gestion-donnees-recherche-bilodeau.pdf (https://www.ulaval.ca/sites/default/files/recherche-creation/documents/conduite%20responsable/gestion-donnees-recherche-bilodeau.pdf)

Boyko, E. et Watkins, W. (2011). *The Canadian data liberation initiative: An idea worth considering*. International Household Survey Network, IHSN Working Paper (006). http://www.ihsn.org/sites/default/files/resources/IHSN-WP006.pdf (http://www.ihsn.org/sites/default/files/resources/IHSN-WP006.pdf)

Cheung, M., Cooper, A., Dearborn, D., Hill, E., Johnson, E., Mitchell, M. et Thompson, K. (2022). Practices before policy: Research data management behaviours in Canada. *Partnership: The Canadian Journal of Library and Information Practice and Research, 17*(1), 1-80. https://doi.org/10.21083/partnership.v17i1.6779

Cooper, A., Steeleworthy, M., Paquette-Bigras, È., Clary, E., MacPherson, E., Gillis, L., Wilson, L. et Brodeur,

J. (2021). *Guide pour la curation dans Dataverse*. Zenodo. https://doi.org/10.5281/zenodo.5579820 (https://zenodo.org/records/5579827)

Hill, E. et Gray, S. V. (2016). The academic data librarian profession in Canada: History and future directions. Dans L. M. Kellam et K. Thompson (dir.), *Databrarianship: The Academic Data Librarian in Theory and Practice* (p. 321-334). ACRL. http://ir.lib.uwo.ca/wlpub/49 (http://ir.lib.uwo.ca/wlpub/49)

Hackett, Y. (2001). A national research data management strategy for Canada: The work of the National Data Archive Consultation Working Group. *IASSIST Quarterly, 25*(3), 13-16. https://doi.org/10.29173/iq91 (https://doi.org/10.29173/iq91)

Heidorn, P. B. (2008). Shedding light on the dark data in the long tail of science. *Library Trends, 57*(2), 280–299. https://doi.org/10.1353/lib.0.0036 (https://doi.org/10.1353/lib.0.0036)

Humphrey, C. (2005). Collaborative training in statistical and data library services: Lessons from the Canadian data liberation initiative. *Resource Sharing and Information Networks, 18*(1–2), 167–181. https://doi.org/10.1300/J121v18n01_13 (https://doi.org/10.1300/J121v18n01_13)

Humphrey, C. (2012a, 5 décembre). Canada's long tale of data. *Preserving Research Data in Canada*. http://preservingresearchdataincanada.net/2012/12/05/hello-world (http://preservingresearchdataincanada.net/2012/12/05/hello-world)

Humphrey, C. (2012b, 13 décembre). Research data management infrastructure II. *Preserving research data in Canada*. https://preservingresearchdataincanada.net/2012/12/13/research-data-management-infrastructure-ii/ (https://preservingresearchdataincanada.net/2012/12/13/research-data-management-infrastructure-ii/)

Humphrey, C. (2020). The CARL portage partnership story. *Partnership: The Canadian Journal of Library and Information Practice and Research, 15*(1), 1–7. https://doi.org/10.21083/partnership.v15i1.5825

Humphrey, C., Shearer, K. et Whitehead, M. (2016). Towards a collaborative national research data management network. *International Journal of Digital Curation, 11*(1), 195–207. https://doi.org/10.2218/ijdc.v11i1.411 (https://doi.org/10.2218/ijdc.v11i1.411)

Liss, S. N. (2018, 5 septembre). Addressing gaps in Canadian research data management: A comprehensive guide of the Portage Network. *University Affairs*. https://www.universityaffairs.ca/magazine/sponsored-content/addressing-gaps-in-canadian-research-data-management/ (https://www.universityaffairs.ca/magazine/sponsored-content/addressing-gaps-in-canadian-research-data-management/)

Moon, J. (2014). Developing a research data management service – a case study. *Partnership: The Canadian*

*Journal of Library and Information Practice and Research, 9*(1), 1–14. https://doi.org/10.21083/partnership.v9i1.2988

Roy, J. et Bégin, J.-O. (1969). *Enquête relative à un plan de coordination: rapport*. Comité de coordination des bibliothèques de la CREPUQ.

Steeleworthy, M. (2014). Research data Management and the Canadian Academic library: An organizational consideration of data management and data stewardship. *Partnership: The Canadian Journal of Library and Information Practice and Research, 9*(1), 1–11. https://doi.org/10.21083/partnership.v9i1.2990

Strong, D. F. et Leach, P. B. (2005). *National consultation on access to scientific research data* (Rapport final). Gouvernement du Canada. https://publications.gc.ca/site/eng/272526/publication.html (https://publications.gc.ca/site/eng/272526/publication.html)

Vézina, M.-H. (2022). *Métadonnées bibliographiques des thèses et mémoires du dépôt institutionnel de l'Université de Montréal [Canada]* [Jeu de données]. Borealis. https://doi.org/10.5683/SP3/SJJACL (https://doi.org/10.5683/SP3/SJJACL)

## À propos des auteurs

### Eugene Barsky

Eugene Barsky est bibliothécaire en données de recherche à l'Université de la Colombie-Britannique (UBC). Il travaille avec les chercheuses et chercheurs de l'UBC pour conserver et gérer les données de recherche, de la planification au dépôt en passant par la préservation. M. Barsky a participé à la mise sur pied du service de Dépôt fédéré de données de recherche et collabore avec l'Alliance de recherche numérique du Canada (l'Alliance) ainsi que l'Union européenne (OpenAIRE). Il est le chercheur principal du projet national Geodisy financé par l'Alliance. Parmi les prix qui lui ont été décernés par ses pairs, mentionnons ceux de l'Association des bibliothèques de recherche du Canada, de l'American Society for Engineering Education et la Special Library Association. Il a publié plus de 30 articles révisés par des pairs et il a fait des présentations à plus de 70 conférences. Il est professeur associé à l'iSchool de l'UBC. Il enseigne la gestion des données de recherche et compte parmi les personnes qui ont fondé le réseau d'expertes et d'experts de Portage (maintenant l'Alliance) au Canada. Courriel : eugene.barsky@ubc.ca (mailto:eugene.barsky@ubc.ca) | ORCID : 0000-0002-5119-2271 (https://orcid.org/0000-0002-5119-2271)

### Elizabeth Hill

Elizabeth Hill est la bibliothécaire des données à l'Université Western. Elle y donne des formations en littératie

des données ainsi qu'en accès aux sources de données. Elle agit à titre de conseillère externe auprès de Statistiques Canada. Mme Hill est active dans plusieurs communautés de données et groupes de travail et ce, tant à titre de participante que de cheffe. Ses domaines d'intérêt en recherche comprennent le soutien aux chercheuses et chercheurs. Elle a publié des ouvrages au sujet des systèmes de diffusion de données et de la bibliothéconomie des données au Canada. ORCID : 0000-0002-9715-238X (https://orcid.org/0000-0002-9715-238X)

## Tatiana Zaraiskaya

Tatiana Zaraiskaya est une bibliothécaire en STIM aux bibliothèques de l'Université du Nouveau-Brunswick (UNB) où elle est également responsable de la gestion des données de recherche. Depuis 2016, elle est membre du Groupe d'experts sur la recherche et l'intelligence de l'Alliance de recherche numérique du Canada. Elle a participé à plusieurs sondages menés par ce groupe de travail et compte parmi les personnes ayant réalisé le sondage sur la GDR de l'Université Queen's et de l'UNB. Elle a corédigé plusieurs conférences et autres publications savantes en lien avec la GDR et elle a participé à l'élaboration du modèle de PGD de base de Portage. Elle a obtenu un doctorat en biophysique à l'Université de Guelph et un MLIS à l'Université Western en Ontario. Courriel : t.zaraiskaya@unb.ca (denied:about:blank) | Google Scholar : https://scholar.google.com/citations?user=BB6c8XQAAAAJ&hl=en (https://scholar.google.com/citations?user=BB6c8XQAAAAJ&hl=en) | ORCID : 0000-0001-9294-6052 (https://orcid.org/0000-0001-9294-6052)

## Minglu Wang

Minglu Wang est une bibliothécaire en gestion des données de recherche (GDR) à l'Université York. Elle a publié des articles de recherche, des chapitres de livre, des documents de travail et des communications lors de colloques au sujet des bibliothèques universitaires et des services de GDR. Mme Wang est une membre active de l'Association of College & Research Libraries (ACRL), une division de l'American Library Association. Pendant plusieurs années, elle a rédigé des articles et des livres blancs pour les publications *Top Trends* et *Environmental Scan* de l'ACRL. Elle fait partie du Groupe d'experts sur la recherche et l'intelligence de l'équipe de GDR de l'Alliance de recherche numérique du Canada. Elle a participé à la conception et à la rédaction du rapport sur le sondage sur les capacités en GDR des établissements canadiens. Courriel : mingluwa@yorku.ca (denied:about:blank) | ORCID : 0000-0002-0021-5605 (https://orcid.org/0000-0002-0021-5605)

## Lucia Costanzo

Lucia Costanzo est la bibliothécaire en GDR à l'Université de Guelph. Récemment, elle a terminé un prêt de

service auprès de l'Alliance de recherche numérique du Canada (l'Alliance) en tant que coordonnatrice de l'évaluation, de la recherche et de l'intelligence. Dans le cadre de ce rôle, Mme Costanzo a coordonné les activités du Groupe d'experts sur la recherche et l'intelligence. Ces activités comprenaient informer et conseiller l'équipe de GDR et la direction de l'Alliance relativement aux nouveaux développements et aux nouvelles directions et ce, tant à l'échelle nationale qu'internationale, au chapitre de la GDR et des écosystèmes plus vastes d'infrastructure de recherche numérique. Avant sa période de prêt de service, elle a travaillé pendant plus de 20 ans à l'Université de Guelph à appuyer, rendre accessible et contribuer au processus d'apprentissage et de recherche sur le campus. Courriel : lcostanz@uoguelph.ca (denied:about:blank) | ORCID : 0000-0003-4785-660X (https://orcid.org/0000-0003-4785-660X)

5.

# PARTAGE ET RÉUTILISATION DES DONNÉES DE RECHERCHE AU CANADA : PRATIQUES ET POLITIQUES

Meghan Goodchild; Shahira Khair; Amber Leahey; Kaitlin Newson; et Lee Wilson

**Objectifs d'apprentissage**

À la fin de ce chapitre, vous pourrez :

1. Comprendre les pratiques, les politiques et les services qui guident le partage et la réutilisation des données de recherche au Canada.
2. Déterminer les éléments de l'infrastructure de recherche numérique canadienne, notamment les options de stockage comme les dépôts de données et les plateformes de préservation à long terme, ainsi que les services qui soutiennent l'accès à ces infrastructures et leur utilisation.
3. À l'aide d'études de cas, définir les soutiens et obstacles au partage et à la réutilisation des données tout au long du cycle de vie des données de recherche, en plus des secteurs qui doivent être développés.

## Introduction

Les chercheuses et chercheurs au Canada, toutes disciplines et tous domaines confondus, produisent des quantités de données encore jamais vues (Baker *et al*., 2019). Grâce aux progrès de la science ouverte et des politiques de données ouvertes des maisons d'édition, des organismes de financement de la recherche, des groupes disciplinaires et des établissements, les chercheuses et chercheurs réalisent de plus en plus la nécessité

de gérer leurs données conformément aux politiques connexes en matière de dépôt et de partage de données. Ces politiques soutiennent des buts plus larges en ce qui a trait à la transparence, à la reproductibilité et à la réutilisation (Groupe de travail de l'Alliance en gestion des données de recherche [GT GDR de l'Alliance], 2020). (Consultez le chapitre 12, « Planification de la gestion des données pour les processus de travail en science ouverte, » pour obtenir un aperçu de la science ouverte et des données ouvertes).

Accélérer le progrès scientifique et éviter les collectes de données dispendieuses constituent des éléments importants en faveur du partage et de la réutilisation des données. Le partage des données permet également de reproduire les résultats de recherche, ce qui améliore l'intégrité des résultats publiés et le degré de confiance à leur égard. Lorsqu'il est facile de découvrir des données de recherche et d'y accéder, cela accroît la visibilité et l'impact de la recherche. Qui plus est, le partage des données, des environnements de recherche et des outils favorise et améliore la collaboration, ce qui se traduit par une plus grande **interopérabilité** et des économies en recherche.

Dans le but d'optimiser les avantages du partage et de la réutilisation des données, les résultats des données de recherche doivent être guidés par les **principes FAIR** – Facile à trouver, Accessible, Interopérable, Réutilisable – abordés au chapitre 2 (Wilkinson *et al.*, 2016). De plus, ils doivent être appuyés par une infrastructure et des services de soutien en recherche numérique selon les principes TRUST – Transparence, Responsabilité, Orientation vers l'utilisateur, Durabilité et Technologie (*Transparency, Responsibility, User focus, Sustainability and Technology*) (Lin *et al*., 2020). Par conséquent, le partage de données devient une partie intégrante de la recherche de haute qualité, ce qui exige la mise en pratique continue de la **gestion des données de recherche** (GDR). Des services de GDR émergent au Canada dans toutes les disciplines, dans les établissements ainsi qu'aux paliers régional et national afin d'appuyer les chercheuses et chercheurs en matière de partage et de réutilisation des données.

Dans le cadre de ce chapitre, vous apprendrez au sujet des politiques et des pratiques, de l'infrastructure de recherche numérique ainsi que des outils et des services permettant le partage et la réutilisation des données de recherche au Canada. Nous examinerons les éléments qui soutiennent le **cycle de vie des données** ainsi que les services relatifs à la curation et à la préservation des données. Enfin, nous aborderons des études de cas afin de mettre en évidence des pratiques de partage et de réutilisation des données et des défis disciplinaires.

# Politiques et pratiques au Canada

## Organismes de financement de la recherche

Les organismes de financement et les gouvernements de partout dans le monde ont reconnu la nécessité d'établir des politiques de GDR afin de soutenir l'accès aux données financées par des fonds publics. Les

mandats des organismes de financement qui exigent le partage des données influencent le comportement des chercheuses et chercheurs ainsi que la demande pour une infrastructure et des services en GDR (GT GDR de l'Alliance, 2020). La Politique des trois organismes sur la GDR au Canada (https://science.gc.ca/site/science/fr/financement-interorganismes-recherche/politiques-lignes-directrices/gestion-donnees-recherche/politique-trois-organismes-gestion-donnees-recherche) (2021) alimente une culture de changement pour le dépôt et le partage de données, car elle définit les exigences en vertu desquelles les chercheuses et chercheurs « sont tenus de déposer dans un dépôt numérique les données de recherche, les métadonnées et les codes qui appuient directement les conclusions de la recherche publiées dans des revues de même que les **préimpressions** découlant de la recherche financée par les organismes subventionnaires » (Gouvernement du Canada, 2021); la mise en oeuvre de cette mesure est à venir. Les titulaires de subventions doivent offrir un accès convenable aux données pour autant que les exigences éthiques, culturelles, juridiques et commerciales le permettent, conformément aux principes FAIR et aux normes propres à leurs disciplines. La souveraineté des données autochtones (abordée en détail au chapitre 3) reconnaît les droits inhérents des communautés autochtones de gouverner la collecte, la propriété et l'utilisation de leurs données, ce qui peut se traduire par des pratiques distinctes en ce qui a trait au partage de leurs données de recherche.

# Politiques des organismes de financement

## Locales et régionales

Les établissements de recherche canadiens peuvent définir leurs propres exigences pour la gestion et le partage des données en fonction de politiques internes qui régissent les pratiques de la recherche et la propriété intellectuelle. De plus, ils doivent publier une stratégie indiquant comment les pratiques de GDR seront prises en charge (Gouvernement du Canada, s.d.).

## Nationales

- Politique des trois organismes sur la GDR (https://science.gc.ca/site/science/fr/financement-interorganismes-recherche/politiques-lignes-directrices/gestion-donnees-recherche/politique-trois-organismes-gestion-donnees-recherche) (2021)
    - Certaines demandes de subvention doivent comprendre un **plan de gestion des données** (mise en œuvre progressive depuis le printemps 2022).
    - Les titulaires de subventions doivent verser dans un dépôt numérique les **données de recherche**, les **métadonnées** et les codes qui appuient directement les conclusions de la recherche publiées dans des revues, de même que les prépublications préimpressions découlant de la recherche financée par les organismes subventionnaires. Le dépôt doit être effectué au moment de la publication (mise en œuvre à venir).

    - Bien que le partage de données ne soit pas exigé, les organismes subventionnaires s'attendent à ce que les chercheuses et chercheurs donnent un accès approprié aux données lorsque les exigences éthiques, culturelles, juridiques et commerciales le permettent et conformément aux principes FAIR ainsi qu'aux normes de leurs disciplines. Dans la mesure du possible, ces données, ces métadonnées et ces codes doivent être reliés à la publication à l'aide d'un **identifiant unique pérenne** (IUP).
- Déclaration de principes des trois organismes en GDR au Canada (https://science.gc.ca/site/science/fr/financement-interorganismes-recherche/politiques-lignes-directrices/gestion-donnees-recherche/declaration-principes-trois-organismes-gestion-donnees-numeriques) (2016)
    - Les données doivent être collectées et stockées en utilisant des logiciels et des formats qui permettent leur stockage sûr ainsi que leur préservation et leur accès bien au-delà de la durée du projet.
- Politique des trois organismes sur le libre accès aux publications (https://science.gc.ca/site/science/fr/financement-interorganismes-recherche/politiques-lignes-directrices/libre-acces/politique-trois-organismes-libre-acces-aux-publications) (2015)
    - Les chercheuses et chercheurs dont les travaux sont financés par les Instituts de recherche en santé du Canada (IRSC) devraient déposer certains types de données (p. ex., bio-informatique) dans des bases de données publiques appropriées.
- Politique sur l'archivage des données de recherche (https://www.sshrc-crsh.gc.ca/about-au_sujet/policies-politiques/statements-enonces/edata-donnees_electroniques-fra.aspx) (1990)
    - Les données de recherche doivent être conservées et rendues disponibles dans les deux années qui suivent l'achèvement du projet (Conseil de recherche en sciences humaines, s.d.).

## Internationales

Plusieurs organismes publics de financement de la recherche dans d'autres pays qui soutiennent les chercheuses et chercheurs du Canada exigent que les jeux de données qui sous-tendent leurs publications de recherche soient publiés. C'est le cas notamment des organisations suivantes :

- Organismes de financement aux É.-U., comme les National Institutes of Health (NIH (https://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm)) et la National Science Foundation (NSF (https://www.nsf.gov/bfa/dias/policy/dmp.jsp))
- UK Research and Innovation funders (https://www.dcc.ac.uk/guidance/policy/overview-funders-data-policies)
- European Commission Horizon 2020

Plusieurs sources privées de financement de la recherche ont leurs propres attentes en matière de partage de

données (p. ex., Wellcome Trust (https://wellcome.ac.uk/grant-funding/guidance/data-software-materials-management-and-sharing-policy), Bill & Melinda Gates Foundation (https://docs.gatesfoundation.org/documents/faq.pdf)).

## Autres politiques et pratiques

Des maisons d'édition ont également encouragé l'adoption de pratiques en GDR. Lorsqu'une déclaration sur la disponibilité des données est exigée, il est beaucoup plus probable que les données de recherche soient partagées en ligne. Lorsque les politiques sont moins rigoureuses, comme le fait de recommander l'archivage de données, les taux d'archivage n'augmentent que légèrement comparativement au fait de ne pas disposer d'une telle politique (Vines *et al*., 2013). Le partage et la disponibilité des données varient selon la discipline. Par exemple, les domaines de la biologie, des sciences de la Terre, des sciences médicales et des sciences physiques présentent un taux supérieur de partage de données (Stuart *et al*., 2018); toutefois, les données sont moins faciles d'accès dans des documents en lien avec l'énergie et la catalyse, la psychologie, l'optique et l'optoélectronique et la foresterie (Tedersoo *et al*., 2021).

Au cours des 20 dernières années, le partage des données s'est amélioré (Tedersoo *et al*., 2021), mais les études démontrent que les résultats ne sont pas toujours entièrement reproductibles à partir des données partagées en raison d'une documentation et de métadonnées inadéquates (Rieseberg *et al*., 2021). Des efforts importants ont été déployés pour atténuer ce phénomène. Par exemple, le *Journal of Molecular Ecology* encourage les autrices et auteurs à utiliser la base de données en **libre accès** GEOME (https://geome-db.org/) pour créer des liens permanents entre les données génétiques et les métadonnées géographiques et écologiques afin que les données versées respectent les principes FAIR (Rieseberg *et al*., 2021). La Public Library of Science (2022) a annoncé le lancement d'un projet pilote de « données accessibles » où certains articles mettront en évidence les liens vers des jeux de données dans des dépôts spécifiques dans le but d'accroître le partage et la découverte de données de recherche et de souligner l'avantage des modèles de science ouverte. L'*American Journal of Political Science*, en partenariat avec le Odum Institute for Research in Social Science, fournit des services de curation et de vérification de données pour faire en sorte que les jeux de données reproduisent les résultats des articles correspondants (Jacoby *et al*., 2017). Par conséquent, les politiques, à elles seules, ne suffisent pas; le recours à des solutions propres aux disciplines est requis pour que les données partagées soient accessibles et réutilisables.

# Infrastructure, outils et services

Une gamme d'infrastructures est nécessaire pour soutenir la production, le partage et la réutilisation des

données tout au long de leur cycle de vie. Ainsi, elles travaillent de concert afin que les données adhèrent aux principes FAIR au-delà de la durée du projet de recherche.

Il existe trois types de stockage de données de recherche : **actif**, **dépôt** et **archive**. La figure 1 définit le stockage actif au cours de la phase de recherche, le stockage dans un dépôt pour la phase d'accès et de publication et le stockage de type archive pour la phase de préservation, laquelle nécessite un traitement supplémentaire afin de soutenir l'accessibilité à long terme.

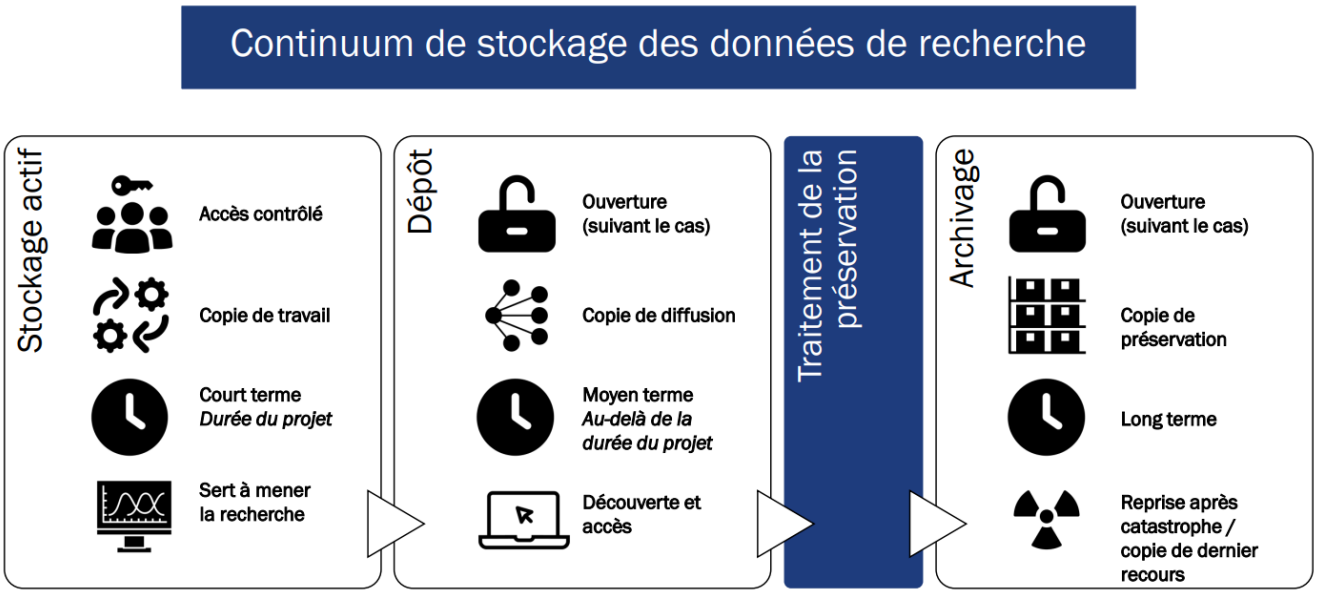



Figure 1. Spectre des possibilités de stockage de données de recherche (GT GDR de l'Alliance, 2020). © Tous droits réservés; réutilisé avec permission.

Le tableau 1 donne des détails sur le stockage actif, de type dépôt ou sous forme d'archive et donne des exemples de ce qui est utilisé au Canada. Le tableau 2 aborde les diverses infrastructures de recherche qui aident au partage, à la réutilisation et à l'accès.

**Tableau 1 : Types de stockage de données de recherche**

| Type | Attributs | Exemples |
|---|---|---|
| Stockage actif | • Prend en charge les données qui doivent changer ou avec lesquelles il faut interagir souvent, de constamment (chaque seconde) à périodiquement (chaque semaine). | • Stockage d'analyses et de calculs scientifiques (p. ex., calcul de haute performance régional et national)<br>• Stockage d'entreprise institutionnel et personnel (p. ex., disques durs)<br>• Stockage infonuagique commercial (p. ex., Microsoft Azure, OneDrive, Google |

| Type | Attributs | Exemples |
|---|---|---|
| | | Cloud, services Web d'Amazon)<br>• Service de stockage et de partage de fichiers (p. ex., Open Science Framework (https://osf.io/), Code Ocean (https://codeocean.com/)) |
| Stockage de type dépôt | • Prend en charge l'intendance et la maintenance de données, des métadonnées et d'autres objets, y compris le code, qui constituent une copie fiable dans le registre scientifique.<br>• Comporte quatre fonctions principales : l'ingestion, la curation, la conservation et l'accès (GT GDR de l'Alliance, 2020).<br>• Assure habituellement un accès via des plateformes logicielles, notamment des portails et des passerelles de recherche. | • Plateformes de dépôt (p. ex., CKAN (https://ckan.org/), InvenioRDM (https://inveniosoftware.org/products/rdm/), The Dataverse Project (https://dataverse.org/), HUBzero (https://hubzero.org/))<br>• Services hébergés (p. ex., GitHub (https://github.com/), Zenodo (https://zenodo.org/), Dépôt fédéré des données de recherche (DFDR) (https://www.frdr-dfdr.ca/repo/?locale=fr), Borealis (https://borealisdata.ca/fr/), dépôts institutionnels ou disciplinaires) |
| Stockage de type archive | • Prend en charge la préservation à long terme; peut ne pas constituer le point d'accès principal à des fins de réutilisabilité, mais offre une fiabilité en ce qui a trait à l'accès et à la réutilisation.<br>• Permet aux associations de bibliothèques régionales d'offrir cette infrastructure aux établissements membres. | • Archivage institutionnel<br>• Stockage employé par les services de bibliothèques universitaires (p. ex., le Ontario Library Research Cloud (OLRC) (https://cloud.scholarsportal.info/) du Ontario Council of University Libraries (OCUL), offert à l'échelle nationale; et WestVault (https://coppul.ca/preservation/westvault/) du Council of Prairie and Pacific University Libraries (COPPUL)) |

**Tableau 2 : Infrastructures de données de recherche au Canada**

| Type | Attributs | Exemples |
|---|---|---|
| Dépôts multidisciplinaires | • Le recours aux dépôts disciplinaires est encouragé lorsqu'ils sont disponibles.<br>• Autrement, il est possible d'utiliser des dépôts institutionnels ou généralistes qui peuvent prendre en charge plusieurs types de fichiers et des cas d'utilisation. | • Consultez le tableau 3 au sujet des dépôts canadiens.<br>• Plateformes internationales et services hébergés (p. ex., Mendeley Data (https://data.mendeley.com/), Figshare (https://figshare.com/), Dryad (https://datadryad.org/stash), Zenodo (https://zenodo.org/), dépôt Dataverse de Harvard (https://dataverse.harvard.edu/)) |

| | | |
|---|---|---|
| Dépôts et infrastructures disciplinaires | • Mettent l'accent sur des types de données en particulier (p. ex., génomique) et peuvent utiliser des normes spécialisées.<br>• Peuvent servir de base de connaissances, offrant la curation, l'extraction, l'organisation, l'annotation et la création de liens vers des corpus littéraires ou de données.<br>• Peuvent servir de portails dédiés à un projet pour recueillir et partager les données de recherche à des fins d'échange, de connaissances et de mobilisation; peuvent comprendre des liens vers des dépôts ou d'autres options de stockage. | • Consultez le tableau 3 au sujet des dépôts canadiens.<br>• Projets de recherche de grande envergure, notamment Linked Infrastructure for Networked Cultural Scholarship (LINCS) (https://lincsproject.ca/), Ocean Networks Canada, (https://www.oceannetworks.ca/) Génome Canada (https://genomecanada.ca/fr/), Guichet de soutien à l'accès aux données (GSAD) (https://www.hdrn.ca/fr/gsad), Linked Parliamentary Data Project (LiPaD) (https://www.lipad.ca/) |
| Outils et services de préservation | • Prennent en charge l'entretien et la préservation à long terme des objets numériques utiles pour la recherche.<br>• Ont recours à des logiciels spécialisés pour préparer les données de recherche en prévision d'une préservation à long terme à l'aide de techniques comme la **normalisation** de fichier, le **contrôle d'intégrité** et la **création de paquets de données**. | • Archivematica — intégration avec des dépôts (p. ex., DFDR, Borealis)<br>• Services de préservation consortiaux (p. ex., Archivematica-as-a-service de COPPUL (https://coppul.ca/preservation/archivematica-as-a-service/), service Permafrost (https://permafrost.scholarsportal.info/) d'OCUL)<br>• Logiciel de préservation national (p. ex., DuraCloud, hébergé pour prendre en charge la préservation numérique pour les abonnés de l'OLRC) |
| Outils et services de reproduction des données de recherche et de logiciels | • Permettent à d'autres de consulter, de manipuler et d'interpréter les données afin d'appuyer la réutilisation et la reproductibilité.<br>• Utilisés afin que d'autres personnes puissent reproduire les données (p. ex., aux fins de collecte, d'analyse, de visualisation). | • Services et plateformes de reproduction de logiciels et de code (p. ex., Code Ocean (https://codeocean.com/), Syzygy (https://syzygy.ca/), Jupyter Hub (https://jupyter.org/hub), GitHub (https://github.com/))<br>• Outils qui facilitent la reproduction du code et des environnements de traitement (p. ex., Jupyter Notebooks (https://jupyter.org/), Docker (https://www.docker.com/)) |
| Services de découverte de données | • Relient les métadonnées et les données à l'aide d'un schéma, d'un format et d'une structure communs afin d'aider les chercheuses et chercheurs à trouver et à réutiliser les données.<br>• Améliorent la découverte dans les dépôts dont les normes et les | • Services de recherche de données de recherche canadiens et internationaux (p. ex., Lunaris (https://www.lunaris.ca/fr), Données ouvertes Canada (https://ouvert.canada.ca/fr/donnees-ouvertes), OpenAIRE (https://www.openaire.eu/), Google Dataset Search (https://datasetsearch.research.google.com/), DataCite |

| | | |
|---|---|---|
| | niveaux d'interopérabilité varient. | Commons (https://commons.datacite.org/), Data Citation Index (https://clarivate.com/webofsciencegroup/solutions/webofscience-data-citation-index/))<br>• Services disciplinaires (p. ex., iReceptor Commons (http://ireceptor.irmacs.sfu.ca/repositories), Global Biodiversity Information Facility (https://www.gbif.org/), Plateforme canadienne de neurosciences ouvertes (https://conp.ca/fr/platform-du-canadian-neuroscience-ouvert/)) |
| Interopérabilité et normes | • Prend en charge un des quatre types d'interopérabilité : technique, sémantique, organisationnelle et juridique (Corcho *et al*., 2021). | • Identifiants uniques pérennes (IUP) (p. ex., Identifiant numérique d'objet (https://www.doi.org/) (DOI) pour les données et les articles, ORCID iD (https://orcid.org/) pour les chercheuses et chercheurs, ROR (https://ror.org/) pour les organisations, RAiD (https://raid.org/) pour les projets de recherche)<br>• Normes de métadonnées (p. ex., Dublin Core (https://www.dublincore.org/), Data Documentation Initiative (https://ddialliance.org/) (DDI), DataCite Schema (https://schema.datacite.org/), Data Catalog Vocabulary (https://www.w3.org/TR/vocab-dcat/))<br>• Ontologies et classement par sujet (p. ex., Répertoire de vedettes-matières (https://rvmweb.bibl.ulaval.ca/rvmweb/accueil.do) (RVM), normes ISO, vocabulaires de W3C)<br>• Licence de données (p. ex., Creative Commons (https://creativecommons.org/), Licence du gouvernement ouvert (https://ouvert.canada.ca/fr/licence-du-gouvernement-ouvert-canada))<br>• Licence de logiciel (p. ex, MIT (https://opensource.org/license/mit/), GNU (https://www.gnu.org/licenses/gpl-3.0.en.html), Apache (https://www.apache.org/licenses/LICENSE-2.0))<br>• Protocole ouvert et normes d'échange (p. ex., OAI-PMH (https://www.openarchives.org/pmh/), SWORD (https://sword.cottagelabs.com/)) |

# Dépôts de données canadiens

Les dépôts de données sont essentiels à l'infrastructure de recherche au Canada. De tels outils nationaux et institutionnels sont mis sur pied pour aider les communautés de recherche à déposer, partager et préserver à long terme leurs données afin d'offrir des services de GDR ouverts, équitables et connectés. Ce faisant, nous évitons les intérêts commerciaux croissants et réduisons la dépendance aux solutions personnalisées comme des sites Web de projet de recherche qui exigent souvent une maintenance et des ressources à long terme. Grâce au financement fédéral, provincial et institutionnel, les dépôts canadiens sont mis à la disposition des chercheuses et chercheurs sans frais supplémentaires et peuvent offrir une plus longue durée de vie que le projet de recherche. Le tableau 3 donne un aperçu des types de dépôts de données au Canada dont plusieurs peuvent être découverts via les registres internationaux comme le Registry of Research Data Repositories (http://re3data.org/) (re3data), FAIRSharing (https://fairsharing.org/) et OpenDOAR (https://v2.sherpa.ac.uk/opendoar/).

**Tableau 3 : Dépôts de données au Canada**

| Type | Attributs | Exemples |
|---|---|---|
| Dépôts multidisciplinaires | • Prennent en charge les données provenant de plusieurs disciplines.<br>• Peuvent offrir des services de curation.<br>• Peuvent agréger des données provenant de différents jeux de données. | • Institutionnels (p. ex., Dataverse de l'Université du Nouveau-Brunswick, Dépôt de données (https://data.upei.ca/) de l'Université de l'Île-du-Prince-Édouard).<br>• Nationaux (p. ex., Borealis (https://borealisdata.ca/), DFDR (https://www.frdr-dfdr.ca/repo/)) |
| Dépôts disciplinaires | • Prennent en charge les données relatives à des disciplines en particulier.<br>• Peuvent offrir des services de curation.<br>• Peuvent agréger des données provenant de différents jeux de données. | • Disciplinaires (p. ex., Polar Data Catalogue (https://www.polardata.ca/), Barcode of Life Data System (http://www.boldsystems.org/), Système intégré d'observation des océans du Canada (https://cioos.ca/fr/accueil/)) |
| Dépôts gouvernementaux | • Conçus pour les données recueillies ou compilées par les ministères.<br>• Axés sur une discipline (c'est-à-dire, ce ne sont pas des sites de données ouvertes génériques). | • BC Data Conservation Centre (https://www2.gov.bc.ca/gov/content/environment/plants-animals-ecosystems/conservation-data-centre)<br>• Centre mondial de données sur l'ozone et le rayonnement ultraviolet (https://woudc.org/home.php?lang=fr)<br>• Archive de données climatiques nationales (https://climat.meteo.gc.ca/in |

| | | |
|---|---|---|
| | | dex_f.html)<br>• Système de gestion des données d'observation de la Terre de RNCan (https://www.eodms-sgdot.nrcan-rncan.gc.ca/index-fr.html) |
| Bases de connaissances | • Extraient, recueillent et font la curation de données d'un domaine d'étude spécifique.<br>• Reposent sur des jeux de données de bases pour relier des corpus d'information. | • Avibase (https://avibase.bsc-eoc.org/avibase.jsp?lang=EN)<br>• DrugBank (https://www.drugbank.ca/)<br>• BioGRID (https://thebiogrid.org/) |
| Dépôts de données universitaires | • Élaborés ou pris en charge par les universités pour héberger des données sous licence et ouvertes.<br>• Peuvent également comprendre des données gouvernementales. | • Services de données de bibliothèque (p. ex., Odesi (https://search.odesi.ca/), Abacus Data Network (https://abacus.library.ubc.ca/), Scholars GeoPortal (http://geo.scholarsportal.info/), Géoindex (https://geoapp.bibl.ulaval.ca/)) |

# Services de soutien

Pour produire des jeux de données au potentiel de réutilisation élevé, les chercheuses et chercheurs doivent adopter de bonnes pratiques de curation alors que les données sont nettoyées, documentées, interreliées, stockées puis partagées. Plusieurs services sont à leur disposition pour élaborer ces pratiques de GDR (voir le tableau 4).

L'évaluation des besoins de l'infrastructure de recherche numérique réalisée en 2021 par l'Alliance de recherche numérique du Canada (l'Alliance) a découvert que les chercheuses et chercheurs ont des niveaux variables d'accès et de sensibilisation au support offert par rapport aux processus de travail de recherche au palier local, provincial et national; l'accès le plus grand se trouve au palier local (Pérez-Jvostov *et al*., 2021).

- Soutiens internes : le premier point de soutien pour plusieurs chercheuses et chercheurs se trouve dans leurs propres groupes de recherche. Par exemple, plusieurs ont recours à des gestionnaires de données pour soutenir les membres de l'équipe avec la gestion et la publication des données. Habituellement, les chercheuses et chercheurs découvrent et sélectionnent les outils et services sur recommandation de leurs pairs (Pérez-Jvostov *et al*., 2021).
- Établissements d'enseignement supérieur : ils offrent des services et un soutien formels par le biais des bureaux de la recherche, des bibliothèques universitaires et des services de calcul informatique (Pérez-Jvostov *et al*., 2021). L'exigence des trois organismes subventionnaires de disposer de stratégies de GDR

aidera à unir le soutien à l'échelle du campus.

- Modèles de soutien partagé : ils peuvent améliorer l'efficacité, l'accès et l'équité tout en répondant aux demandes des chercheuses et chercheurs. Ils sont souvent coordonnés par un consortium régional ou national. L'étude de cas 1 illustre une communauté de pratique formant un réseau de soutien pour les gestionnaires d'un dépôt institutionnel.
- Services et soutiens par discipline : ils répondent aux besoins de communautés de recherche précises et sont souvent promus à l'échelle nationale et internationale par le biais d'organismes de recherche et de maisons d'édition. Ils sont essentiels pour l'adoption de pratiques et d'outils normalisés dans les disciplines connexes, car ils sont adaptés à des processus de travail de recherche particuliers.

**Tableau 4 : Services de soutien au Canada**

| Catégorie | Services |
|---|---|
| Planification de la gestion des données (PGD) | L'Alliance soutient l'infrastructure et supervise l'élaboration de l'Assistant PGD, un outil de gestion des données en ligne.<br>Les bibliothèques et bureaux de la recherche universitaires collaborent pour soutenir les chercheuses et chercheurs locaux à développer des PGD conformément à la politique de GDR des trois organismes. |
| Découverte et accès aux données | Les bibliothèques universitaires soutiennent la découverte et l'accès aux données par le biais de services de référence et d'abonnement à des bases de données. Certains de ces services sont partagés entre les établissements (p. ex., Odesi, dépôt Dataverse Abacus).<br>Des organisations nationales et provinciales permettent l'accès aux données et l'utilisation d'informations démographiques à des fins de recherche. En raison du caractère sensible de ces données, le soutien exige souvent la signature d'un accord avec le fournisseur de service (p. ex., RCCDR et centres de données StatCan, ICES, Population Data BC).<br>L'Alliance soutient un service de découverte national, Lunaris, pour accroître l'exposition aux dépôts de données et aux jeux de données canadiens. Les travaux exploratoires appuient l'accès aux jeux de données partagés sur les infrastructures de calcul de haute performance (p. ex., jeux de données en bio-informatique). |
| Calcul et stockage | Les services de calcul informatique locaux et les TI offrent du soutien aux chercheuses et chercheurs en gestion des données pour le calcul ainsi que de l'infrastructure de stockage pour les données pendant la phase active de la recherche.<br>L'Alliance et sa Fédération nationale de partenaires (https://alliancecan.ca/fr/services/calcul-informatique-de-pointe/la-federation) accordent une importance au fait d'accroître la prise en charge de la gestion des données pour le stockage actif. Les chercheuses et chercheurs peuvent obtenir du soutien par le biais du soutien technique (https://docs.alliancecan.ca/wiki/Technical_support/fr) national de l'Alliance. |
| Curation et publication des données | Une gamme de flux de travail et de guides ont été développés pour aider les personnes responsables de la curation des données, notamment : |

| | |
|---|---|
| | • *Guide pour la curation dans Dataverse* (https://zenodo.org/record/5579827#.YrtH63bMJD8)<br>• Réseau de curation des données (https://datacurationnetwork.org/curator-resources/) : modèles et documents de base (en anglais uniquement)<br>• Guides et listes de vérification de DCC (https://www.dcc.ac.uk/guidance/how-guides) (en anglais uniquement)<br>• *Guide de survie pour la curation des données canadiennes* (bêta) (https://portage-ceg.github.io/fr/)<br><br>Pour appuyer la publication en libre accès, certaines bibliothèques universitaires offrent un service de curation aux chercheuses et chercheurs qui déposent des données et d'autres objets de recherche dans des dépôts institutionnels ou d'autres systèmes de gestion d'actifs numériques.<br>Borealis offre des services locaux par le biais d'un modèle de soutien distribué; l'infrastructure est hébergée de manière centralisée, mais l'aide pour la curation est offerte localement aux chercheuses et chercheurs selon les capacités et l'offre de service des établissements (consultez l'étude de cas 1 ci-dessous).<br>L'Alliance offre du soutien à la curation aux chercheuses et chercheurs qui utilisent le DFDR, un dépôt accessible à l'échelle nationale. Elle aide également les chercheuses et chercheurs à créer et à déployer des portails de recherche sur des infrastructures de calcul informatique de pointe.<br>D'autres dépôts agissent comme des ressources de confiance pour gérer les données de recherche et offrir des services qui soutiennent leurs plateformes (p. ex., Centre canadien des données astronomiques, Ocean Networks Canada, Polar Data Catalogue).<br>Des maisons d'édition commerciales, notamment Springer Nature and Elsevier, offrent des services de soutien à la curation et la publication de jeux de données. D'autres disposent de partenariats avec des dépôts tiers afin d'aider les autrices et auteurs à publier des jeux de données qui soutiennent leurs publications (p. ex., le partenariat entre Wiley et Dryad). |
| Formation | Les chercheuses et chercheurs bénéficient d'une formation auprès de services élaborés au sein des communautés et des établissements dans leur discipline (Pérez-Jvostov *et al.*, 2021). Ces services sont souvent dirigés par des pairs et des spécialistes en soutien qui agissent à titre de « **responsables de l'intendance des données,** » développant des activités pour promouvoir la sensibilisation, la compréhension, le perfectionnement et l'adoption d'outils de GDR, de pratiques exemplaires et de ressources. Les principaux événements au Canada comprennent :<br>• Ateliers et camps d'entraînement estivaux<br>• Cours de type « former le formateur » et ressources<br>• Modules de formation en ligne |
| Domaines de services émergents | Les services de soutien au partage et à la réutilisation des données sont mis sur pied en réaction aux besoins des chercheuses et chercheurs en matière de GDR. Les domaines de services émergents comprennent :<br>• **Préservation numérique** (voir le chapitre 11)<br>• Curation des données sensibles (voir le chapitre 13)<br>• Curation des logiciels de recherche<br>• Souveraineté des données autochtones |

## Étude de cas 1 : mettre sur pied un service et une communauté de dépôt de type Dataverse au Canada

### Contexte

Le projet Dataverse (https://dataverse.org/) est un logiciel ouvert de dépôt de données de recherche qui permet aux personnes utilisatrices de partager, de citer, d'explorer et d'analyser des données de recherche. Il est développé par l'Institute for Quantitative Social Science de l'Université Harvard avec des partenaires de partout dans le monde. Borealis, le dépôt Dataverse canadien (https://borealisdata.ca/fr/), repose sur le logiciel Dataverse et a commencé en tant que dépôt de données de recherche régional pour l'Ontario Council of University Libraries. Au cours des 10 dernières années, il est devenu un service national, bilingue, qui compte plus de 60 établissements membres. L'infrastructure est hébergée par l'Université de Toronto; les fichiers de données sont stockés en toute sécurité sur le Ontario Library Research Cloud (https://cloud.scholarsportal.info/). Borealis offre une option de dépôt aux chercheuses et chercheurs qui ne disposent pas d'un dépôt disciplinaire et qui pourraient bénéficier d'une flexibilité dans les choix de partage des données (p. ex., d'un accès libre à restreint), d'outils d'exploration dans le navigateur ainsi que d'actions et de stockage propices à la préservation.

### Analyse

Bien que Borealis soit hébergé de manière centrale, les bibliothèques et les établissements universitaires gèrent leurs collections; ils appuient ainsi leurs chercheuses et chercheurs dans le dépôt et le partage de jeux de données. Puisque la capacité locale varie selon les établissements et les régions (Goddard *et al.*, 2018), il est essentiel de cultiver une communauté de pratique pour renforcer les capacités de chaque établissement et pour développer de façon collaborative les ressources et le matériel de formation nécessaires pour soutenir les chercheuses et chercheurs. En plus des efforts déployés pour mettre sur pied l'infrastructure technique, l'équipe de Borealis a travaillé avec le  Groupe d'experts Dataverse Nord de l'Alliance sur des initiatives de développement de la communauté telles que la création de ressources bilingues, de documents de sensibilisation et de formation à l'intention des gestionnaires et des personnes utilisatrices, la tenue de rencontres communautaires mensuelles et le maintien d'une liste de diffusion pour partager librement les connaissances, l'expertise et les besoins des chercheuses et chercheurs (Goodchild et Huck, 2022).

### Discussion

Pour que Borealis existe, il est essentiel de créer des espaces et du soutien pour la communauté. La rétroaction aide à définir les priorités en matière d'élaboration technique et de service; la participation de la communauté à la préparation de guides pour l'utilisation et l'administration de Borealis et d'autres projets fait en sorte que les ressources répondent aux besoins de la communauté de recherche. Le but global de la communauté (c'est-à-dire, favoriser le partage et la réutilisation des données de recherche) est en harmonie avec les efforts nationaux pour consolider l'infrastructure de recherche numérique et la communauté de GDR au Canada (GT GDR de l'Alliance, 2020).

## Éléments à prendre en considération pour le partage de données

Le partage de données exige de la planification. Dès le début du projet, dans le cadre du plan de gestion des données, les chercheuses et chercheurs doivent réfléchir aux logiciels et aux outils nécessaires pour créer ou collecter, analyser et documenter les données; au stockage approprié et aux procédures de sauvegarde; à la manière dont les données seront déposées et, si possible, partagées; la manière dont les données seront gérées pour assurer la conformité aux exigences éthiques et légales.

Les différences disciplinaires, notamment l'attitude et la culture, peuvent exercer une influence sur le partage et la réutilisation des données. Certains domaines de recherche disposent de traditions à cet égard et peuvent avoir adopté des normes ainsi que des outils pour soutenir ce travail. C'est particulièrement le cas des sciences humaines où les résultats ne correspondent pas toujours aux définitions traditionnelles des données de recherche; les chercheuses et chercheurs peuvent alors penser à utiliser des approches différentes pour favoriser le partage. Des services et des outils sont souvent mis sur pied afin de répondre à des besoins particuliers à une discipline et il peut s'avérer difficile de les adopter ou de les réorienter dans d'autres disciplines ou contextes. Bien que les outils et services généraux peuvent être utiles, ils n'ont souvent pas le contexte disciplinaire nécessaire pour permettre leur réutilisation et leur adoption. Parmi les éléments disciplinaires à prendre en considération, mentionnons :

- Les formats de fichier (libre vs propriétaire, outils et logiciels standards au sein de la discipline);
- Les normes de métadonnées utilisées pour la  documentation et la découverte de jeux de données;

- Le stockage de données actives, les outils de transfert de données et les dépôts de données pour soutenir les besoins disciplinaires (p. ex., données massives, données sensibles);
- Le choix de dépôt en fonction des caractéristiques et de la communauté qui va l'utiliser;
- La disponibilité de la curation des données :
  - Examen de la qualité des données;
  - Documentation des données aux fins de réutilisation;
  - Transformation des données (p. ex., nettoyage, anonymisation, dépersonnalisation);
- Les modalités d'accès et les licences de réutilisation;
- Les outils d'exploration et de visualisation des données;
- Les avantages de partager divers types de données.

Les études de cas suivantes examinent des projets de recherche ou des considérations d'ordre disciplinaire dans les domaines des **humanités numériques** (étude de cas 2), des sciences de la santé (étude de cas 3) et des sciences naturelles (étude de cas 4). Elles mettent en évidence les enjeux auxquels sont confrontés les chercheuses et chercheurs et mettent de l'avant des solutions ainsi que les leçons apprises.

## Étude de cas 2 : humanités numériques

### Contexte

La bibliothèque de l'Université Queen's a organisé l'exposition virtuelle de la collection Diniacopoulos (https://virtual-exhibits.library.queensu.ca/diniacopoulos-collection/) (en anglais uniquement), le point culminant d'un projet de recherche qui présente des films en réalité virtuelle et des modèles 3D à l'échelle d'artéfacts archéologiques grecs et égyptiens de la collection de la faculté d'études classiques. L'exposition virtuelle a été construite sur WordPress et utilise le logiciel Object2VR pour créer une expérience interactive qui permet d'examiner et de faire tourner les objets en réalité virtuelle 3D dans le navigateur.

### Analyse

L'équipe de recherche voulait partager et préserver les données du projet pour une utilisation future, car le domaine de la réalité virtuelle ne cesse d'évoluer. Les visionneuses en ligne et les systèmes de gestion de contenu exigent un entretien continu de logiciels et d'outils dont la durée de vie est inconnue, ce qui met en lumière des éléments à prendre en considération au chapitre de la durabilité et de l'accès à long terme. Parmi les défis rencontrés, mentionnons le

choix du dépôt, étant donné la taille du jeu de données (60 Go), l'important nombre de fichiers (plus de 6500) et la complexité de la structure de dossiers, sans compter que ce domaine dispose de peu d'options et de pratiques exemplaires. Qui plus est, il était essentiel d'inclure la documentation et les métadonnées disciplinaires pour faire en sorte que les données puissent être réutilisées et comprises hors de leur contexte d'origine.

### Discussion

L'équipe de recherche a déposé le jeu de données dans la collection Dataverse de Queen's (https://doi.org/10.5683/SP/T7ZJAF) (Jones *et al.*, 2017), qui fait partie de Borealis, afin de bénéficier du soutien de la bibliothèque de l'Université Queen's et de caractéristiques comme des champs de métadonnées exhaustifs et la capacité d'attribuer un **identificateur d'objets numériques (DOI)** qui pourrait être lié à l'exposition virtuelle. L'équipe de Borealis a pris en charge le dépôt de gros dossiers d'archives compressés de type ZIP pour chaque artéfact. Le débat se poursuit au sujet de la compréhension des données de recherche en sciences humaines. Il faut continuer à étudier la question par le biais de statistiques d'utilisation et de citations des jeux de données pour déterminer s'il existe des défis à la réutilisation de ces données contextuelles et si des outils et des plateformes améliorés pourraient mieux gérer, partager et conserver ces types de projets en humanité numériques.

## Étude de cas 3 : partage de données sensibles

### Contexte

Les données sensibles font référence aux données qui peuvent causer préjudice si rendues publiques. Habituellement, il s'agit de données recueillies à propos d'êtres humains et peuvent comprendre de l'information sensible, confidentielle ou personnelle en lien, entre autres, avec la santé, l'ethnicité, les opinions politiques ou l'emplacement géographique d'une personne. Les données de recherche qui impliquent des êtres humains doivent être gérées conformément aux lignes directrices du comité d'éthique de la recherche (CÉR) et en recevoir l'approbation. Plusieurs établissements fournissent des normes de sécurité et des lignes directrices en matière de protection pour gérer les données sensibles et confidentielles.

Au Canada, la recherche financée par les trois organismes fédéraux de financement de la recherche (**les organismes subventionnaires**) qui implique des êtres humains est encadrée par l'*Énoncé de politique des trois conseils : Éthique de la recherche avec des êtres humains (https://ethics.gc.ca/fra/policy-politique_tcps2-eptc2_2022.html)*(EPTC 2) (https://ethics.gc.ca/fra/policy-politique_tcps2-eptc2_2022.html) (Groupe en éthique de la recherche, 2022). Les chercheuses et chercheurs doivent se conformer à la politique, laquelle aborde les enjeux de consentement, de la vie privée et de l'équité en lien avec divers types de recherche humaine, notamment les essais cliniques, la recherche génétique et celle impliquant les Premières Nations, les Inuits et les Métis. La recherche portant sur les peuples autochtones peut ne pas être sujette aux lignes directrices de l'**EPTC 2**, selon les circonstances et les modalités convenues ou qui régissent les données considérées sous le contrôle des personnes participantes ou des groupes communautaires (consultez le chapitre 3, « Souveraineté des données autochtones » ; consultez les principes de PCAP (https://fnigc.ca/fr/les-principes-de-pcap-des-premieres-nations/)® pour un modèle de gestion des données au sujet des Premières Nations). La manipulation et l'utilisation de données sensibles peuvent être régies par d'autres cadres légaux et éthiques du programme de recherche (p. ex., IRSC, CRSH) ou de l'établissement, ou au palier provincial (p. ex., *Loi sur l'accès à l'information et la protection de la vie privée (https://www.ontario.ca/fr/lois/loi/90f31)*) ou fédéral (p. ex., *Loi sur la protection des renseignements personnels et les documents électroniques (https://lop.parl.ca/sites/PublicWebsite/default/fr_CA/ResearchPublications/200744E#:~:text=les%20documents%20%C3%A9lectroniques-,La%20Loi%20sur%20la%20protection%20des%20renseignements%20personnels%20et%20les,cadre%20d'activit%C3%A9s%20commerciales%2024)*).

En 2021, les trois conseils ont émis des lignes directrices à l'intention des chercheuses et chercheurs intitulées *Lignes directrices pour verser des données existantes dans des dépôts publics (https://ethics.gc.ca/fra/depositing_depots.html)* (Groupe en éthique de la recherche, s.d.). Le document indique que les chercheuses et chercheurs peuvent déposer et partager des données dans un dépôt si les personnes participantes ont consenti à cet effet ou si un CÉR a donné son approbation. Les chercheuses et chercheurs doivent être conformes à l'EPTC 2 avant le dépôt et le partage des données et obtenir l'approbation du CÉR avant de faire la collecte ou la réutilisation de la recherche qui implique des êtres humains.

## Analyse

L'infrastructure et les services de soutien pour le stockage, le dépôt et le partage de données sensibles demeurent une lacune importante au Canada. La complexité entourant les données

sensibles exige un croisement entre plusieurs services et unités administratives d'un établissement, notamment les lignes directrices du CÉR, les contrats et services juridiques, les pratiques en matière de GDR ainsi que l'infrastructure et les processus de travail pour gérer les données sensibles tout au long de leur cycle de vie.

Dans le cadre de la recherche en sciences de la santé, plusieurs options sont offertes pour publier ou partager des données; les éléments à prendre en considération varient. La dépersonnalisation ou l'anonymisation des jeux de données comprend la suppression de données identifiables d'un jeu de données. Toutefois, certains d'entre eux ne peuvent pas être dépersonnalisés sans compromettre l'utilité des données. Ils peuvent être partagés par des portails à accès restreint grâce à des ententes de partage/transfert de données. Cette approche présente certains inconvénients:  les frais administratifs généraux et le besoin potentiel d'avoir un portail fait sur mesure.

## Discussion/conclusions

Des efforts sont continuellement déployés pour améliorer les outils, l'infrastructure, les processus de travail et les ressources en ce qui a trait à la gestion et au partage de données sensibles. Des logiciels sécuritaires et faciles d'utilisation, comme Research Electronic Data Capture (REDCap), sont de plus en plus populaires en tant qu'outils pour la saisie de données en recherche clinique et pour la création de bases de données et de projets conformes aux lignes directrices légales (Patridge et Bardyn, 2018). Le projet de dépôt de données sensibles de l'Alliance a mené à la création d'un outil de cryptage à divulgation nulle de connaissance pour faciliter le dépôt sécuritaire et l'accès contrôlé aux données sensibles au sein de la plateforme DFDR. Pour la prochaine phase du projet, l'équipe de GDR de l'Alliance dirige la participation collaborative entre établissements afin d'élaborer un cadre politique ayant pour but de préciser et de simplifier le flux de travail pour le dépôt et le partage de données sensibles. Le Groupe d'experts en données sensibles de l'Alliance a publié des documents visant à encadrer les pratiques de GDR dans le contexte de l'éthique de la recherche, notamment la boîte à outils pour les données sensibles.

- Partie 1 : *Glossaire terminologique sur l'utilisation des données sensibles à des fins de recherche (https://zenodo.org/record/4088986)*
- Partie 2 : *Matrice de risque lié aux données de recherche avec des êtres humains (https://zenodo.org/record/4107119)*
- Partie 3 : *Langage en matière de gestion des données de recherche pour le consentement éclairé (https://zenodo.org/record/4107186)*

Les chercheuses et chercheurs ont besoin d'un leadership permanent pour trouver des solutions nationales afin de garantir un accès équitable au soutien, aux outils et à l'infrastructure pour la gestion et le partage des données sensibles.

## Étude de cas 4 : soutenir les grands producteurs de données au Canada – SuperDARN et le Dépôt fédéré de données de recherche (DFDR)

### Contexte

Le Super Dual Auroral Radar Network (SuperDARN) est un réseau composé de 36 radars scientifiques déployés partout dans le monde par des universités et des laboratoires gouvernementaux de 10 pays. SuperDARN Canada (dont le siège social se trouve à l'Université de la Saskatchewan) exploite cinq radars au Canada, lesquels produisent des données précieuses que les chercheuses et chercheurs peuvent utiliser pour comprendre la météorologie de l'espace, la radiocommunication et la physique dans la haute atmosphère terrestre. Toutefois, en raison des saisies de qualité supérieure et des taux de collecte rapides des radars, SuperDARN génère des données à très grande échelle; leur stockage de manière sécuritaire, consultable et accessible constitue un défi. En 2018, SuperDARN Canada a commencé à rencontrer l'équipe du DFDR.

### Analyse

La taille, l'échelle et la portée des données, en plus de la complexité du cadre organisationnel de SuperDARN en tant que partenaire de recherche international, présentaient de nombreux défis. La collecte de données de SuperDARN a commencé en 1993; elles existent sous forme brute et traitée. SuperDARN Canada et le DFDR ont réfléchi au format de données qu'il conviendrait le mieux de publier (environ 80 To de données brutes ou environ 10 To de données traitées par version algorithmique) et, parmi les données traitées, quelle génération d'algorithme choisir : l'algorithme le plus ancien, largement utilisé, ou le plus récent. La création de **versions** des jeux de données pour mettre à jour l'algorithme obsolète signifiait doubler la taille de la collection.

Les données sont collectées au fil du temps, des régions et des instruments par des installations de radars qui fonctionnent dans les deux hémisphères. Par conséquent, les équipes devaient prendre en considération la manière de subdiviser les données en unités publiables les mieux adaptées à la découverte, à la réutilisation, au suivi de l'utilisation et à la création de rapports. Les équipes devaient également réfléchir à la taille des jeux de données et au nombre de fichiers, sans oublier les limites relatives au navigateur Web. Bien que les fichiers soient petits, les jeux de données pouvaient atteindre plusieurs téraoctets en fonction de la manière dont les données étaient organisées.

Puisque les données brutes et traitées étaient offertes uniquement sous forme de fichiers binaires, l'équipe de curation du DFDR ne pouvait pas réaliser de vérification de la qualité. La complexité des données signifiait aussi que sans documentation exhaustive, les jeux de données ne seraient utiles qu'à un nombre restreint de personnes qui participent à la recherche.

## Discussion/conclusions

### Format

L'équipe a décidé de publier les données sous forme brute depuis 1993.

### Curation

L'équipe de curation du DFDR a collaboré avec SuperDARN Canada pour examiner les jeux de données et préparer des fichiers **LISEZ-MOI** qui saisissent les métadonnées descriptives et techniques pour que la communauté élargie de chercheuses et chercheurs puisse les utiliser. Des liens vers les publications et la documentation connexes ont été ajoutés et les jeux de données ont été reliés à un logiciel d'analyse et de visualisation créé par SuperDARN.

### Leçons tirées

En plus des solutions abordées précédemment, ce projet a permis de tirer les leçons suivantes:

- La consultation sur les besoins en matière de publication des données peut prendre du temps et le processus est continu. Il s'est écoulé plusieurs années entre la première conversation et l'intégration des premiers jeux de données. Après la publication, le DFDR et SuperDARN Canada continuent de se rencontrer régulièrement.
- Il est important d'avoir une communication cohérente, surtout lorsque les décisions exigent des échéances plus longues. Il faut organiser des rencontres régulières,

documenter les discussions et les décisions pour faire en sorte que les parties prenantes demeurent sur la même longueur d'onde et que les fils de discussion ne soient pas perdus.
- La durabilité et la planification sont essentielles. Dans le cadre de sa collaboration avec SuperDARN, le DFDR devait réfléchir aux besoins en matière de publication des données en lien avec la collecte ainsi que son engagement pour l'avenir.

# L'avenir du partage de données au Canada

Plusieurs développements pourraient mieux soutenir les chercheuses et chercheurs du Canada pour tirer pleinement profit des avantages du partage de données. Quelques possibilités sont suggérées ci-après, notamment l'amélioration de l'accès et de l'inclusion, le renforcement des plateformes de recherche qui prennent en charge le cycle de vie des données, l'élaboration d'outils et de technologies pour automatiser les processus de travail de curation et l'amélioration de l'**intégration** et de l'interopérabilité entre les systèmes et les plateformes.

## Accès et inclusion

Les obstacles systémiques à l'inclusion de l'ensemble des chercheuses et chercheurs de toutes les disciplines pour l'accès et l'utilisation des outils et des services de partage des données doivent être supprimés. Ceci permettrait de favoriser une adoption plus équitable des politiques et pratiques de partage de données. De nouvelles façons de concevoir le partage des données sont nécessaires pour transformer les infrastructures qui prennent en charge tous les types de données de recherche, à la fois en matière de formats et de normes, mais aussi en ce qui a trait aux modèles et processus de travail encore théoriques.

Au fur et à mesure que les processus de travail de partage de données évoluent, il faut veiller à créer des modèles d'édition équitables. Étant donné le coût élevé du stockage, particulièrement pour les gros jeux de données, nous devons équilibrer durabilité et équité.

### Exemples

- Davantage d'options de personnalisation des dépôts de données; des outils et des normes flexibles;

- Des normes d'accessibilité Web dans les logiciels et plateformes;
- Des ententes d'accès libre entre établissements de recherche, maisons d'édition et dépôts.

## Plateformes de cycle de vie de la recherche

Les processus de travail habituels pour téléverser ou télécharger des données d'un dépôt exigent le transfert de données entre les plateformes et entre les emplacements de stockage. Cette façon de procéder est inefficace et dispendieuse, voire impossible pour les gros ensembles de données en raison du coût, du temps nécessaire pour le transfert ou des limites de l'infrastructure. De plus, certains jeux de données dépendent de logiciels ou d'environnements informatiques spécialisés pour réaliser des analyses. Les plateformes de recherche et les grappes de stockage qui prennent en charge le cycle de vie complet des données, où il serait possible d'analyser les données, d'en faire la curation et où une version sûre serait partagée, sont nécessaires.

### Exemples

- Des outils faciles à utiliser pour redistribuer les jeux de données entre plusieurs couches de stockage diverses (p. ex., déplacer des données depuis et vers un dépôt et un stockage actif);
- Des plateformes infonuagiques complètes permettant l'analyse, la curation et le partage de données.

## Automatisation de la curation

Pour faire progresser la science ouverte, il ne suffit pas de rendre les données accessibles. Il faut temps et argent pour que les jeux de données soient conformes aux principes FAIR. Les nouveaux outils et les nouvelles technologies pourraient réduire cet investissement et soutenir les chercheuses, les chercheurs ainsi que les personnes responsables de la curation à produire des résultats de recherche de qualité supérieure.

### Exemples

- Des algorithmes d'intelligence artificielle qui génèrent des métadonnées de qualité supérieure à partir des données;
- Des logiciels pour le couplage automatisé de données, à l'intérieur des jeux de données et entre eux;
- Des logiciels qui guident les chercheuses et chercheurs dans la documentation de leurs jeux de données, avec des normes et des taxonomies intégrées;
- Des logiciels qui vérifient la reproductibilité et la qualité des jeux de données.

## Intégration et interopérabilité

Comme l'illustre la gamme de politiques, d'outils et de services qui soutiennent le partage des données de recherche, l'impulsion est grande pour faire progresser ces infrastructures. Toutefois, plusieurs sont offerts et développés en silos, reliés par trop peu d'éléments de logiciel médiateur ou de politiques-cadres. Alors que ces infrastructures sont mises sur pied, l'interopérabilité (p. ex., relier la politique à la plateforme, la plateforme au service, le service à la politique) et l'intégration aux processus de travail de recherche et d'édition se trouveront au cœur des activités visant à améliorer la facilité d'utilisation et l'adoption accrue de pratique de partage de données.

### Exemples

- Des cadres politiques pour le partage de données au-delà des limites des juridictions;
- L'intégration des plans de gestion des données à l'infrastructure de recherche et de partage;
- La connexion des jeux de données dans un réseau plus vaste de résultats de recherche.

# Conclusion

L'infrastructure, les outils et services canadiens qui soutiennent le partage de données de recherche sont importants, surtout à la lumière des politiques qui exigent un accès aux données financées par des fonds publics. Le domaine d'étude d'une chercheuse ou d'un chercheur et les préoccupations éthiques ont un impact sur la manière dont les données sont partagées et influencent l'élaboration de politiques et d'infrastructures qui pourraient faire progresser le partage de données au Canada.

**Questions de réflexion**

1. Quels sont les défis en matière de partage de données de recherche?
2. Quels sont les types de stockage de données? Donnez un exemple pour chacun d'eux.
3. Que faut-il prendre en considération en matière de partage des données? Quel rôle jouent les différences relatives à la discipline à cet égard?
4. Quels types de services de données (local, spécifique à un domaine ou national) pourraient

être mis sur pied pour aborder les défis et obstacles mentionnés dans ce chapitre?

**Éléments clés à retenir**

- Les organismes de financement et les maisons d'édition peuvent définir des exigences qui favorisent le partage de données de recherche; toutefois, les politiques à elles seules ne suffisent pas à créer des résultats reproductibles. Des solutions techniques et particulières à la discipline sont nécessaires pour rendre les données accessibles et réutilisables.
- Les options de stockage, les infrastructures et les dépôts de données au Canada soutiennent la production, le partage et la réutilisation des données de recherche tout au long de leur cycle de vie. Le stockage de données de recherche peut être divisé en trois types : actif, de dépôt et archivistique. Les établissements de recherche canadiens offrent souvent des infrastructures de stockage à leurs chercheuses et chercheurs, bien que la disponibilité varie selon la capacité de l'établissement.
- Des services de soutien existent pour les chercheuses et chercheurs du Canada qui élaborent des pratiques de GDR, qui publient des données ou qui planifient la réutilisation de données, y compris des services provenant de leurs propres groupes de recherche ou établissements d'enseignement supérieur et des services uniques pour répondre aux besoins de communautés de recherche en particulier.
- Les chercheuses et chercheurs devraient prendre en considération les différences disciplinaires et le contexte relatif au partage des données. Traditionnellement, certains domaines sont ouverts au partage et à la réutilisation des données. Si certaines disciplines ont adopté des normes et des outils pour soutenir ce travail, d'autres peuvent en avoir besoin pour aborder des sujets comme les métadonnées, la taille des fichiers, le type de fichier et les exigences relatives aux données sensibles.
- Le partage et la réutilisation des données sont soutenus par l'intégration et l'interopérabilité des systèmes et des plateformes, notamment celles qui prennent en charge le cycle de vie et les technologies qui facilitent les processus de travail liés à la curation des données.

# Lectures et ressources supplémentaires

Barsky, E., Laliberté L. W., Leahey, A. et Trimble, L. (2017). Chapter 3. Collaborative Research Data Curation Services: A View from Canada. Dans L. R. Johnston (dir.), *Curating research data, volume one: Practical strategies for your digital repository* (p. 79-101). Association of College and Research Libraries. https://dx.doi.org/10.14288/1.0340778 (https://dx.doi.org/10.14288/1.0340778)

Cheung, M., Cooper, A., Dearborn, D., Hill, E., Johnson, E., Mitchell, M. et Thompson, K. (2022). Practices before policy: Research data management behaviours in Canada. *Partnership: The Canadian Journal of Library and Information Practice and Research, 17*(1), 1–80. https://doi.org/10.21083/partnership.v17i1.6779

First Nations Information Governance Centre. (2014, 23 mai). *Ownership, control, access and possession (OCAP™): The path to First Nations information governance*. https://achh.ca/wp-content/uploads/2018/07/OCAP_FNIGC.pdf (https://achh.ca/wp-content/uploads/2018/07/OCAP_FNIGC.pdf)

Garnett, A., Leahey, A., Savard, D., Towell. B. et Wilson, L. (2017). Open metadata for research data discovery in Canada. *Journal of Library Metadata, 17*(3-4), 201-217. https://doi.org/10.1080/19386389.2018.1443698 (https://doi.org/10.1080/19386389.2018.1443698)

Thompson, K. et Kellam, L. M. (2016). Introduction to databrarianship: The academic data librarian in theory and practice. Dans L. M. Kellam et K. Thompson (dir.), *Databrarianship: The academic data librarian in theory and practice*. Association of College and Research Libraries. https://scholar.uwindsor.ca/cgi/viewcontent.cgi?article=1047&context=leddylibrarypub (https://scholar.uwindsor.ca/cgi/viewcontent.cgi?article=1047&context=leddylibrarypub)

Rice, R. et Southall, J. (2016). *The data librarian's handbook*. Facet Publishing.

# Bibliographie

Baker, D., Bourne-Tyson, D., Gerlitz, L., Haigh, S., Khair, S., Leggott, M., Moon, J., Ridsdale, C., Tourangeau, R. et Whitehead, M. (2019). *Research data management in Canada: A backgrounder*. Zenodo. https://doi.org/10.5281/zenodo.3574685 (https://doi.org/10.5281/zenodo.3574685)

Conseil de recherche en sciences humaines. (s.d.). *Politique sur l'archivage des données de recherche*. Gouvernement du Canada. https://www.sshrc-crsh.gc.ca/about-au_sujet/policies-politiques/statements-enonces/edata-donnees_electroniques-fra.aspx (https://www.sshrc-crsh.gc.ca/about-au_sujet/policies-politiques/statements-enonces/edata-donnees_electroniques-fra.aspx)

Corcho, O., Eriksson, M., Kurowski, K, Ojsteršek, M., Choirat, C. van de Sanden, M. et Coppens, F. (2021). *EOSC interoperability framework: Report from the EOSC executive board working groups FAIR and architecture*. Publications Office of the European Union. https://data.europa.eu/doi/10.2777/620649 (https://data.europa.eu/doi/10.2777/620649)

Goddard, L., Barsky, E., Cooper, A., Darnell, A., Davis, C., Doiron, J. et Taylor, S. (2018). *Dataverse north working group: Year 1 recommendations*. UBC Faculty Research and Publications. https://doi.org/10.14288/1.0386773 (https://doi.org/10.14288/1.0386773)

Goodchild, M. et Huck, J. (2022, 29 mars). *Building a shared open research data repository community in Canada*. Open Science Framework. https://osf.io/b9vyt (https://osf.io/b9vyt)

Gouvernement du Canada. (2021). *Politique des trois organismes sur la gestion des données de recherche*. https://science.gc.ca/site/science/fr/financement-interorganismes-recherche/politiques-lignes-directrices/gestion-donnees-recherche/politique-trois-organismes-gestion-donnees-recherche (https://science.gc.ca/site/science/fr/financement-interorganismes-recherche/politiques-lignes-directrices/gestion-donnees-recherche/politique-trois-organismes-gestion-donnees-recherche)

Gouvernement du Canada. (s.d.). *Stratégies institutionnelles de gestion des données de recherche publiées*. https://science.gc.ca/site/science/fr/financement-interorganismes-recherche/politiques-lignes-directrices/gestion-donnees-recherche/strategies-institutionnelles-gestion-donnees-recherche-publiees (https://science.gc.ca/site/science/fr/financement-interorganismes-recherche/politiques-lignes-directrices/gestion-donnees-recherche/strategies-institutionnelles-gestion-donnees-recherche-publiees)

Groupe de travail sur la gestion des données de recherche de l'Alliance. (2020). *État actuel de la gestion des données de recherche au Canada*. Zenodo. https://zenodo.org/record/6647045 (https://zenodo.org/record/6647045)

Groupe en éthique de la recherche. (s.d.). *Lignes Directrices pour verser des données existantes dans des dépôts publics*. Gouvernement du Canada. https://ethics.gc.ca/fra/depositing_depots.html (https://ethics.gc.ca/fra/depositing_depots.html)

Groupe en éthique de la recherche. (2022). *Énoncé de politique des trois conseils : Éthique de la recherche avec des êtres humains – EPTC 2 (2022)*. Gouvernement du Canada. https://ethics.gc.ca/fra/policy-politique_tcps2-eptc2_2022.html (https://ethics.gc.ca/fra/policy-politique_tcps2-eptc2_2022.html)

Jacoby, W. G., Lafferty-Hess, S. et Christian, T-M. (2017). *Should journals be responsible for reproducibility?* Inside Higher Ed Blog. https://www.insidehighered.com/blogs/rethinking-research/should-journals-be-responsible-reproducibility (https://www.insidehighered.com/blogs/rethinking-research/should-journals-be-responsible-reproducibility)

Jones, K., Bevan, G. et Monette, M. (2017). *The Diniacopoulos ceramics display, Department of Classics – 2016* [Jeu de données].Borealis. https://doi.org/10.5683/SP/T7ZJAF (https://doi.org/10.5683/SP/T7ZJAF)

Lin, D., Crabtree, J., Dillo, I., Downs, R. R., Edmunds, R., Giaretta, D., De Giusti, M., L'Hours, H., Hugo, W., Jenkyns, R., Khodiyar, V., Martone, M. E., Mokrane, M., Navale, V., Petters, J., Sierman, B., Sokolova, D. V., Stockhause, M. et Westbrook, J. (2020) The TRUST Principles for digital repositories. *Sci Data, 7*, 144. https://doi.org/10.1038/s41597-020-0486-7 (https://doi.org/10.1038/s41597-020-0486-7)

Patridge, E. F. et Bardyn, T. P. (2018). Research electronic data capture (REDCap). *JMLA, 106*(1), 142–144. https://doi.org/10.5195/jmla.2018.319 (https://doi.org/10.5195/jmla.2018.319)

Pérez-Jvostov, F., Iron, K., Khair, S., Sahrakorpi, S. et Zhang, Q. (2021). *Évaluation des besoins de la communauté de recherche: résumé des commentaires reçus*. Alliance de recherche numérique du Canada. https://alliancecan.ca/sites/default/files/2022-04/EvaluationBesoins_Alliance_20220126.pdf (https://alliancecan.ca/sites/default/files/2022-04/EvaluationBesoins_Alliance_20220126.pdf)

Public Library of Science. (2022, 29 mars). *PLOS launches new feature to promote data sharing and access*. The Official PLOS Blog. https://theplosblog.plos.org/2022/03/plos-launches-new-feature-to-promote-data-sharing-and-access/ (https://theplosblog.plos.org/2022/03/plos-launches-new-feature-to-promote-data-sharing-and-access/)

Rieseberg, L., Warschefsky, E., O'Boyle, B., Taberlet, P., Ortiz-Barrientos, D., Kane, N. C. et Sibbett, B. (2021). Editorial 2021. *Molecular Ecology, 30*(1), 1-25. https://doi.org/10.1111/mec.15759 (https://doi.org/10.1111/mec.15759)

Stuart, D., Baynes, G., Hrynaszkiewicz, I., Allin, K., Penny, D., Lucraft, M. et Astell, M. (2018). *Whitepaper: Practical challenges for researchers in data sharing*. Figshare. https://doi.org/10.6084/m9.figshare.5975011 (https://doi.org/10.6084/m9.figshare.5975011.v1)

Tedersoo, L., Küngas, R., Oras, E., Köster, K., Eenmaa, H., Leijen, Ä., Pedaste, M., Raju, M., Astapova, A., Lukner, H., Kogermann, K. et Sepp, T. (2021). Data sharing practices and data availability upon request differ across scientific disciplines. *Scientific data, 8*, 192. https://doi.org/10.1038/s41597-021-00981-0 (https://doi.org/10.1038/s41597-021-00981-0)

Vines, T. H., Andrew, R. L., Bock, D. G., Franklin, M. T., Gilbert, K. J., Kane, N. C., Moore, J-S., Moyers, B. T., Renaut, S., Rennison, D. J., Veen, T. et Yeaman, S. (2013), Mandated data archiving greatly improves access to research data. *The FASEB Journal, 27*(4), 1304-1308. https://doi.org/10.1096/fj.12-218164 (https://doi.org/10.1096/fj.12-218164)

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N.,

Boiten, J-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., ... Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific data, 3*, 160018. https://doi.org/10.1038/sdata.2016.18 (https://doi.org/10.1038/sdata.2016.18)

## À propos des auteurs

### Meghan Goodchild

Meghan Goodchild est la bibliothécaire responsable de la gestion des données de recherche à l'Université Queen's et à Scholars Portal, un service du Ontario Council of University Libraries. À la bibliothèque de l'Université Queen's, Meghan est la principale personne-ressource pour la gestion des données de recherche et collabore avec les partenaires du campus afin d'améliorer les processus de travail et les services à l'appui au cycle de vie des données de recherche. À Scholars Portal, Meghan dirige l'équipe qui soutient Borealis, le dépôt Dataverse canadien. Elle est titulaire d'un doctorat en théorie de la musique et d'une maîtrise en sciences de l'information de l'Université McGill.

### Shahira Khair

Shahira Khair (elle/elle) est bibliothécaire aux bibliothèques de l'Université de Victoria (UVic), responsable de l'analyse organisationnelle et de la gestion des données. Avant de rejoindre l'UVic, elle a travaillé avec des organisations nationales qui font progresser les initiatives numériques dans le domaine de la recherche et de l'enseignement supérieur, notamment l'Association des bibliothèques de recherche du Canada et l'Alliance de recherche numérique du Canada. Elle est titulaire d'une maîtrise en biologie et d'une maîtrise en sciences de l'information de l'Université d'Ottawa.

### Amber Leahey

Amber Leahey est bibliothécaire de données et des systèmes d'information géographique (SIG) ainsi que directrice des services pour Borealis, le dépôt Dataverse canadien, un dépôt de données national sécurisé et bilingue fourni en partenariat avec les bibliothèques universitaires et les établissements de recherche à travers le Canada. Dans son rôle, elle soutient les bibliothèques, les établissements et les chercheuses et chercheurs dans la gestion, le partage, la préservation et la réutilisation des données grâce au développement continu des services de soutien en lien avec les données et la recherche à Scholars Portal et aux bibliothèques de l'Université de Toronto. Elle est titulaire d'une maîtrise en bibliothéconomie et en sciences de l'information de l'Université de Toronto.

## Newson Kaitlin

Kaitlin Newson est consultante en recherche numérique auprès d'ACENET à l'Université de l'Île-du-Prince-Édouard. Auparavant, Kaitlin était bibliothécaire de projets numériques au sein de Scholars Portal, un service du Ontario Council of University Libraries, où elle soutenait l'infrastructure numérique pour la gestion des données de recherche, l'édition savante et les services de stockage infonuagique pour les bibliothèques universitaires canadiennes. Elle est titulaire d'une maîtrise en information de l'Université de Toronto.

## Lee Wilson

Lee Wilson est directeur de la gestion des données de recherche (GDR) à l'Alliance de recherche numérique du Canada (l'Alliance). À ce titre, Lee supervise l'équipe nationale de gestion des données de recherche de l'Alliance ainsi que la fourniture et le développement de services dans le cadre de partenariats avec divers établissements et organisations canadiennes. Auparavant, Lee a occupé le poste de gestionnaire des plateformes et services de GDR à l'Alliance, a travaillé comme consultant en recherche pour la gestion des données au Canada atlantique avec ACENET et a fait partie de l'équipe de gestion des données pour le Marine Environmental Observation Prediction and Response Network, soutenant les chercheuses et chercheurs qui travaillaient avec des données océaniques. Il est titulaire d'une maîtrise en bibliothéconomie et en sciences de l'information de l'Université de Dalhousie.

6.

# LE MODÈLE D'ÉVALUATION DE LA MATURITÉ DE LA GDR AU CANADA (MEMAC)

Jane Fry; Jennifer Abel; Dylanne Dearborn; Alison Farrell; et Chantal Ripp

**Objectifs d'apprentissage**

À la fin de ce chapitre, vous pourrez :

1. Expliquer ce qu'est un modèle d'évaluation de la maturité.
2. Comprendre l'importance d'une évaluation de la maturité pour la gestion des données de recherche.
3. Comprendre le pourquoi et le comment du développement d'un modèle proprement canadien d'évaluation de la maturité.
4. Utiliser le modèle d'évaluation de la maturité au Canada pour évaluer la maturité du service de gestion des données de recherche d'un établissement canadien de recherche.
5. Encourager la prise de décision fondée sur les preuves grâce aux résultats tirés du modèle d'évaluation de la maturité au Canada

## Introduction

Vous savez maintenant que la **gestion des données de recherche** (GDR) comprend tout un ensemble de pratiques et de services, tels que la planification de la gestion des données, la curation, la découverte et la préservation. Les établissements de recherches qui réfléchissent à la GDR – tant les universités, les collèges que les centres hospitaliers – devraient examiner tous les services, ressources et effectifs qui soutiennent la GDR pour chacun des projets de recherche, particulièrement lors de l'officialisation de leurs services. Cette

démarche a été entreprise par plusieurs établissements canadiens au moment de la rédaction de ce chapitre (printemps 2022) en réponse à la **Politique des trois organismes sur la gestion des données de recherche**.

Mais comment les établissements de recherche peuvent-ils déterminer si toutes les étapes du **cycle de vie des données de recherche** sont prises en charge et ensuite, qui est responsable de ces étapes? Afin de soutenir les établissements canadiens de recherche dans leurs réflexions, les autrices de ce chapitre se sont réunies au cours de l'été 2021 pour développer le *Modèle d'évaluation de la maturité de la GDR au Canada (https://zenodo.org/record/5745894#.Y-vbphOZPaq)*, ou **MEMAC** (Fry *et al*., 2021). Il s'agit d'une initiative pour aider les partenaires de GDR à mieux comprendre les services et ressources qui soutiennent la gestion des données dans leur établissement.

Dans ce chapitre, nous examinerons pourquoi certains établissements canadiens pourraient vouloir mener chez eux une **évaluation de la maturité de la GDR** en tenant compte, particulièrement, des exigences en matière de stratégie institutionnelle de GDR mises en place par les trois organismes fédéraux de financement de la recherche au Canada (les **trois organismes subventionnaires**). Nous étudierons également le développement du MEMAC, comment le compléter et comment utiliser ses résultats. Pour terminer, nous soulignerons l'importance des efforts de la communauté dans la création de cet outil.

Accéder au MEMAC ici: anglais (https://zenodo.org/record/5745493#.Y08bdGjMKUk), français (https://zenodo.org/record/5745894#.Y08k_GjMKUk)

## Le besoin: comment évaluer les services de GDR d'un établissement

Au printemps 2021, les Instituts de recherche en santé du Canada (IRSC), le Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG) et le Conseil de recherches en sciences humaines du Canada (CRSH) ont publié leur Politique des trois organismes sur la GDR (https://science.gc.ca/site/science/fr/financement-interorganismes-recherche/politiques-lignes-directrices/gestion-donnees-recherche/politique-trois-organismes-gestion-donnees-recherche). Cette politique tant attendue cherche à appuyer l'excellence en recherche au Canada en veillant à ce que les chercheuses et chercheurs appliquent de bonnes pratiques de GDR et d'intendance des données et que leurs établissements d'attache les soutiennent. Les trois organismes subventionnaires visent les normes d'excellence les plus rigoureuses – ils s'attendent à ce que la recherche soit effectuée d'une manière éthique, que les fonds accordés soient bien employés, que les expériences et les études

soient généralisables et que les recherches soient aussi accessibles que possible (Gouvernement du Canada, 2021). Pour en faire la preuve, les établissements doivent créer et publier une stratégie de GDR qui fait état de leur engagement envers les principes de GDR et qui explique la façon dont ils appuieront les chercheuses et chercheurs dans la mise en œuvre de ces principes (voir la partie 3.1 de la politique).

Puisqu'une seule stratégie ne pourra convenir à toutes les situations, chaque établissement doit tenir compte des circonstances qui lui sont propres, telles que sa taille, l'intensité de sa recherche et sa capacité existante de GDR. Mais comment un établissement détermine-t-il sa capacité de GDR ou à quoi sa stratégie doit-elle ressembler? Pour aider dans ce processus, l'Alliance de recherche numérique du Canada (anciennement le réseau Portage) a publié en 2018 son Modèle pour l'élaboration de stratégie institutionnelle de GDR (https://zenodo.org/record/5745923#.Y-v76BOZPao). Mis à jour en novembre 2021, ce modèle présente un processus en cinq étapes pour renseigner et façonner l'élaboration d'une stratégie de GDR qui répond aux besoins locaux et aux capacités des ressources. Nous nous attarderons sur la deuxième étape du processus, celle qui encourage les établissements à évaluer l'état de leur GDR en utilisant des modèles et outils d'évaluation.

## Qu'est-ce qu'un modèle d'évaluation de la maturité? Et pourquoi le Canada en a-t-il besoin?

Les outils et **modèles d'évaluation de la maturité** évaluent la maturité et le niveau de préparation d'un établissement à fournir des services en GDR et aident à déterminer le degré de sophistication d'un service ou d'un produit. Une des caractéristiques couramment utilisée par ces modèles est l'échelle. Celle-ci sert à représenter la maturité d'un organisme relativement à certaines capacités particulières – autrement dit, dans quelle mesure la mise en œuvre du processus de l'organisme est fiable (Rans et White, 2017). La grille d'évaluation de la maturité permet à l'utilisateur de quantifier les capacités et favorise l'amélioration continue du processus.

Sur le plan international, la GDR est déjà bien implantée pour favoriser l'excellence en recherche et plusieurs modèles de maturité ont été développés. Toutefois, quand les établissements canadiens ont commencé à utiliser ces modèles pour évaluer l'état de la GDR sur leur campus, ils se sont aperçus que ces outils ne cadraient pas bien au contexte canadien en GDR. Par exemple, les établissements canadiens ne sont pas tenus d'avoir des politiques en GDR, contrairement aux établissements dans d'autres pays.

Après la publication de la politique des trois organismes sur la GDR en 2021, les membres de la communauté nationale de GDR ont entamé des discussions informelles sur la façon dont les établissements devraient s'y

prendre pour l'élaboration de leur stratégie en GDR. Le Groupe d'experts national sur la formation (GENF)[1] réunissant des membres de la communauté de recherche et des gens qui les soutiennent a décidé de créer une série de webinaires et d'ateliers qui serait présentée en octobre 2021 afin de rassembler des personnes représentant différents établissements pour discuter du développement de leur stratégie. En préparant cette série pour l'automne, plusieurs membres ont constaté l'absence d'un modèle canadien d'évaluation de la maturité qui pourrait servir aux établissements dans la deuxième étape du développement de leur stratégie. Le GENF a convenu qu'un outil proprement canadien de modèle d'évaluation de la maturité serait un bon point de départ pour discuter des capacités institutionnelles de GDR qui doivent être incluses dans les stratégies afin de s'aligner avec la politique des trois organismes sur la GDR. En avril 2021, un plus petit groupe a entamé l'élaboration de ce qui allait devenir la première version du MEMAC, à temps pour la tenue de l'atelier en octobre. Nous – les autrices de ce chapitre – en plus de Shahira Khair de l'Université de Victoria, faisions partie de ce groupe.

# Comment le MEMAC a-t-il été créé?

## L'analyse du contexte des modèles d'évaluation de la maturité

Comme première étape, nous avons analysé plusieurs outils internationaux d'évaluation. Bien que les modèles disponibles soient excellents, ils comprenaient des sections qui n'étaient pas applicables au Canada, en plus de comporter certaines lacunes – des éléments que nous devions alors inclure dans le modèle canadien, dont l'exigence des organismes subventionnaires pour une stratégie institutionnelle de GDR. À la suite de cette analyse, nous nous sommes concentrées sur les éléments des trois modèles les plus populaires pour nous aider à développer le MEMAC – l'outil du *Research Infrastructure Self-Evaluation Framework (https://www.dcc.ac.uk/guidance/how-guides/RISE)* (RISE) publié en 2017 par le Digital Curation Centre (https://www.dcc.ac.uk/) (DCC); l'outil d'évaluation de l'offre en GDR (*Evaluate your RDM Offering* (https://sparceurope.org/evaluate-your-rdm-offering/)) de SPARC Europe (https://sparceurope.org/) (s.d.); et le *Data Management Framework* (https://web.archive.org/web/20220309174711/https://www.ands.org.au/__data/assets/pdf_file/0005/737276/Creatinga-data-management-framework.pdf) du Australian National Data Service (ANDS, 2018)[2]. Notre modèle canadien s'est surtout construit à partir du RISE, avec l'ajout de certains éléments inspirés du modèle SPARC et des cadres du ANDS.

1. Le GENF fait partie du réseau d'expertes et experts et est affilié à l'Alliance de recherche numérique du Canada.
2. Depuis le développement du MEMAC, le ANDS a été intégré au Australian Research Data Commons (ARDC) https://ardc.edu.au/ (https://ardc.edu.au/)

## Un aperçu du MEMAC

Dans notre outil canadien, nous détaillons la raison d'être, les objectifs et la définition du MEMAC, et nous fournissons une section sur la marche à suivre pour le compléter. Il y a quatre tableaux à remplir par les partenaires de recherche :

- Politiques et procédures de l'établissement;
- Infrastructure informatique;
- Services de soutien;
- Soutien financier.

Chaque tableau comporte cinq colonnes:

- l'élément à évaluer;
- la définition de l'élément;
- son niveau de maturité (l'état d'avancement de la GDR de l'établissement);
- son échelle (qui peut accéder au service ou au soutien);
- tout commentaire nécessaire pour expliquer la note attribuée à l'élément.

Sous chaque tableau, il y a un espace pour indiquer la date d'achèvement ainsi que le nom et le rôle de chaque personne qui a participé à le remplir; de cette façon, les gens qui consultent le tableau peuvent savoir avec qui communiquer s'ils veulent poser des questions ou émettre des commentaires. Le MEMAC a aussi été conçu pour une utilisation future, il est donc important de savoir qui a participé aux versions antérieures.

Nous voulions aussi nous assurer que les termes utilisés soient bien définis alors nous avons inclus une page de définitions des niveaux de maturité et des échelles spécifiques à chacun des tableaux, en y ajoutant quelques pistes de réflexion pour aider à les remplir.

## La première version du MEMAC

Après avoir reçu une série de commentaires des membres de la communauté de GDR, nous avons terminé une ébauche de la version originale anglophone (MAMIC ou *Maturity Assessment Model in Canada*) qui a ensuite été traduite en français et présentée en octobre 2021 aux participantes et participants des ateliers sur les stratégies institutionnelles.

À noter que certains éléments n'ont pas été traités dans cette version originale, des modifications devront donc être faites pour les versions futures. Par exemple, une nouvelle version devrait tenir compte de la souveraineté

des données autochtones. De nouveaux moyens de présentation de l'outil pourraient également être explorés, tels que le développement d'un outil en ligne qui permettrait aux personnes qui l'utilisent de produire différents types de graphiques, au même titre que l'outil du SPARC.

Ces révisions seront particulièrement utiles pour les établissements qui prévoient mener ce type d'évaluation régulièrement, dans le cadre du processus de révision de leur stratégie institutionnelle de GDR ou en guise d'amélioration continue de leurs services. Ce serait également utile de pouvoir appliquer le MEMAC à l'échelle nationale afin de relever et de traiter des lacunes et de présenter les contextes où certains établissements pourraient faire appel à des ressources nationales.

# L'utilisation du MEMAC

Le MEMAC peut être utilisé pour déterminer ce qui existe en matière de ressources et de services en GDR, mais aussi qui est responsable de ces différentes formes de soutien. Cette connaissance permet aux établissements d'aider leurs chercheuses et chercheurs à améliorer leur gestion des données et de cibler les modifications à apporter pour bonifier leur offre de services. L'utilisation du modèle implique une coordination entre différents services sur le campus, dont la bibliothèque, le bureau de la recherche, le bureau de l'éthique et les responsables des TI.

## Les catégories et mesures

Avant de commencer, les partenaires de recherche qui complètent le MEMAC devraient discuter du processus pour que tous comprennent bien la façon d'appliquer les échelles ainsi que les mesures et pour s'assurer que les décisions principales liées au processus soient documentées. L'évaluation de chacune des catégories (Politiques et procédures de l'établissement, Infrastructure informatique, Services de soutien, Soutien financier) peut se faire dans son propre tableau – voir l'Annexe 2 pour un exemple complet du tableau de la catégorie Politiques et procédures de l'établissement – avec les trois différentes mesures :

Mesure 1 : Le **niveau de maturité** de l'élément de l'établissement est noté sur une échelle de 0 à 5, allant de « n'existe pas » OU « inconnu » à « robuste et se concentre sur l'évaluation continue ». À noter que la première note de cette échelle est bien 0 et non 1, parce que certains établissements ne sont peut-être pas en mesure de fournir un service ou du soutien ou ils estiment ne pas en avoir besoin. La catégorie « n'existe pas » ne sert pas à indiquer le niveau de maturité, mais plutôt à reconnaître que l'élément n'est pas disponible aux chercheuses et chercheurs de l'établissement.

Mesure 2 : L'**échelle** est utilisée pour identifier celles et ceux qui peuvent accéder au service ou au soutien. L'élément pourrait ne pas s'appliquer à certaines personnes ou ne pas être disponible à toute la population.

On peut alors déterminer si les services de l'établissement sont offerts de façon équitable et appropriée ou s'il existe des problèmes d'accessibilité.

Mesure 3 : Les **commentaires** sont probablement la plus importante des mesures parce qu'ils peuvent identifier certaines forces ou faiblesses, en plus de fournir certaines pistes de discussion. C'est aussi l'endroit où lister des outils régionaux, nationaux, consortiaux ou autres qui servent de complément à la maturité de la GDR de l'établissement. Déterminer le niveau de maturité ou l'échelle d'un élément peut être difficile s'il existe plusieurs initiatives à l'intérieur de cet élément (p. ex., plusieurs unités qui offrent des services semblables de gestion des données). La section des commentaires est donc l'endroit idéal pour expliquer ce type de cas.

## Compléter le MEMAC

Les données recueillies dans le MEMAC sont à l'usage exclusif des établissements qui le remplissent ; aucun autre organisme ne recueillera ces données. De plus, celles et ceux qui complètent le MEMAC décident des moyens employés pour recueillir et utiliser leurs données.

Le MEMAC peut être complété par un seul individu, mais nous recommandons l'implication d'un groupe de partenaires de recherche concernés qui représentent les domaines évalués. Par exemple, l'infrastructure informatique devrait être évaluée par des gens du service des TI et le soutien financier devrait être évalué par des personnes qui fournissent les services et le soutien en GDR (p. ex., les bibliothèques, les TI, les services de recherche).

Après que le MEMAC ait été rendu public, la communauté de GDR a partagé avec nous quatre exemples d'un processus achevé du MEMAC; tous se ressemblaient. Trois des quatre établissements avaient constitué un groupe de travail qui regroupait des bibliothécaires, des responsables en TI, des membres du bureau de la recherche et soit des chercheuses ou chercheurs ou des partenaires d'industrie. Pour l'autre établissement, seule la bibliothécaire de données a complété le document, en faisant appel à ses collègues du bureau de la recherche pour combler les informations manquantes. Cette méthode s'est avérée moins efficace et beaucoup plus laborieuse que les autres. Les résultats du MEMAC de chaque cas ont été présentés à un comité de GDR plus large à des fins de discussion.

# Les avantages du MEMAC

En élaborant des stratégies et du soutien en GDR, les partenaires doivent réfléchir à l'état et à la portée des services et du soutien en GDR de leur établissement, ainsi qu'aux besoins et aux souhaits futurs. Un modèle d'évaluation de la maturité, comme le MEMAC, peut aider à identifier les lacunes, les forces, les faiblesses, les

difficultés et les opportunités qui existent dans le paysage des données de recherche. Les établissements peuvent ainsi déterminer où les ressources et les efforts devraient être dirigés pour pouvoir mettre en place le soutien nécessaire au succès de leurs chercheuses et chercheurs.

Une utilisation efficace de cet outil permet une évaluation complète et représentative. Le processus nécessite toutefois la collaboration et la contribution d'une variété de partenaires de recherche; un des avantages de l'utilisation du MEMAC se rapporte donc à cette opportunité de discussion qui ouvre la voie à la création ou au renforcement de relations dans l'équipe. L'environnement de GDR institutionnel en sort consolidé et de nouvelles opportunités de dialogues, de collégialité et de partenariats à l'extérieur du cadre de la GDR apparaissent. Par exemple, une communication peut s'établir entre les services des TI et l'unité interne des TI de la bibliothèque, permettant ainsi une plus grande intégration des services de la bibliothèque et des ressources en TI.

Rassembler différents partenaires de recherche autour d'un outil partagé peut renseigner sur la complexité de la GDR et l'étendue des efforts déployés à travers l'établissement. Le processus peut contribuer à défaire les silos et à identifier les domaines d'expertise dans l'établissement, à établir des contacts ainsi qu'à faire ressortir les domaines qui pourraient profiter de la collaboration et des discussions autour de la stratégie et des services institutionnels, de l'allocation de ressources et de considérations budgétaires. L'utilisation d'un même outil sur une longue période peut aussi aider à suivre l'évolution des développements et des progrès dans l'établissement.

Plus largement, le MEMAC peut aussi favoriser les échanges entre établissements canadiens. En relevant les endroits où des ressources externes sont disponibles ou en développement, les établissements peuvent mieux décider où ils veulent investir au niveau local. De plus, l'identification de lacunes à travers les établissements peut mener à la création de nouvelles initiatives nationales. Ainsi, nous réduisons le dédoublement des efforts pour corriger chaque problème au niveau institutionnel, ce qui est long, coûteux et nécessite du personnel dédié à la tâche.

## Conclusion

Ce chapitre a présenté le MEMAC sous deux angles : comme un outil que le personnel et les établissements de GDR peuvent utiliser dans leurs travaux actuels ou futurs, et comme un exemple d'initiative où les membres de la communauté canadienne de GDR ont pu se doter d'outils qui permettent à toutes et tous d'être plus efficaces et performants. Nous avons identifié un besoin et avons entrepris des démarches pour y répondre en utilisant les mêmes compétences et techniques utilisées ailleurs dans notre travail : mener des analyses de contexte et des revues de littérature, travailler en équipe et développer du matériel testé et commenté par des utilisatrices et utilisateurs. Nous avons également fait appel aux ressources et aux individus

qui étaient disponibles pour nous appuyer dans le développement et la diffusion de l'outil, notamment, la communauté nationale du réseau d'expertes et experts de GDR et l'équipe de GDR de l'Alliance canadienne de recherche numérique du Canada.

### Questions de réflexion

Choisissez une catégorie du MEMAC (https://zenodo.org/record/5745894#.Y-vbphOZPaq) à examiner, et complétez ce qui suit :

- Identifiez les partenaires de recherche d'un établissement dont l'implication serait pertinente pour dresser un portrait juste de l'état du soutien de la GDR dans cette catégorie. Comment feriez-vous pour encourager leur participation?
- Énumérez quatre façons dont le MEMAC peut aider à évaluer le niveau de soutien en GDR dans un établissement.

### Éléments clés à retenir

- Un modèle d'évaluation de la maturité est un outil qui détermine le degré de sophistication d'un service ou d'un produit.
- Des modèles d'évaluation de la maturité propres à la GDR ont été développés par différents organismes internationaux et ont été utilisés depuis de nombreuses années pour l'évaluation des services de soutien en GDR.
- Le MEMAC a été développé pour traiter des besoins particuliers des établissements canadiens dans l'élaboration de leurs stratégies institutionnelles de GDR.
- Compléter le MEMAC permet aux partenaires de recherche de discuter et d'évaluer l'état de la GDR de leur établissement, de mieux comprendre l'étendue de l'offre et du soutien en GDR et de travailler en collaboration avec d'autres services.
- Il existe de nombreuses façons d'utiliser le MEMAC pour contribuer aux discussions et à la

prise de décisions des établissements en matière de GDR. Le MEMAC peut favoriser l'avancement et le progrès en permettant aux partenaires de recherche de l'établissement de prendre des décisions fondées sur les preuves qui auront un impact sur le développement futur de ses services et ses ressources en GDR.

## Lectures et ressources supplémentaires

Alliance de recherche numérique du Canada. (s.d.). *Gestion des données de recherche.* https://alliancecan.ca/fr/services/gestion-des-donnees-de-recherche (https://alliancecan.ca/fr/services/gestion-des-donnees-de-recherche)

Alliance de recherche numérique du Canada. (s.d.). *Réseau d'experts nationaux pour la gestion des données de recherche*. https://alliancecan.ca/fr/services/gestion-des-donnees-de-recherche/reseau-dexperts (https://alliancecan.ca/fr/services/gestion-des-donnees-de-recherche/reseau-dexperts)

Alliance de recherche numérique du Canada. (2021). Webinaires sur l'application de la politique des trois organismes :

- Session 1: Introduction à la politique des trois organismes en GDR et plans de gestion des données
  - anglais: https://www.youtube.com/watch?v=FftT68eXlNQ (https://www.youtube.com/watch?v=FftT68eXlNQ)
  - français: https://www.youtube.com/watch?v=lD6TrMukSBQ (https://www.youtube.com/watch?v=lD6TrMukSBQ)
- Session 2: Dépôt de données
  - anglais:: https://www.youtube.com/watch?v=oBYdbpZkXNg (https://www.youtube.com/watch?v=oBYdbpZkXNg)
  - français: https://www.youtube.com/watch?v=h7rUIrEceoI (https://www.youtube.com/watch?v=h7rUIrEceoI)
- Session 3: Stratégies institutionnelles
  - anglais: https://www.youtube.com/watch?v=TR8yv1dzHyI (https://www.youtube.com/watch?v=TR8yv1dzHyI)
  - français: https://www.youtube.com/watch?v=K2jh1HhjXb8 (https://www.youtube.com/watch?v=K2jh1HhjXb8)
- Session 4: Panel sur les stratégies institutionnelles

- anglais: https://www.youtube.com/watch?v=mPnE2KDHl0M (https://www.youtube.com/watch?v=mPnE2KDHl0M)
- français: https://www.youtube.com/watch?v=XUmGotnnslY (https://www.youtube.com/watch?v=XUmGotnnslY)

Australian Research Data Commons (ARDC). https://ardc.edu.au/ (https://ardc.edu.au/)

Borghi, J. (2016, 12 septembre). Building a user-friendly RDM maturity model. *University of California Curation Center (UC3).* https://uc3.cdlib.org/2016/09/12/building-a-user-friendly-rdm-maturity-model/ (https://uc3.cdlib.org/2016/09/12/building-a-user-friendly-rdm-maturity-model/)

Fry, J., Doiron, J., Létourneau, D., Perrier, L., Perry, C. et Watkins, W. (2017, 31 janvier). *Portrait de la formation sur la gestion des données de recherche au Canada : Livre blanc*. https://dx.doi.org/10.14288/1.0372050 (https://dx.doi.org/10.14288/1.0372050)

Groupe de travail chargé de la révision du modèle de stratégie institutionnelle en matière de GDR. (2021). *Modèle pour l'élaboration de stratégie institutionnelle de gestion des données de la recherche (3.0)*. Zenodo. https://zenodo.org/record/5745923 (https://zenodo.org/record/5745923)

Jacob, B., Whyte, A., Meyer, A., D'haenens, S., Hartmann, N. K. et Weiß, N. (2019, 2 octobre). *Using RISE, an international perspective* [Résumé]. 15th International Digital Curation Conference (IDCC), Dublin, Irlande. https://doi.org/10.5281/zenodo.3565440 (https://doi.org/10.5281/zenodo.3565440)

Jones, S., Pryor, G. et Whyte, A. (2012). Developing research data management capability: The view from a national support service. Dans R. Moore, K. Ashley, et S. Ross (dir.), *Proceedings of the 9th International Conference on Preservation of Digital Objects (iPRES)*, (pp. 142-149). University of Toronto Faculty of Information. https://phaidra.univie.ac.at/detail/o:293775 (https://phaidra.univie.ac.at/detail/o:293775)

Jones, S., Rans, J., Sisu, D. et Whyte, A. (2014). Reshaping the DCC institutional engagement programme. *International Journal of Digital Curation, 9*(2), 47-64.https://doi.org/10.2218/ijdc.v9i2.334 (https://doi.org/10.2218/ijdc.v9i2.334)

Kouper, I., Fear, K., Ishida, M., Kollen, C. et Williams, S. C. (2017). *Research data services maturity in academic libraries*. http://dx.doi.org/10.14288/1.0343479 (http://dx.doi.org/10.14288/1.0343479)

Perry, C. l., Fry, J. et Doiron, J. (2017, 1 juin). *Portaging the landscape: Developing and delivering a national RDM training infrastructure in Canada* [Présentation]. IASSIST 2017 https://doi.org/10.5281/zenodo.4551708 (https://doi.org/10.5281/zenodo.4551708)

SPARC Europe. (s.d.). *How Open Are You*? https://sparceurope.org/what-we-do/open-access/sparc-europe-

open-access-resources/open-research-checklist-institutions/ (https://sparceurope.org/what-we-do/open-access/sparc-europe-open-access-resources/open-research-checklist-institutions/)

# Bibliographie

ANDS. (2018, 23 mars). *Creating a data management framework*. https://web.archive.org/web/20220309174711/https://www.ands.org.au/__data/assets/pdf_file/0005/737276/Creatinga-data-management-framework.pdf (https://web.archive.org/web/20220309174711/https://www.ands.org.au/__data/assets/pdf_file/0005/737276/Creatinga-data-management-framework.pdf)

Fry, J., Dearborn, D., Farrell, A., Khair, S. et Ripp, C. (2021, 30 novembre). *Modèle d'évaluation de la maturité de la GDR au Canada (MEMAC) (1.0)*. Zenodo. https://zenodo.org/record/5745894 (https://zenodo.org/record/5745894)

Gouvernement du Canada. (2021). *Politique des trois organismes sur la gestion des données de recherche*. https://science.gc.ca/site/science/fr/financement-interorganismes-recherche/politiques-lignes-directrices/gestion-donnees-recherche/politique-trois-organismes-gestion-donnees-recherche (https://science.gc.ca/site/science/fr/financement-interorganismes-recherche/politiques-lignes-directrices/gestion-donnees-recherche/politique-trois-organismes-gestion-donnees-recherche)

Rans, J. et Whyte, A. (2017). *Using RISE, the research infrastructure self-evaluation framework v.1.1*. Digital Curation Centre. www.dcc.ac.uk/guidance/how-guides/RISE (http://www.dcc.ac.uk/guidance/how-guides/RISE)

SPARC Europe. (s.d.) *Evaluate your RDM offering*. https://sparceurope.org/evaluate-your-rdm-offering/ (https://sparceurope.org/evaluate-your-rdm-offering/).

## À propos des auteurs

### Jane Fry

Jane Fry est bibliothécaire des services de données à la bibliothèque MacOdrum de l'Université Carleton, où la gestion des données de recherche est l'une de ses responsabilités. Elle est également responsable du groupe de travail sur la stratégie institutionnelle de GDR de l'Université Carleton. Elle a également présidé le Groupe d'experts sur la formation de l'Alliance de recherche numérique pendant quatre ans et en est encore membre aujourd'hui.

## Jennifer Abel

Jennifer Abel est spécialiste de la gestion des données de recherche au sein du bureau des services de recherche de l'Université de Calgary et coordonne le développement de la stratégie en GDR de l'Université de Calgary. Elle a précédemment travaillé en tant que coordinatrice de la formation pour le réseau Portage et l'équipe GDR de l'Alliance de recherche numérique du Canada. Elle est titulaire d'un doctorat en linguistique et d'un MLIS de l'Université de Colombie-Britannique.

## Dylanne Dearborn

Dylanne Dearborn est coordinatrice de la gestion des données de recherche et bibliothécaire des données de recherche pour les sciences et l'ingénierie à l'Université de Toronto dans la bibliothèque des cartes et des données. Elle a présidé le Groupe d'experts sur la recherche et l'intelligence de l'Alliance de recherche numérique pendant trois ans et en est encore membre aujourd'hui.

## Alison Farrell

Alison Farrell est bibliothécaire à la gestion des données de recherche et aux services publics à la bibliothèque des sciences de la santé de l'Université Memorial de Terre-Neuve et membre du groupe de travail sur la stratégie institutionnelle de GDR de l'Université Memorial. Elle a été coprésidente du groupe de travail sur les stratégies institutionnelles de Portage, dont le mandat était de fournir du matériel pédagogique sur l'élaboration de stratégies institutionnelles.

## Chantal Ripp

Chantal Ripp est bibliothécaire de recherche au sein de l'équipe interdisciplinaire de données de l'Université d'Ottawa. Elle est membre du groupe consultatif de GDR de l'université chargé d'élaborer une stratégie institutionnelle. Elle est également membre d'un certain nombre d'autres comités locaux et nationaux, notamment le comité de développement professionnel de l'Initiative de démocratisation des données (IDD) et le Groupe d'experts sur la planification de la gestion des données de l'Alliance de recherche numérique du Canada.

PARTIE III

# MÉTHODES DE TRAVAIL AVEC LES DONNÉES DE RECHERCHE

7.

# LE NETTOYAGE DE DONNÉES DANS LE PROCESSUS DE GESTION DES DONNÉES DE RECHERCHE

Lucia Costanzo

**Objectifs d'apprentissage**

À la fin de ce chapitre, vous pourrez :

1. Expliquer pourquoi il est important de nettoyer vos données.
2. Vous rappeler les tâches courantes de nettoyage des données.
3. Mettre en œuvre les tâches courantes de nettoyage des données en utilisant OpenRefine.

## Qu'est-ce que le nettoyage des données?

Vous avez peut-être entendu parler de la règle du 80/20? La plupart des chercheuses et chercheurs consacrent 80% de leur temps à la recherche, au nettoyage et à la réorganisation de grandes quantités de données et seulement 20% de leur temps à analyser ces données.

En entamant un projet de recherche, vous utiliserez soit des données primaires générées à partir de vos propres expériences, soit des données secondaires issues des expériences d'une autre équipe de recherche. Une fois les données obtenues pour répondre à votre ou vos question(s) de recherche, vous aurez besoin de temps pour les examiner et les comprendre. Les données peuvent se retrouver dans des formats difficiles à analyser. Au cours de l'étape du nettoyage des données, vous utiliserez des pratiques de **gestion des données de recherche** (GDR). Le processus de nettoyage des données peut être long et fastidieux, mais il est essentiel pour assurer la précision et la qualité de votre recherche.

Le **nettoyage des données** peut sembler évident, mais c'est une étape où plusieurs chercheuses et chercheurs éprouvent des difficultés. George Fuechsel, un programmeur et instructeur de la compagnie IBM a été le premier à utiliser l'expression « *garbage in, garbage out* » (déchet qui entre, déchet qui sort) (Lidwell *et al.*, 2010) pour rappeler à ses étudiantes et étudiants que l'ordinateur ne traite que ce qu'on lui donne – peu importe que l'information soit bonne ou mauvaise. Le même principe s'applique aux chercheuses et chercheurs; peu importe la qualité de vos méthodes, l'analyse ne dépend que de la qualité des données. Autrement dit, les résultats et conclusions d'une étude seront aussi fiables que les données utilisées.

L'utilisation de données nettoyées vous permet de ne pas perdre de temps en analyses inutiles.

# Six actions principales de nettoyage et de préparation

Le nettoyage et la préparation des données peuvent se résumer à six actions principales : découvrir, structurer, nettoyer, enrichir, valider et publier. Elles sont menées tout au long du projet de recherche afin de maintenir l'organisation des données. Examinons chacune des actions de plus près.

## 1. Découvrir les données

Cette étape importante consiste à découvrir ce que les données peuvent révéler. Elle est désignée comme l'**analyse exploratoire des données** (AED). Le concept de l'AED a été développé dans les années 70 par le mathématicien américain John Tukey. D'après un mémoire, « Tukey a souvent comparé l'AED à un travail de détective. Le rôle de l'analyste de données est d'écouter les données de toutes les manières possibles jusqu'à ce qu'une histoire plausible se dégage des données[1] » [traduction] (Behrens, 1997). L'AED est une approche employée pour mieux comprendre les données par le biais de méthodes quantitatives et graphiques.

Les méthodes quantitatives synthétisent les caractéristiques des variables en utilisant des mesures de tendance centrale, dont le mode, la médiane et la moyenne arithmétique qui est la plus courante d'entre elles. Les mesures de dispersion indiquent à quelle distance du centre il est vraisemblable de retrouver des points de données. La variance, l'écart-type, l'étendue et l'écart interquartile constituent tous des mesures de dispersion. D'un point de vue quantitatif, la distribution peut être évaluée en utilisant une mesure d'asymétrie. Des histogrammes, des boîtes à moustaches et parfois aussi des diagrammes à tiges et à feuilles sont utilisés pour

1. "*Tukey often likened EDA to detective work. The role of the data analyst is to listen to the data in as many ways as possible until a plausible 'story' of the data is apparent*"

visualiser rapidement chacune des variables en fonction de la tendance centrale, de la dispersion, de la modalité, de l'étendue et des observations aberrantes.

L'examen des données par le biais de techniques d'AED permet d'identifier des tendances sous-jacentes et des anomalies, aide à établir des hypothèses et vérifie des suppositions liées à l'analyse. Examinons maintenant l'action de structurer les données.

## 2. Structurer les données

En fonction de la ou des question(s) de recherche, vous pourrez avoir à organiser les données de différentes façons pour différents types d'analyses. Prenons comme exemple les données par mesures répétées – quand chaque unité expérimentale ou sujet est mesuré à plusieurs moments ou dans différentes conditions.

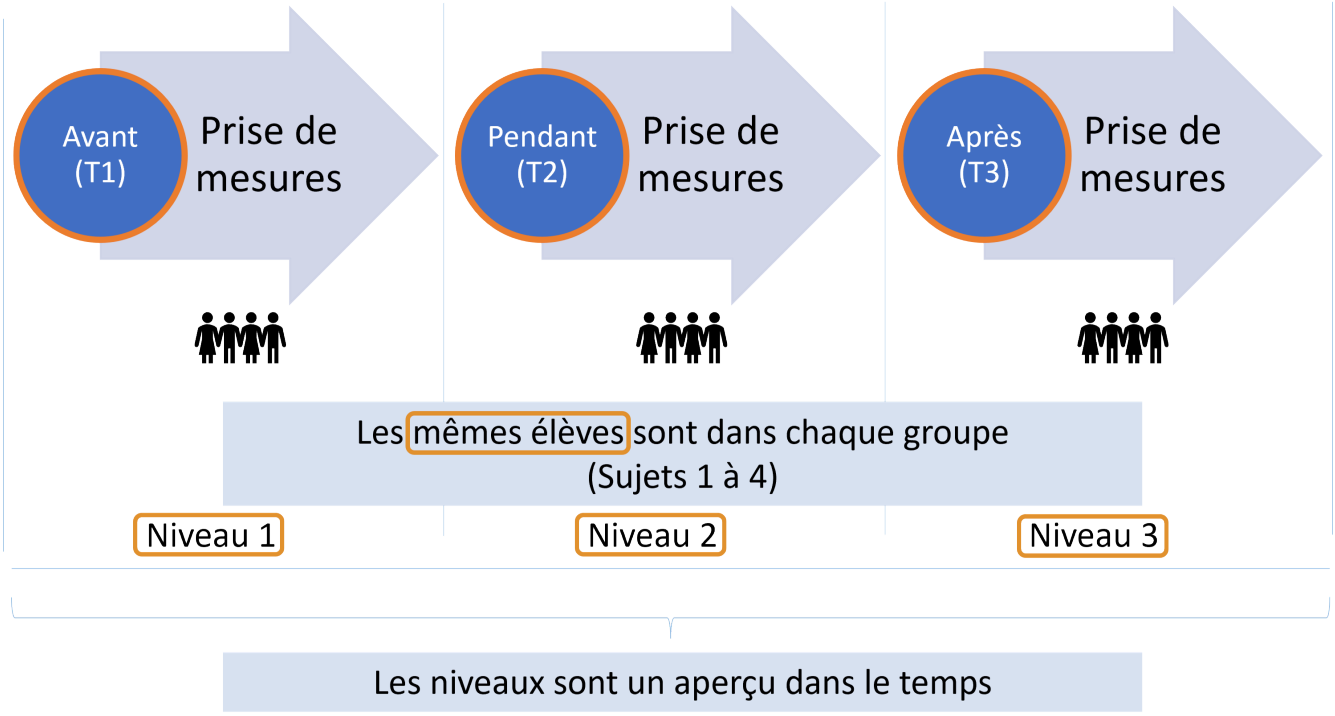



Figure 1. Une enquête sur les effets d'un programme de petit-déjeuner.

Imaginons, pour la figure 1, que des chercheuses étudient les effets d'un programme de petits-déjeuners auprès d'élèves de 6e année et qu'elles veulent recueillir des résultats de tests à trois moments différents, dont avant (T1), pendant (T2) et après (T3) la tenue du programme matinal. À noter que chaque groupe est constitué des mêmes élèves et que chaque élève est mesuré aux différents moments. Chaque mesure au cours de l'étude est un aperçu dans le temps. Il y a deux façons de structurer des données de mesures répétées : les formats longs et larges.

Le tableau 1 illustre des données structurées en format long avec chaque élève de l'étude représenté par trois rangées de données, une pour chacun des moments dans le temps où des résultats ont été recueillis. Dans la première rangée, Élève 1 au Temps 1 (avant le programme de petits-déjeuners) a eu un résultat de 50 pour le test. Dans la deuxième rangée, Élève 1 au Temps 2 (à mi-chemin du programme de petits-déjeuners) a eu un résultat plus élevé, de 65. Et dans la troisième rangée, Élève 1 au Temps 3 (après le programme) a obtenu un résultat de 80.

| ID | TEMPS | NOTE |
|---|---|---|
| 1 | 1 | 50 |
| 1 | 2 | 65 |
| 1 | 3 | 80 |
| 2 | 1 | 70 |
| 2 | 2 | 75 |
| 2 | 3 | 90 |
| : | : | : |

Tableau 1. Données structurées en format long.

Le format large, tel qu'illustré par le tableau 2, utilise une seule rangée pour chacune des observations ou des personnes participantes, et chaque mesure ou résultat se trouve dans sa colonne distincte.

| ID | TEST1 | TEST2 | TEST3 |
|---|---|---|---|
| 1 | 50 | 65 | 80 |
| 2 | 70 | 75 | 90 |
| : | : | : | : |

Tableau 2. Données structurées en format large.

Dans le format large, l'ensemble des résultats des tests d'un seul élève se trouve dans la même rangée, et chaque résultat de test individuel de l'élève est dans sa propre colonne. En consultant la première rangée, Élève 1 a obtenu un résultat de 50 sur le test avant le programme de petits-déjeuners, il a ensuite obtenu 65 à mi-chemin du programme et 80 après le programme.

Bref, un format long de données utilise plusieurs rangées pour chaque observation ou personne participante, tandis que le format large de données utilise une rangée par observation. La façon de structurer vos données (format long ou large) sera déterminée par le modèle ou l'analyse statistique choisie. Il est possible que vous ayez à structurer vos données dans les deux formats pour réaliser vos objectifs d'analyse.

Structurer les données est une activité fondamentale de nettoyage et de préparation des données; l'étape vise à réorganiser les données en vue d'une analyse statistique particulière. Les données peuvent contenir des irrégularités ou des anomalies, ce qui peut avoir un impact sur la fiabilité des modèles de la chercheuse ou du chercheur. Examinons de plus près le nettoyage des données pour que votre analyse puisse fournir des résultats plus précis.

## 3. Le nettoyage des données

Le nettoyage des données est essentiel pour assurer la qualité de votre analyse. Les neuf conseils suivants abordent, à l'aide d'exemples pratiques, une série de problèmes couramment rencontrés lors du nettoyage des données.

# Conseil n° 1: la vérification de l'orthographe

CONSEIL **01**

VÉRIFICATION DE L'ORTHOGRAPHE

Trouver les erreurs de frappe ou les variations orthographiques

Le repérage de fautes de frappe ou de variations orthographiques est l'une des tâches les plus importantes dans le nettoyage des données. Vous pouvez utiliser un logiciel de vérification orthographique pour identifier et corriger les erreurs d'orthographe ou de saisie des données.

Ces logiciels de vérification peuvent aussi être utilisés pour uniformiser les noms propres. Si, par exemple, un jeu de données contient des entrées pour « Université de Guelph » et « UOG » et « U of G » et « Guelph University » (tableau 3), chaque variation orthographique sera considérée comme un établissement différent. L'orthographe choisie importe peu; l'important, c'est qu'elle soit uniforme à travers le jeu de données.

**Exercice du conseil n° 1 : la vérification orthographique**

Parcourez le tableau 3 et uniformisez le nom dans la colonne ETABLISSEMENT à « Université de Guelph ».

| ID | AGE | ETABLISSEMENT | NOTE |
|---|---|---|---|
| 1 | 17 | Universté de Guelph | 88 |
| 2 | 21 | UOG | 60 |
| 3 | 18 | Université de Guelph | 80 |
| 4 | 19 | Université de Guelph | 75 |
| 8 | 18 | Université de Guelph | 72 |
| 12 | 21 | Université de Guelph | 60 |
| 13 | 18 | Université de Guelph | 80 |
| 14 | 19 | Guelph University | 77 |
| 15 | 18 | Université de Guelph | 49 |
| 16 | 21 | U of G | 60 |
| 17 | 18 | Université de Guelph | 88 |
| 19 | 19 | Guelph University | 73 |
| 20 | 18 | Université de Guelph | 72 |

Tableau 3. Des données qui nécessitent une vérification orthographique.

Voir le solutionnaire pour les réponses. Les fichiers de données pour les exercices de ce chapitre sont disponibles (en anglais uniquement) dans le dépôt Borealis (https://borealisdata.ca/dataverse/datacleaning).

## Conseil n° 2: les doublons

Parfois, les données sont entrées ou générées manuellement en utilisant des méthodes qui peuvent entraîner des dédoublements de rangées. Vérifiez les rangées pour déterminer si certaines données sont en double et doivent être supprimées. Si chaque rangée comporte un numéro d'identification, celui-ci devrait être unique à chacune des observations. Dans cet exemple, deux observations ont le 3 comme numéro d'identification (tableau 4). Puisqu'elles ont toutes les deux des valeurs identiques, une des rangées devrait être supprimée.


**Exercice du conseil n° 2: les doublons**

Parcourez le tableau 4 et supprimez les entrées en double.

Indice : Si vous utilisez Excel, cherchez et utilisez la fonction « Valeurs en double ».

| ID | AGE | ETABLISSEMENT | NOTE |
|---|---|---|---|
| 1 | 17 | Université de Guelph | 88 |
| 2 | 21 | UOG | 60 |
| 3 | 18 | Université de Guelph | 80 |
| 4 | 19 | Université de Guelph | 75 |
| 3 | 18 | Université de Guelph | 80 |
| 12 | 21 | Université de Guelph | 60 |
| 13 | 18 | Université de Guelph | 80 |
| 14 | 19 | Guelph University | 77 |
| 15 | 18 | Université de Guelph | 49 |
| 16 | 21 | U of G | 60 |
| 17 | 18 | Université de Guelph | 88 |
| 19 | 19 | Guelph University | 73 |
| 20 | 18 | Université de Guelph | 72 |

Tableau 4. Des données avec des rangées en double, dont une à supprimer.

Voir le solutionnaire pour les réponses.

# Conseil n° 3: trouver et remplacer

CONSEIL **03**

TROUVER ET REMPLACER

Trouver et remplacer du texte

Avec certains remplacements bien choisis, il est possible d'obtenir des données relativement propres et de les organiser dans une forme intéressante en cherchant des tendances ou des répétitions dans un fichier. Cet exemple illustre le nombre d'observations d'oiseaux dans la ville de Guelph. En consultant la colonne LIEU, les abréviations « ch » et « Ch » sont remplacées par le mot complet « chemin » (tableau 5). Cette opération s'effectue à l'aide de la fonction trouver et remplacer.

**Exercice du conseil n° 3: trouver et remplacer**

Parcourez le tableau 5. Trouvez et remplacez tous les « ch » et « Ch » avec « chemin » dans la colonne LIEU.

INDICE: Faites attention en utilisant la fonction trouver et remplacer. Dans l'exemple ci-dessous, il y a des instances où les lettres « ch » ou « Ch » NE se rapportent PAS au mot chemin (p. ex., « Church » ou « March »); ces données seront donc remplacées par erreur. Ce type de modification peut être évité en insérant un espace avant la chaîne que vous cherchez (donc, « (*espace*)ch »). Expérimentez aussi la fonction de respect de la casse (majuscules ou minuscules) si elle est disponible. Conservez toujours une copie de vos données originales en cas de problèmes.

| ID | OISEAU | LIEU | TOTAL |
|---|---|---|---|
| 1 | 17 | ch Québec | 6 |
| 2 | 21 | chemin Cork | 5 |
| 3 | 18 | ch March | 8 |
| 4 | 19 | chemin Victoria | 5 |
| 5 | 18 | ch Steffler | 8 |
| 6 | 21 | ch Extra | 0 |
| 7 | 18 | ch Doyle | 2 |
| 8 | 19 | chemin Oxford | 7 |
| 9 | 18 | ch Dublin | 4 |
| 10 | 21 | chemin First | 6 |
| 11 | 18 | chemin Church | 1 |
| 12 | 19 | Ch North | 3 |
| 13 | 18 | ch Dulac | 2 |

Tableau 5. Des données avec un étiquetage non uniforme.

Voir le solutionnaire pour les réponses.

## Conseil n° 4: majuscules et minuscules

Le texte peut être tout en lettres minuscules, tout en majuscules ou seulement avoir une majuscule au début de chaque mot. Le texte peut être converti tout en minuscules, comme pour les adresses courriel; tout en majuscules, comme pour les abréviations de provinces, et dans les deux casses, comme pour les noms propres. Dans l'exemple de ce tableau, l'utilisation des majuscules et minuscules n'est pas uniforme. Parfois, les noms et courriels sont inscrits en majuscules, parfois en minuscules et parfois dans un format pour les noms propres (tableau 6).


**Exercice du conseil n° 4: majuscules et minuscules**

Dans la colonne NOM du tableau 6, modifiez le texte en remplaçant la première lettre du prénom et du nom avec une majuscule. Convertissez ensuite le texte dans la colonne COURRIEL à des lettres minuscules.

INDICE: Si vous utilisez Excel, cherchez les fonctions MAJUSCULE, MINUSCULE et NOMPROPRE.

| ID | AGE | NOM | COURRIEL |
|---|---|---|---|
| 1 | 17 | James Smith | JSMITH@GMAIL.COM |
| 2 | 21 | Michael Smith | MSMITH@GMAIL.COM |
| 3 | 18 | Robert SMITH | SMITHR@AOL.COM |
| 4 | 19 | Maria Garcia | mgarcia@hotmail.com |
| 8 | 18 | David SMITH | DAVIDSMITH@GMAIL.COM |
| 12 | 21 | Maria Rodriguez | mariaR@gmail.com |
| 13 | 18 | Mary SMITH | MARYSMITH@GMAIL.COM |
| 14 | 19 | Maria Hernandez | hernandez@outlook.com |
| 15 | 18 | Maria Martinez | mmartinez@mail.com |
| 16 | 21 | James Johnson | james@gmail.com |
| 17 | 18 | Lee Hartman | hartman@mail.com |
| 19 | 19 | Patricia SMITH | SMITHP@MAIL.CA |
| 20 | 18 | Ben SMITH | BENSMITH@MAIL.COM |

Tableau 6. Des données où l'utilisation des majuscules et minuscules n'est pas uniforme.

Voir le solutionnaire pour les réponses.

## Conseil n° 5: espaces et caractères qui ne s'impriment pas

Les espaces et les caractères qui ne s'impriment pas peuvent avoir des résultats imprévus quand vous effectuez des fonctions de tri, de filtrage ou de recherche. Les espaces en début ou en fin de mot, de multiples espaces intercalés ou les caractères qui ne s'impriment pas sont tous invisibles. Ils peuvent se faufiler lorsque vous importez des données à partir de pages Web, de documents Word ou PDF.


## Conseil n° 6: chiffres et signes

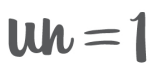
Il y a deux éléments à surveiller :

1. les données peuvent comprendre du texte
2. le signe négatif peut ne pas être standardisé

CONSEIL **06**

CHIFFRES ET SIGNES

Convertir les chiffres en valeurs numériques et normaliser le signe négatif




Vous pourriez obtenir un jeu de données qui comporte des variables définies comme des chaînes de caractères (elles peuvent inclure des chiffres, des lettres ou des symboles). Les fonctions numériques, comme les additions et les soustractions, ne peuvent pas être utilisées avec les variables en chaîne. Pour pouvoir effectuer une analyse quantitative des données, vous devez donc convertir les valeurs d'un format en chaîne à des valeurs numériques. Le tableau des observations d'oiseaux (tableau 7) comporte une colonne indiquant s'il s'agit d'un oiseau juvénile. Pour une analyse quantitative des données, vous devrez convertir toutes les valeurs en chaîne « non » à la valeur numérique 0 ainsi que les valeurs en chaîne « oui » à la valeur numérique 1. Laissez telle quelle la colonne originale JUVENILE en guise de référence et créez une nouvelle colonne avec les valeurs numériques. La colonne originale ne sert qu'à vérifier la transformation de la nouvelle colonne et pourra ensuite être supprimée lorsque la transformation sera jugée complète et exacte. Dans cet exemple, la nouvelle colonne JUVENILE_NUM contient les valeurs numériques qui correspondent aux valeurs en chaîne de la colonne JUVENILE (l'exercice du conseil n° 6 illustre cet exemple).

Les chiffres peuvent être formatés de différentes façons, surtout pour les données financières. Par exemple, les valeurs négatives peuvent être représentées avec un trait d'union, placées à l'intérieur de parenthèses ou même surlignées en rouge. Ces valeurs négatives ne pourront pas toutes être lues par un ordinateur, notamment quand il s'agit de couleur. En nettoyant les données, choisissez et appliquez une approche claire et uniforme pour le formatage de toutes les valeurs négatives. Le signe négatif est le choix le plus courant.

**Exercice du conseil n° 6: chiffres et signes**

Créez une nouvelle colonne intitulée JUVENILE_NUM dans le tableau 7. Inscrivez une valeur de 0 dans cette nouvelle colonne quand « non » apparaît dans la colonne JUVENILE. Inscrivez une valeur de 1 dans cette même colonne quand « oui » apparaît dans la colonne JUVENILE.

| ID | OISEAU | LIEU | JUVENILE |
|---|---|---|---|
| 1 | rouge-gorge | ch Québec | non |
| 2 | hirondelle | chemin Cork | oui |
| 3 | corbeau | ch March | non |
| 4 | pigeon | chemin Victoria | non |
| 5 | corbeau | ch Steffler | non |
| 6 | corbeau | ch Extra | oui |
| 7 | rouge-gorge | ch Doyle | oui |
| 8 | rouge-gorge | chemin Oxford | non |
| 9 | corbeau | ch Dublin | non |
| 10 | pigeon | chemin First | non |
| 11 | pigeon | chemin Sixth | oui |
| 12 | pigeon | Ch Church | non |
| 13 | hirondelle | ch Dulac | oui |

Tableau 7. Les données dans un format en chaîne.

Voir le solutionnaire pour les réponses.

## Conseil n° 7: date et heure

Il y a de nombreuses façons de formater les dates d'un jeu de données. Elles sont parfois formatées en chaîne. Mais si les données de dates sont nécessaires à des fins d'analyse, le type de champ devrait être changé de « chaîne » à « date » pour que les dates puissent être reconnues dans l'outil d'analyse choisi. Pour les valeurs liées à l'heure, vous devrez choisir une convention et l'appliquer uniformément à travers le jeu de données. Par exemple, vous pouvez choisir d'utiliser l'horloge de 12 ou de 24 heures pour préciser l'heure dans votre jeu de données. Peu importe ce que vous choisissez, vous devez veiller à ce que l'application se fasse uniformément. Vous pouvez aussi avoir à modifier un format pour vous assurer que tous les formats de date et d'heure sont uniformes.

# Conseil n° 8: fusionner et fractionner les colonnes


Après un examen approfondi d'un nouveau jeu de données, il est possible de soit (1) fusionner deux colonnes ou plus en une seule ou (2) de fractionner une colonne en deux ou plus. Conservez les colonnes originales qui ont été utilisées pour fusionner ou fractionner les colonnes. Utilisez ensuite les colonnes originales pour vérifier la transformation de la nouvelle colonne et supprimez l'originale lorsque la transformation est confirmée comme étant exacte. Par exemple, vous pouvez vouloir fractionner une colonne qui contient le nom complet en deux colonnes; une pour le prénom et l'autre, le nom de famille (tableau 8). Ou vous pouvez vouloir fractionner la colonne qui comporte l'adresse en colonnes distinctes pour la rue, la ville, la région et le code postal. L'inverse peut aussi s'appliquer; vous pouvez vouloir fusionner les colonnes du prénom et du nom de famille en une seule ou combiner les colonnes pour l'adresse.

**Exercice du conseil n° 8: fusionner et fractionner les colonnes**

Dans le tableau 8, fractionnez la colonne du NOM en deux; une pour le prénom et l'autre, pour le nom de famille.

CONSEIL : Avec Excel, cherchez les fonctions pour combiner le texte de plusieurs cellules en une seule et pour fractionner le texte en différentes colonnes.

| ID | AGE | NOM | COURRIEL |
|---|---|---|---|
| 1 | 17 | James Smith | jsmith@gmail.com |
| 2 | 21 | Michael Smith | msmith@gmail.com |
| 3 | 18 | Robert Smith | smithr@aol.com |
| 4 | 19 | Maria Garcia | mgarcia@hotmail.com |
| 8 | 18 | David Smith | davidsmith@gmail.com |
| 12 | 21 | Maria Rodriguez | mariar@gmail.com |
| 13 | 18 | Mary Smith | marysmith@gmail.com |
| 14 | 19 | Maria Hernandez | hernandez@outlook.com |
| 15 | 18 | Maria Martinez | mmartinez@mail.com |
| 16 | 21 | James Johnson | james@gmail.com |
| 17 | 18 | Lee Hartman | hartman@mail.com |
| 19 | 19 | Patricia Smith | smithp@mail.ca |
| 20 | 18 | Ben Smith | bensmith@mail.com |

Tableau 8. Des données dont les colonnes peuvent être fractionnées.

Voir le solutionnaire pour les réponses.

## Conseil n° 9: données d'un sous-ensemble


Certains fichiers de données peuvent parfois contenir des informations qui ne sont pas nécessaires à l'analyse; vous pourriez donc vouloir créer un nouveau fichier qui comporte uniquement les variables ou les observations qui vous intéressent. Vous aurez donc à faire une élimination sélective des colonnes ou des rangées superflues. Dans cet exemple, le chercheur a retiré la colonne JUVENILE (tableau 9). Vous pouvez également avoir besoin d'analyser uniquement certaines observations dans le fichier, ce qui vous permet de supprimer certaines rangées du jeu de données. Dans ce tableau, toutes les observations en lien avec les hirondelles seront éliminées. L'avantage de ce type de nettoyage, c'est qu'il limite la grosseur du fichier de données, permettant ainsi aux logiciels de fonctionner plus efficacement.

### Exercice du conseil n° 9: données d'un sous-ensemble

Créez un sous-ensemble de données dans le tableau 9 pour inclure uniquement les observations de juvéniles (JUVENILE =1).

CONSEIL: Comme toujours, il est important de conserver une copie de vos données originales.

| ID | OISEAU | LIEU | JUVENILE |
|---|---|---|---|
| 1 | rouge-gorge | chemin Québec | 1 |
| 2 | hirondelle | chemin Cork | 0 |
| 3 | corbeau | chemin March | 1 |
| 4 | pigeon | chemin Victoria | 1 |
| 5 | corbeau | chemin Steffler | 1 |
| 6 | corbeau | chemin Extra | 0 |
| 7 | rouge-gorge | chemin Doyle | 0 |
| 8 | rouge-gorge | chemin Oxford | 1 |
| 9 | corbeau | chemin Dublin | 1 |
| 10 | pigeon | chemin First | 1 |
| 11 | pigeon | chemin Sixth | 0 |
| 12 | pigeon | chemin Church | 1 |
| 13 | hirondelle | chemin Dulac | 0 |

Tableau 9. Données d'un sous-ensemble.

Voir le solutionnaire pour les réponses.

Le nettoyage des données est une action importante qui mise sur l'élimination d'incohérences et d'erreurs qui peuvent avoir un impact sur la précision des modèles. Le processus de nettoyage des données donne aussi l'occasion d'examiner les données de plus près afin de déterminer si des modifications sont nécessaires, si les données doivent être codées différemment ou s'il y a lieu d'ajouter des données supplémentaires.

## 4. Enrichir les données

Un jeu de données peut parfois contenir des informations manquantes, ce qui nuit à la capacité de bien répondre à la question de recherche. Vous pourrez donc avoir à chercher d'autres jeux de données pour les fusionner à vos données. Il peut s'agir d'ajouter des données géographiques, telles qu'un code postal ou des coordonnées de longitude et de latitude, ou de données démographiques, telles que le revenu, l'état civil, l'éducation, l'âge ou le nombre d'enfants. L'enrichissement des données permet d'obtenir des réponses plus complètes à vos questions de recherche.

Il est également important de vérifier la qualité et l'uniformité des données dans le jeu de données. Examinons maintenant la validation des données pour que vos modèles puissent fournir des résultats plus précis.

## 5. Valider les données

La validation des données est essentielle pour assurer la propreté, l'exactitude et l'utilité des données. Rappelez-vous l'adage de Fuechsel: « *garbage in, garbage out* ». Si des données inexactes sont intégrées à une analyse statistique, les réponses qui en découlent seront elles aussi inexactes. Un logiciel n'est pas doté de raison et ne peut traiter que les données qu'il reçoit, qu'elles soient bonnes ou mauvaises. Bien que la validation des données soit fastidieuse, elle permet d'optimiser la capacité des données à répondre aux questions de recherche posées. Voici quelques vérifications couramment utilisées pour valider des données :

1. Vérifier les types de données des colonnes et les données sous-jacentes pour s'assurer qu'ils sont bien ce qu'ils sont censés être. Par exemple, une variable de date peut avoir besoin d'être convertie d'un format en chaîne à un format de date. En cas de doute, mieux vaut convertir la valeur à un format en chaîne; elle pourra être modifiée plus tard, au besoin.
2. Examiner l'étendue et l'exactitude des données en en passant en revue les principales fonctions d'agrégation telles que la somme, le décompte, le minimum, le maximum, la moyenne ou d'autres opérations connexes. Cette étape est particulièrement importante dans le contexte de l'analyse des données. Par exemple, Statistiques Canada peut attribuer des codes aux valeurs manquantes pour l'âge en utilisant des chiffres bien au-delà de l'échelle d'âge de la vie humaine (p. ex., en utilisant un chiffre comme 999). Si ces valeurs sont incluses par inadvertance dans l'analyse (en raison des « valeurs manquantes » qui n'ont pas été clairement identifiées), tout résultat lié à l'âge sera erroné. Faire le calcul et la révision de la moyenne, du minimum, du maximum, etc. aidera à identifier et à éviter de telles erreurs.
3. S'assurer que les variables sont normalisées. Par exemple, en enregistrant les coordonnées de longitude et de latitude pour les emplacements en Amérique du Nord, vérifiez que les coordonnées de latitude soient positives et que celles de longitude soient négatives pour éviter de désigner par erreur des endroits de

l'autre côté du globe.

La validation des données est importante pour assurer la qualité et l'uniformité. Une fois toutes les questions de recherche répondues, les bonnes pratiques prônent le partage des données nettoyées avec d'autres équipes de recherche, conformément aux ententes de confidentialité ou tout autre type de restrictions. Examinons maintenant l'étape de la publication des données, ce qui leur permet d'être partagées avec d'autres chercheuses et chercheurs.

## 6. Publier les données

Après tout l'effort déployé au nettoyage et à la validation des données, ainsi qu'à l'examen approfondi de votre question de recherche, une des meilleures pratiques en GDR est de rendre vos données disponibles à d'autres qui souhaitent les utiliser à leur tour. Cet objectif est incarné dans les **principes FAIR**, abordés ailleurs dans ce manuel; ces principes visent à rendre les données faciles à trouver, accessibles, **interopérables** et réutilisables. La publication des données permet de réaliser cet objectif.

Les logiciels propriétaires peuvent être utiles pour la collecte, la gestion et l'analyse des données, mais les données devraient être enregistrées dans des formats ouverts à l'étape de la publication. Généralement, des fichiers texte sont utilisés. Pour les tableurs et feuilles de calculs simples, le meilleur format pour la conversion des données est le **CSV** (*comma separated values*) tandis que le langage XML est mieux adapté aux structures de données plus complexes. Ainsi, les fichiers sont protégés contre l'obsolescence rapide des formats et un accès plus universel aux données est assuré pour d'autres équipes de recherche. Pour en savoir plus, consultez le chapitre 9, « Un aperçu du fascinant monde des formats de fichiers et des métadonnées. »

Si vos données impliquent des êtres humains ou des renseignements confidentiels, vous pourrez avoir à anonymiser ou à dépersonnaliser vos données (le sujet est abordé plus en détail dans le chapitre 13, « Les données sensibles » ). Gardez à l'esprit que la suppression des références explicites à des personnes ne garantit pas qu'elles ne pourront pas être identifiées. Si la divulgation non voulue de renseignements personnels est impossible, vous devrez peut-être publier un sous-ensemble des données qui, lui, sera sans danger pour les personnes participantes.

Pour que d'autres équipes de recherche puissent utiliser les données, ajoutez de la documentation et des **métadonnées**, incluant de la documentation au niveau du projet, des fichiers de données et des éléments de données. Un dictionnaire des données établit le nommage, la définition et les attributs des éléments d'un jeu de données; le sujet est discuté au chapitre 10. Vous devriez aussi documenter les scripts et les méthodes qui ont été développés pour l'analyse des données.

# Logiciel de nettoyage des données

**OpenRefine** (https://openrefine.org/ (https://openrefine.org/) ) est un puissant outil de manipulation des données qui nettoie, réorganise et fait l'édition en lot des données qui manquent d'ordre ou de structure. Il fonctionne mieux avec des données en **format tabulaire** simple, notamment les feuilles de calcul dont les valeurs sont séparées par des virgules (CSV) ou des **tabulations** (TSV). OpenRefine est aussi simple à utiliser que Excel et dispose de puissantes fonctions de bases de données comme Microsoft Access. Il s'agit d'une application de bureau qui utilise un navigateur Web comme interface graphique. Tout traitement des données se fait localement à même votre ordinateur. En utilisant OpenRefine pour nettoyer et transformer les données, il est possible de filtrer par facette, regrouper, modifier des cellules, faire des concordances et utiliser des services Web plus approfondis pour convertir un jeu de données en format plus structuré. Ce **logiciel ouvert** est gratuit et le code source est librement accessible, de même que les modifications apportées par d'autres. Il y a d'autres outils de nettoyage des données disponibles, mais ils sont souvent coûteux. En plus, OpenRefine est largement utilisé dans le domaine de la GDR. Si vous choisissez d'autres logiciels de nettoyage des données, vérifiez toujours si vos données restent sur votre ordinateur ou si elles sont envoyées ailleurs pour être traitées.

**Exercice: nettoyer et préparer les données pour l'analyse avec OpenRefine**

Le tutoriel « Nettoyer ses données avec OpenRefine (https://programminghistorian.org/fr/lecons/nettoyer-ses-donnees-avec-openrefine) » vous permet de télécharger un jeu de données du Powerhouse Museum composé de métadonnées détaillées sur les objets de la collection dont le titre, la description, les catégories auxquelles les objets appartiennent, des informations sur la provenance et un lien permanent vers l'objet sur le site Web du musée. Vous effectuerez plusieurs tâches de nettoyage des données.

# Conclusion

Nous avons examiné les six actions principales de nettoyage et de préparation des données, soit: découvrir, structurer, nettoyer, enrichir, valider et publier. En appliquant ces pratiques importantes en GDR, vos données seront plus complètes, documentées et accessibles, aussi bien à vous qu'à d'autres chercheuses et

chercheurs. Vous répondrez aux exigences des publications savantes et/ou des organismes subventionnaires, vous rehausserez votre profil de chercheuse ou chercheur et vous répondrez aux attentes toujours croissantes de la communauté de recherche en matière de partage des données. Les pratiques de GDR, telles que le nettoyage des données, sont essentielles pour assurer des recherches précises et de grande qualité.

**Éléments clés à retenir**

- Le nettoyage des données est une tâche importante qui améliore l'exactitude et la qualité des données en amont de l'analyse des données.
- Les six tâches principales de nettoyage des données sont : découvrir, structurer, nettoyer, enrichir, valider et publier.
- OpenRefine est un puissant outil de manipulation des données qui nettoie, réorganise et fait l'édition en lot des données qui manquent d'ordre ou de structure.

# Bibliographie

Behrens, J. T. (1997). Principles and procedures of exploratory data analysis. *Psychological Methods, 2*(2), 131.

Lidwell, W., Holden, K. et Butler, J. (2010). *Universal principles of design, revised and updated: 125 ways to enhance usability, influence perception, increase appeal, make better design decisions, and teach through design*. Rockport Publishers.

## À propos de l'auteur

### Lucia Costanzo

Lucia Costanzo est la bibliothécaire en GDR à l'Université de Guelph. Récemment, elle a terminé un prêt de service auprès de l'Alliance de recherche numérique du Canada (l'Alliance) en tant que coordonnatrice de l'évaluation, de la recherche et de l'intelligence. Dans le cadre de ce rôle, Mme Costanzo a coordonné les activités du Groupe d'experts sur la recherche et l'intelligence. Ces activités comprenaient informer et conseiller l'équipe de GDR et la direction de l'Alliance relativement aux nouveaux développements et aux nouvelles directions et ce, tant à l'échelle nationale qu'internationale, au chapitre de la GDR et des

écosystèmes plus vastes d'infrastructure de recherche numérique. Avant sa période de prêt de service, elle a travaillé pendant plus de 20 ans à l'Université de Guelph à appuyer, rendre accessible et contribuer au processus d'apprentissage et de recherche sur le campus. Courriel : lcostanz@uoguelph.ca (denied:about:blank) | ORCID : 0000-0003-4785-660X (https://orcid.org/0000-0003-4785-660X)

8.

# NOUVELLES AVENTURES EN NETTOYAGE DES DONNÉES: TRAVAILLER AVEC DES DONNÉES DANS EXCEL ET R

Dr. Rong Luo et Berenica Vejvoda

**Objectifs d'apprentissage**

À la fin de ce chapitre, vous pourrez :

1. Expliquer les procédures générales pour se préparer au nettoyage des données.
2. Effectuer les tâches courantes de nettoyage des données à l'aide d'Excel.
3. Importer des données et effectuer les tâches de base de nettoyage des données à l'aide du langage de programmation R.

## Introduction

Le **nettoyage des données** est une partie essentielle du processus de recherche. Au cours du chapitre précédent, quelques tâches de bases courantes de nettoyage des données vous ont été présentées. Dans ce chapitre-ci, nous examinerons plus en profondeur l'exploration, la manipulation et le nettoyage des données en utilisant quelques outils de recherche flexibles et polyvalents. Dans certains cas, ces outils sont les mêmes qu'utilisent les chercheuses et chercheurs pour l'analyse de leurs données; il est donc utile que les personnes responsables de la curation et de la gestion des données les connaissent.

# Les procédures générales pour se préparer au nettoyage des données

Si vous ne faites aucune préparation avant le processus de nettoyage des données, vous prenez le risque de rencontrer de graves problèmes, notamment la perte de vos données. Au cours de cette section, nous discuterons des étapes générales qui devraient être appliquées avant le processus de nettoyage des données.

## Faire une copie de sauvegarde

Les pratiques de **gestion des données de recherche** (GDR) recommandent la création d'une sauvegarde sécurisée de vos données pour assurer que les données originales puissent toujours être restaurées en cas de modifications erronées survenues au cours du processus de nettoyage. Cette copie de sauvegarde des données originales ne devrait en aucun cas être modifiée. Vous devriez aussi tenir un registre / journal de tous les changements effectués. Vous seriez surpris d'apprendre combien de chercheuses et chercheurs créent des erreurs dans leurs données originales en tentant de les « améliorer ». Si une personne a besoin d'un accès aux données originales, vous devriez envoyer ou partager seulement une copie des données, sinon permettre l'accès aux données originales en lecture seule.

## Comprendre les données

La première étape du nettoyage des données est de comprendre les données à nettoyer. Pour les comprendre, il faut commencer par faire une exploration de base des données (ou une **analyse exploratoire des données**) pour se faire une idée des problèmes qui pourraient exister à l'intérieur des données. Vérifiez les valeurs des données par rapport à leur définition dans le fichier de **métadonnées** ou la documentation pour déceler des valeurs impossibles ou qui sortent des limites (p. ex., un âge négatif ou de plus de 200 ans). Assurez-vous d'avoir des noms pratiques pour les colonnes de données et de bien les comprendre. Vérifiez les délimiteurs qui séparent les valeurs dans vos fichiers texte et assurez-vous que les valeurs de vos données n'intègrent pas le délimiteur lui-même. Si vos observations ne sont pas numérotées, vous devriez ajouter un numéro d'enregistrement unique aux observations individuelles dans votre jeu de données pour que vous puissiez plus facilement repérer les enregistrements problématiques en vous rapportant à leur numéro.

## Planifier le processus de nettoyage

Le nettoyage des données doit être fait de façon systématique pour garantir que toutes les données soient nettoyées à l'aide des mêmes procédures. Ainsi, l'intégrité des données est assurée et les données peuvent être

plus facilement traitées pendant l'analyse. Pour créer un plan de nettoyage d'un champ particulier dans un jeu de données, posez-vous les trois questions suivantes :

- Quelles sont les données que vous nettoyez?
- Comment allez-vous identifier un problème dans le jeu de données à nettoyer?
- Comment le champ devrait-il être nettoyé?

## Choisir les bons outils

Une des étapes les plus importantes du nettoyage des données est de choisir le bon outil en fonction d'un objectif précis. Le chapitre précédent a présenté **OpenRefine**, un outil spécialisé et pratique pour le nettoyage des données. Ici, nous discuterons de deux outils logiciels puissants et polyvalents – Excel et R – et nous soulignerons quelques-unes des caractéristiques de nettoyage de chacun d'eux.

# Les outils de nettoyage des données

L'outil de nettoyage des données que vous choisirez dépendra de différents facteurs, dont votre environnement informatique, votre expertise en matière de programmation et des exigences en lien avec la préparation de vos données. Il existe une grande variété de choix de logiciels et de méthodes pour le nettoyage et la transformation des données. Nous examinerons Excel et Google Sheets ainsi que le langage de programmation R.

## Microsoft Excel/Google Sheets

Excel et Google Sheets sont d'excellents outils de nettoyage des données et contiennent une variété de fonctionnalités et de caractéristiques intégrées pour automatiser le nettoyage de vos données. Excel est largement disponible en programme de bureau tant pour Windows et MacOs alors que Google Sheets est disponible en ligne. Ils se ressemblent et sont faciles à apprendre, utiliser et comprendre. Ils peuvent tous deux importer et exporter le format très courant de **fichier de données CSV** et autres formats courants de tableur. Lors de l'exportation, vérifiez que les noms des colonnes de données exportées sont utilisables puisque certains progiciels statistiques peuvent avoir de la difficulté à traiter les noms de colonnes qui contiennent des espaces ou des caractères spéciaux. Les techniques courantes de nettoyage des données utilisées dans Excel et Google Sheets pour l'édition et la manipulation sont résumées dans le tableau ci-dessous.

**Tableau 1. Tableau des fonctions.**

| Fonction | Description |
|---|---|
| = CONCATENER (ou CONCAT dépendamment de votre version Excel) | Permet de joindre plusieurs colonnes. |
| = SUPPRESPACE | Supprime tous les espaces d'une chaîne de texte, à l'exception des espaces simples entre les mots. |
| = GAUCHE | Renvoie le ou les premier(s) caractère(s) d'une chaîne de texte en fonction du nombre de caractères que vous spécifiez. |
| = DROITE | Renvoie le ou les dernier(s) caractère(s) d'une chaîne de texte en fonction du nombre de caractères que vous spécifiez. |
| = STXT | Renvoie un nombre donné de caractères d'une chaîne de texte, à partir d'une position que vous spécifiez, en fonction du nombre de caractères que vous spécifiez. |
| = MINUSCULE | Convertit toutes les lettres d'une chaîne de texte en lettres minuscules. |
| = MAJUSCULE | Convertit toutes les lettres d'une chaîne de texte en lettres majuscules. |
| = NOMPROPRE | Convertit une chaîne de texte à un format de nom propre pour que la première lettre de chaque mot soit en majuscule et les lettres qui suivent en minuscules. |
| = VALEURNOMBRE | Convertit une chaîne de texte en nombre. |
| = TEXTE | Convertit un nombre au format texte. |
| = SUBSTITUE | Remplace un texte précis dans une chaîne de texte. |
| = REMPLACER | Remplace une partie d'une chaîne de texte, selon la position et le nombre de caractères précisés, avec une chaîne de texte différente. |
| = EPURAGE | Supprime tous les caractères non imprimables d'une chaîne de texte. |
| = DATE | Renvoie le nombre qui représente la date dans le code date-heure de Microsoft Excel. |
| = ARRONDI | Arrondi le nombre d'une cellule précise à un nombre spécifié de chiffres. |
| = TROUVE | Renvoie la position initiale d'une chaîne de texte dans une autre chaîne de texte. TROUVE est sensible à la casse. |
| = CHERCHE | Renvoie le numéro de la position initiale d'un caractère spécifique ou d'une chaîne de texte dans une autre chaîne de texte, en lisant de gauche à droite (non sensible à la casse). |

Pour comprendre certaines de ces fonctions, nous examinerons une variété d'erreurs courantes qui surviennent lors de l'importation des données incluant des sauts de lignes à la mauvaise place, des espaces de trop ou aucun espace entre les mots, des majuscules à la mauvaise place ou toutes les lettres en majuscules/minuscules, des valeurs de données mal formatées et des caractères non imprimables.

| | A | B | C |
|---|---|---|---|
| 1 | **CONCATENER et SUPPRESPACE** | | |
| 2 | | | |
| 3 | **Données importées** | **Formule utilisée** | **Résultats** |
| 4 | | =CONCATENER(A5, A6, A7) | Université deWindsor |
| 5 | Université | =SUPPRESPACE(CONCATENER(A5, A6, A7) | Université deWindsor |
| 6 | de | =CONCATENER(SUPPRESPACE(A5), SUPPRESPACE(A6), SUPPRESPACE(A7)) | UniversitédeWindsor |
| 7 | Windsor | =CONCATENER(SUPPRESPACE(A5), " ",SUPPRESPACE(A6), " ",SUPPRESPACE(A7)) | Université de Windsor |

Figure 1. Les fonctions CONCATENER et SUPPRESPACE avec le contenu original à côté du contenu nettoyé.

La figure 1 illustre des combinaisons de CONCATENER ET SUPPRESPACE imbriquées de différentes manières pour trouver la meilleure configuration de sortie selon la façon dont vous voulez que le texte apparaisse[1]. Il s'agit d'un exemple sur la façon de générer une simple ligne de texte à partir du contenu de trois rangées en imbriquant deux fonctions Excel. CONCATENER rassemble les trois cellules en une seule, mais elle n'agit pas sur les espaces supplémentaires que vous voyez dans le texte. SUPPRESPACE supprime tous les espaces à l'exception d'un seul espace entre les mots, mais la fonction ne peut pas ajouter d'autres espaces nécessaires. Nous avons donc besoin d'ajouter des guillemets pour permettre à Excel d'ajouter les espaces nécessaires entre les mots.

| | A | B | C |
|---|---|---|---|
| 8 | **GAUCHE, DROITE, STXT** | | |
| 9 | | | |
| 10 | **Données importées** | **Formule utilisée** | **Résultats** |
| 11 | BUS256XD | =STXT(A11,4,3) | 256 |
| 12 | DRT578XC | =STXT(A12,4,3) | 578 |
| 13 | SACR1373 | =DROITE(A13,4) | 1373 |
| 14 | KINE5301 | =GAUCHE(A14,4) | KINE |

Figure 2. Les fonctions GAUCHE, DROITE et STXT avec le contenu original à côté du contenu nettoyé.

Les fonctions GAUCHE, DROITE et STXT de la figure 2 montrent de quelle façon des données peuvent être traitées selon l'endroit dans la chaîne où se trouve le texte ou les numéros à extraire.

Les rangées 11 et 12 illustrent la façon d'utiliser la fonction STXT pour extraire des numéros à l'intérieur d'une chaîne de texte. La fonction STXT utilise trois **arguments** : une référence à la chaîne avec laquelle vous travaillez, la position du premier caractère à extraire et le nombre de caractères à extraire. Alors STXT(A11,4,3) regarde d'abord le contenu de la cellule A11 et trouve la chaîne « BUS256XD », ensuite elle

1. Les fichiers originaux des feuilles de calcul pour chaque figure de ce chapitre sont disponibles (en anglais uniquement) dans un format accessible dans Borealis (https://borealisdata.ca/dataverse/furtheradventures).

renvoie les 3 caractères en partant du quatrième caractère : 256. Les données du C11 et C12 sont le résultat de la fonction STXT dans les rangées 11 et 12.

Les fonctions GAUCHE et DROITE ne nécessitent que deux arguments : la chaîne et le point de départ. Ces fonctions renvoient ensuite le reste de la chaîne, en partant soit de la gauche ou de la droite. C13 et C14 illustrent des portions de cellules extraites de A13 et A14 en utilisant les fonctions DROITE et GAUCHE.

| | A | B | C |
|---|---|---|---|
| 15 | | **TROUVE et CHERCHE** | |
| 16 | | | |
| 17 | **Données importées** | **Formule utilisée** | **Résultats** |
| 18 | INtrOducTion | =TROUVE("o",A18) | 11 |
| 19 | aux | =TROUVE("u",A19) | 2 |
| 20 | ORDinateurs | =TROUVE("o",A20) | #VALEUR! |
| 21 | INtrOducTion | =CHERCHE("o",A21) | 5 |
| 22 | aux | =CHERCHE("u",A22) | 2 |
| 23 | ORDinateurs | =CHERCHE("o",A23) | 1 |
| 24 | ORDinateurs | =CHERCHE("x",A24) | #VALEUR! |

Figure 3. Les fonctions TROUVE et CHERCHE avec le contenu original à côté du contenu nettoyé.

La figure 3 illustre la différence entre les fonctions TROUVE et CHERCHE. Dans Excel, TROUVE est utilisé pour renvoyer la position d'un caractère ou d'une sous-chaîne spécifique à l'intérieur d'une chaîne de texte; la fonction respecte la casse. La fonction CHERCHE revoie également la position d'un caractère ou d'une sous-chaîne à l'intérieur d'une chaîne de texte. Contrairement à TROUVE, la fonction CHERCHE ne respecte pas la casse. Les deux fonctions renvoient le message d'erreur #VALEUR! si le caractère ou la sous-chaîne spécifique n'apparaît pas dans le texte.

| | A | B | C |
|---|---|---|---|
| 25 | | **MAJUSCULE, MINUSCULE, NOMPROPRE** | |
| 26 | | | |
| 27 | **Données importées** | **Formule utilisée** | **Résultats** |
| 28 | INtrOducTion | =MAJUSCULE(A28) | INTRODUCTION |
| 29 | INtrOducTion | =MINUSCULE(A29) | introduction |
| 30 | anne tremblay | =NOMPROPRE(A30) | Anne Tremblay |

Figure 4. Les fonctions MAJUSCULE, MINUSCULE et NOMPROPRE avec le contenu original à côté du contenu nettoyé.

La figure 4 démontre de quelle façon les fonctions MAJUSCULE, MINUSCULE et NOMPROPRE sont utilisées pour ajuster les données. La fonction MAJUSCULE modifie tout le texte en lettres majuscules. La fonction MINUSCULE modifie tout le texte en minuscule. Et la fonction NOMPROPRE modifie la

première lettre de chaque mot en majuscule avec toutes les autres en minuscules, ce qui est utile avec des noms propres.

| | A | B | C |
|---|---|---|---|
| 31 | **VALEURNOMBRE et TEXTE** | | |
| 32 | | | |
| 33 | **Données importées** | **Formule utilisée** | **Résultats** |
| 34 | 12345 | =VALEURNOMBRE(A34) | 12345 |
| 35 | ABCD | =VALEURNOMBRE(A35) | #VALEUR! |
| 36 | 12345 | =TEXTE(A36,"00000") | 12345 |
| 37 | 12345 | =TEXTE(A37,"0000000") | 0012345 |

Figure 5. Les fonctions VALEURNOMBRE et TEXTE avec le contenu original à côté du contenu nettoyé.

Excel fait l'alignement des chaînes de caractères d'une colonne en fonction de la façon dont elles sont stockées : le texte (y compris des numéros stockés en tant que texte) est aligné à gauche tandis que les numéros sont alignés à droite. Dans la figure 5, la fonction VALEURNOMBRE convertit le texte qui apparaît dans un format reconnu (tel que des formats de numéro, de date ou d'heure) en une valeur numérique. Si un texte ne se retrouve pas dans l'un de ces formats, VALEURNOMBRE renvoie le message d'erreur #VALEUR!. La fonction TEXTE permet de modifier l'affichage d'un nombre en appliquant des codes de format, ce qui est utile dans les situations où vous souhaitez afficher des nombres dans un format plus lisible. Par contre, Excel considère désormais le numéro comme du texte, de sorte que l'exécution de calculs sur ces données risque de ne pas fonctionner ou de donner des résultats inattendus. Il est préférable de conserver la valeur originale dans une cellule, puis d'utiliser la fonction TEXTE pour créer une copie formatée du numéro dans une autre cellule.

| | A | B | C |
|---|---|---|---|
| 38 | **SUBSTITUE et REMPLACER** | | |
| 39 | | | |
| 40 | **Données importées** | **Formule utilisée** | **Résultats** |
| 41 | Taxe | =SUBSTITUE(A41,"t","l") | Taxe |
| 42 | bulle | =SUBSTITUE(A42,"l","t") | butte |
| 43 | bulle | =SUBSTITUE(A43,"l","b",2) | bulbe |
| 44 | Le chien | =SUBSTITUE(A44,"le","un") | Le chien |
| 45 | bulle | =REMPLACER(A45,3,2,"*") | bu*e |

Figure 6. Les fonctions SUBSTITUE et REMPLACER avec le contenu original à côté du contenu nettoyé.

La figure 6 illustre la façon dont la fonction SUBSTITUE remplace une ou plusieurs chaînes de texte avec une autre chaîne. Cette fonction est utile si vous voulez substituer une ancienne version de texte d'une chaîne avec une nouvelle version. Toutefois, elle ne respecte pas la casse. Par exemple, dans la cellule A41, la fonction ne remplacera pas « t » pour « T » dans « Taxe ». La fonction SUBSTITUE est différente de REMPLACER;

vous utilisez SUBSTITUE pour remplacer des caractères spécifiques, peu importe où ils se trouvent dans la chaîne de texte, tandis que vous utilisez REMPLACER pour remplacer tout caractère qui se trouve dans une position particulière d'une chaîne de texte.

| | A | B | C |
|---|---|---|---|
| 46 | | **EPURAGE** | |
| 47 | | | |
| 48 | **Données importées** | **Formule utilisée** | **Résultats** |
| 49 | □Gestion des données de recherche!! | =EPURAGE(A49) | Gestion des données de recherche |
| 50 | !!saut de ligne | =EPURAGE(A50) | !!saut de ligne |
| 51 | !!saut de ligne | =EPURAGE(A51, CAR(127),"") | saut de ligne |

Figure 7. La fonction EPURAGE avec le contenu original à côté du contenu nettoyé.

La fonction EPURAGE illustrée dans la figure 7 élimine du texte les caractères non imprimables tels que les retours chariot (↩) ou d'autres caractères de contrôle (https://fr.wikipedia.org/wiki/American_Standard_Code_for_Information_Interchange#Caract%C3%A8res_de_contr%C3%B4le) représentés par les 32 premiers codes ASCII 7 bits. Les données importées d'une variété de sources peuvent comporter des caractères non imprimables et la fonction EPURAGE aide à les éliminer d'une chaîne de texte. Dans Excel, un caractère non imprimable peut s'afficher comme un symbole de case (□). À noter que la fonction EPURAGE ne peut pas éliminer tous les caractères non imprimables (p. ex., le caractère d'effacement). Vous pouvez préciser un caractère ASCII en utilisant la fonction CAR de Excel et le numéro du code ASCII. Par exemple, CAR(127) correspond au code d'effacement. Pour éliminer un caractère non imprimable, vous pouvez simplement remplacer le caractère non imprimable à éliminer avec des guillemets vides ("").

Dans l'exercice ci-dessous, le caractère non imprimable CAR (19) dans la rangée 10 s'affiche comme « !! ».

**Exercice 1**

Utiliser les fonctions de Excel/Google Sheets pour générer les résultats de nettoyage dans la colonne B à partir des données importées de la colonne A.

| | A | B |
|---|---|---|
| 1 | **Données importées** | **Résultats** |
| 2 | L' | L'école de service social |
| 3 | école de | |
| 4 | service social | |
| 5 | | |
| 6 | 3 ou plus | 3 |
| 7 | | |
| 8 | à la | À La |
| 9 | | |
| 10 | !!Rapport mensuel!! | Rapport mensuel |
| 11 | | |
| 12 | Taxe | taxe |
| 13 | | |
| 14 | SACR-126 | 126 |
| 15 | | |
| 16 | 789 | 00789 |

Voir le solutionnaire pour les réponses.

## Le langage de programmation R

Bien que les tableurs comme Excel et Google Sheets fournissent des fonctions courantes qui peuvent faciliter le nettoyage des données, leur utilisation peut s'avérer difficile pour des jeux de données volumineux. De plus, si Excel et Google Sheets ne disposent pas déjà d'une fonction intégrée particulière, il faudra beaucoup de temps de programmation pour la construire. C'est ici que le programme R peut aider. R est l'un des progiciels statistiques les plus connus et accessibles pour le nettoyage des données. R est un langage de programmation entièrement fonctionnel ayant des fonctionnalités qui permettent de travailler avec des données et des statistiques. Il n'est pas nécessaire de maîtriser la programmation pour utiliser certaines de ses fonctions de base.

Les deux composantes les plus importantes du langage R sont les objets qui stockent les données et les fonctions qui manipulent les données. R utilise également une panoplie d'opérateurs comme +, -, *, / et <- pour effectuer des tâches simples.

Pour créer un **objet R**, choisissez un nom et utilisez ensuite le symbole plus-petit-que suivi du signe moins pour y sauvegarder des données. Cette combinaison ressemble à la tête d'une flèche : <-. Par exemple, vous

pouvez sauvegarder la donnée « 1 » dans un objet « a ». À chaque fois que R rencontre l'objet « a », il sera remplacé par la donnée « 1 » sauvegardée à l'intérieur, tel qu'illustré ici :

```
> a <- 1
```

R est doté de nombreuses fonctions que vous pouvez utiliser pour effectuer des tâches élaborées. Par exemple, vous pouvez arrondir un numéro avec la fonction *round*. L'utilisation d'une fonction est assez simple. Vous n'avez qu'à écrire le nom de la fonction accompagné de la donnée, mise entre parenthèses, sur laquelle vous voulez que la fonction opère.

```
> round (3.1415)
[1] 3
```

Les extensions (ou packages) R sont des collections de fonctions écrites par des programmeurs R. Pour que les extensions puissent bien fonctionner, vous pourriez avoir à en installer d'autres d'avance. En les installant, il est plus simple dès le départ de configurer « dependencies = TRUE ».

```
> install.packages("package name", dependencies = TRUE)
```

Commençons par télécharger et installer les logiciels nécessaires. R est disponible pour Windows, MacOS et Linux.

## 1. Installer R et RStudio

R peut être téléchargé ici : https://cran.rstudio.com/ (https://cran.rstudio.com/).

- Pour les utilisateurs Windows, https://cran.rstudio.com/bin/windows/base/ (https://cran.rstudio.com/bin/windows/base/);
- Pour les utilisateurs Mac, https://cran.rstudio.com/bin/macosx/ (https://cran.rstudio.com/bin/macosx/);
- Pour les utilisateurs Linux, https://cran.r-project.org/bin/linux/ (https://cran.r-project.org/bin/linux/).

Base-R est simplement un **outil de ligne de commande** : vous inscrivez des commandes à l'invite et voyez les résultats affichés à l'écran. RStudio, quant à lui, est un **environnement de développement intégré**, un ensemble d'outils dont un éditeur de script, une invite de commande, une fenêtre de résultats ainsi qu'un menu de commandes pour les fonctions R les plus couramment utilisées. Travailler en R, veut généralement dire utiliser R par le biais de RStudio. À noter qu'avant d'utiliser RStudio, vous devez d'abord installer R.

Téléchargez et installez RStudio Desktop qui est gratuit et disponible pour Windows, Mac et plusieurs

versions de Linux ici : https://posit.co/download/rstudio-desktop (https://posit.co/download/rstudio-desktop/).

## 2. Se familiariser avec RStudio

Avant d'importer des données, familiarisez-vous avec RStudio.

RStudio a quatre zones (voir figure 8) :

**Tableau 2. Rôles des zones de l'interface de RStudio.**

| Section | Objectif |
|---|---|
| Supérieure gauche | Cette section montre le texte en cours d'édition. Un texte R est un ensemble de commandes R et de commentaires. Ils sont généralement utilisés pour faire le suivi des commandes à exécuter et pour fournir des notes, par le biais de commentaires, qui expliquent le pourquoi des commandes. |
| Supérieure droite | L'onglet *Environment* énumère toutes les variables et les fonctions qui ont été définies et utilisées dans une session. |
| | L'onglet *History* énumère toutes les commandes inscrites dans la console R (dans la zone supérieure gauche de RStudio) |
| | L'onglet *Connections* peut aider à se connecter à des bases de données externes pour accéder à des données qui ne se retrouvent pas sur votre ordinateur. |
| Inférieure gauche | L'onglet *Console* affiche une invite de commande qui vous permet d'utiliser R de façon interactive, comme vous le feriez sans RStudio. |
| | L'onglet *Terminal* ouvre une interface système pour effectuer des fonctions plus avancées, telles que l'accès à un système distant. |
| Inférieure droite | L'onglet *Files* vous permet de faire le suivi, d'ouvrir et de sauvegarder les fichiers associés à votre projet R. |
| | L'onglet *Plots* illustre les graphiques en cours de tracé. |
| | L'onglet *Packages* permet de charger et d'installer des extensions qui ajoutent des fonctions R supplémentaires. |
| | L'onglet *Help* fournit des informations utiles sur certaines fonctions. |
| | L'onglet *Viewer* peut être utilisé pour afficher du contenu Web local et interagir avec lui. |

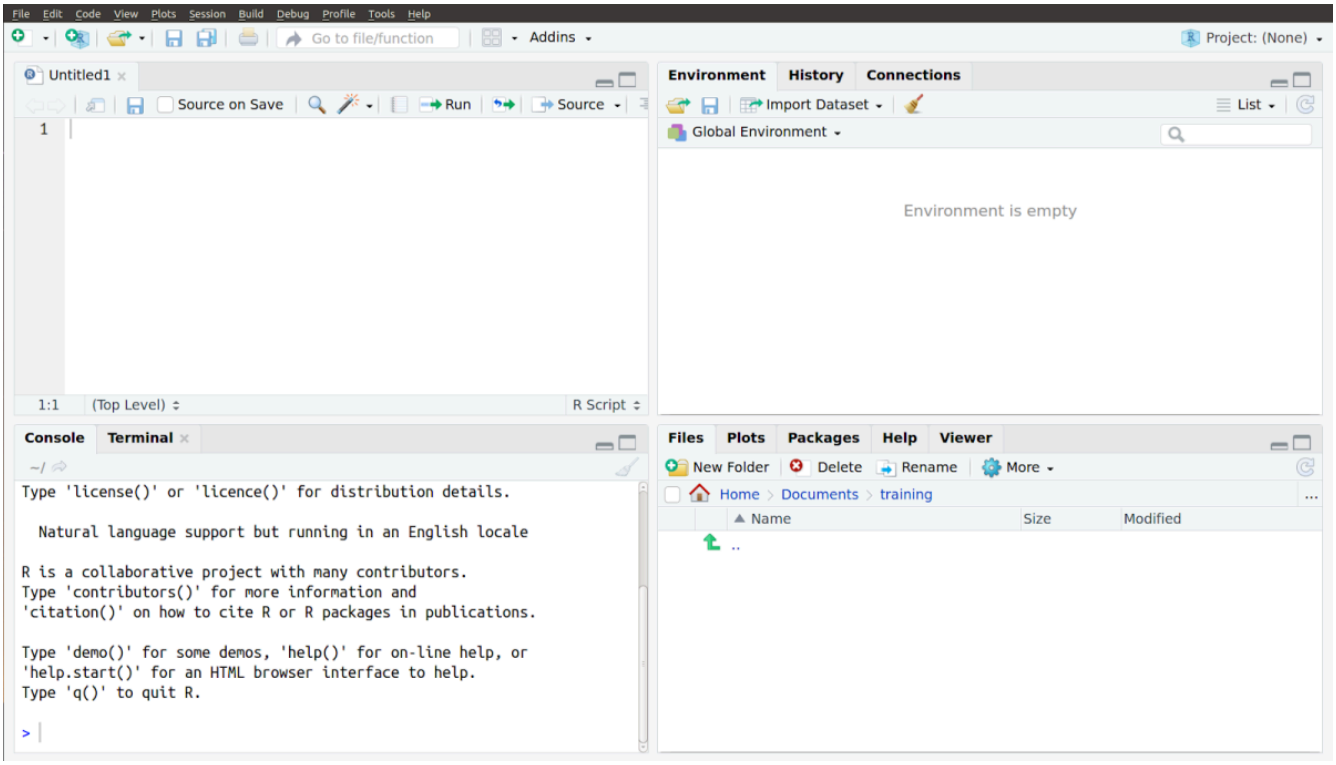



Figure 8. Les zones de l'interface de RStudio.

Vous pouvez exploiter l'invite de commande en inscrivant des commandes R dans la console, tout comme vous le feriez si vous travailliez sans RStudio. Vous pouvez ensuite voir les résultats affichés sous les onglets *History* et *Environment*. Une sauvegarde de votre travail n'est pas faite de façon automatique lorsque vous fermez R. Pour sauvegarder, vous pouvez copier et coller le contenu de votre console dans un fichier texte.

Par exemple, vous pouvez ouvrir RStudio et inscrire la commande suivante dans la console (le texte qui suit l'invite « > ») :

```
print("Bonjour")
```

R renvoie la sortie suivante :

```
[1] "Bonjour"
```

En plus de pouvoir travailler de façon interactive en inscrivant des commandes à l'invite, vous pouvez également créer des scripts R en utilisant l'éditeur de RStudio qui apparaît dans la zone supérieure droite. Les **scripts** sont des fichiers texte qui contiennent une séquence de commandes R pouvant être exécutées de façon consécutive. Vous pouvez sélectionner vos commandes dans les scripts et les exécuter une à la fois ou toutes ensembles. L'écriture et la sauvegarde de vos commandes pour le nettoyage des données dans des scripts vous permettent de mieux faire le suivi de votre travail. Aussi, vous pouvez plus facilement réexécuter le code plus

tard et sur de nouveaux jeux de données. Utiliser cette méthode pour faire le suivi de votre travail est une bonne pratique de GDR.

Pour ouvrir un nouveau script, sélectionnez-le à partir de l'icône supérieure gauche :

Avant d'importer votre jeu de données, vous devriez changer votre répertoire de travail pour le faire correspondre à l'emplacement de votre jeu de données. À partir de RStudio, utilisez le menu pour modifier votre répertoire de travail pour celui où vous avez sauvegardé l'échantillon de votre fichier de données. Dans le menu *Session*, sélectionnez > *Set Working Directory* > *Choose Directory*.

Vous pouvez également utiliser la fonction R setwd(), qui signifie *set working directory* ou « définir le répertoire de travail », dans la fenêtre de la console (ou de l'éditeur de script). Les barres obliques (/), plutôt que les barres obliques inverses, doivent être utilisées dans le chemin d'accès. Si vous avez sauvegardé les données dans « C:\data », il vous faudra alors entrer la commande ainsi :

```
setwd("C:/data")
```

## 3. Importer des données

Vous pouvez importer des données de différents formats en utilisant R. Les fichiers CSV sont couramment utilisés pour les données numériques. Les CSV ressemblent beaucoup aux fichiers Excel typiques, mais ce sont des fichiers texte avec des colonnes séparées par des virgules. Dans Excel, vous pouvez exporter ce type de données en utilisant « Enregistrer sous » ; il s'agit d'un format de préservation courant en gestion des données et il peut être lu par de nombreux programmes.

Dans les prochains exemples, nous utiliserons un échantillon de jeu de données – *sample.csv* – que vous trouverez sur Borealis (https://borealisdata.ca/dataverse/furtheradventures). Veuillez télécharger et sauvegarder le jeu de données dans un nouveau dossier sur votre ordinateur. Les fichiers SPSS et Excel nécessaires pour ces exemples sont également disponibles dans Borealis.

Pour charger le fichier CSV, créez d'abord un nouveau script dans l'éditeur de script. Inscrivez la commande suivante dans le script afin d'utiliser la commande R intégrée *read.csv*. Exécutez ensuite le script.

```
mydata_csv<-read.csv("sample.csv")
```

Le **délimiteur** par défaut de la fonction *read.csv*() est une virgule, mais si vous avez besoin de lire un fichier

qui utilise d'autres types de délimiteurs, vous pouvez le faire en fournissant l'argument *sep* à la fonction (p. ex., ajouter « sep = ';' » pour les fichiers qui utilisent le point-virgule comme séparateur).

```
> mydata_csv<-read.csv("sample.csv", sep=';')
```

Veuillez noter que « mydata_csv » dans la commande ci-dessus fait référence à l'objet (le *data frame* ou tableau de données) qui sera créé lorsque la fonction *read.csv* importera le fichier *sample.csv*. Imaginez le tableau de données *mydata_csv* comme étant le contenant utilisé par R pour conserver les données du fichier CSV.

Les commandes R suivent un certain modèle. Examinons celui-ci. Dans la commande ci-dessus – « mydata_csv<-read.csv("sample.csv", sep=';') » – *read.csv* est une fonction qui lit dans un fichier CSV et qui comporte deux paramètres. Le premier – « sample.csv » – indique à *read.csv* lequel des fichiers il doit lire, tandis que le second – « sep=';' » – lui indique que les données du fichier sont séparées par des points-virgules. Après que *read.csv* ait fait l'analyse du fichier, l'opérateur d'assignation – « <- » – affecte les données à *mydata_csv*, un objet (tableau de données) créé pour contenir les données. Vous pouvez maintenant utiliser et manipuler les données dans le tableau de données.

Dans R, <- est l'opérateur d'assignation le plus courant . Vous pouvez également utiliser le signe égal =. Pour plus d'informations, utilisez la commande d'aide :

```
> ?read.csv
```

Pour lire un fichier Excel, commencez par télécharger et installer l'extension *readxl*. Dans la console R, utilisez la commande suivante:

```
> install.packages("readxl")
```

Veuillez noter que certaines extensions peuvent dépendre d'extensions connexes afin de bien fonctionner. En utilisant des extensions existantes, les personnes qui font de la programmation peuvent économiser du temps dans la création de nouvelles fonctionnalités avec des fonctions qui ont déjà été mises en œuvre. Toutefois, il peut être difficile de savoir si une extension nécessite une autre extension pour bien fonctionner. Conséquemment, il est recommandé d'installer des extensions en ajoutant un énoncé des dépendances (« TRUE » indique à R que les dépendances devraient être incluses). En réglant le paramètre des dépendances à « TRUE », R va télécharger et installer toutes les extensions nécessaires à l'extension qui sera installée.

```
> install.packages("readxl", dependencies = TRUE)
```

Une fois le téléchargement et l'installation de l'extension terminés, utilisez la fonction *library()* pour charger l'extension *readxl*.

```
> library(readxl)
```

Notez que contrairement à la fonction *install.package*, il n'est pas nécessaire de mettre le nom de l'extension entre guillemets pour la fonction *library*.

Vous pouvez maintenant charger le fichier Excel avec la fonction *read_excel()* :

```
> mydata_excel <- read_excel("sample.xlsx")
```

Pour plus d'informations, utilisez la commande d'aide ***?read_excel***.

Les fichiers SAV du logiciel SPSS (Statistical Package for the Social Sciences) peuvent être lus avec R en utilisant l'extension *haven* qui ajoute des fonctions supplémentaires permettant d'importer des données d'autres outils statistiques.

Installez *haven* en utilisant la commande suivante :

```
> install.packages("haven", dependencies = TRUE)
```

Une fois le téléchargement et l'installation de l'extension terminés, utilisez la fonction *library()* pour charger l'extension :

```
> library(haven)
```

Vous pouvez maintenant charger le fichier SPSS avec la fonction *read_sav()* :

```
> mydata_spss <- read_sav("sample.sav")
```

Vous pouvez aussi importer des fichiers SAS et Stata. Pour plus d'informations, utilisez les commandes d'aide *?haven* ou *?read_sav*, ou visitez https://haven.tidyverse.org/ (https://haven.tidyverse.org/).

Les données peuvent aussi être téléchargées directement d'Internet en utilisant les mêmes fonctions que celles énumérées plus haut (à l'exception des fichiers Excel). Vous n'avez qu'à utiliser une adresse Web plutôt que le chemin d'accès.

```
> mydata_web <- read.csv(url("http://quelque.part.net/donnees/
echantillon.csv"))
```

Maintenant que les données ont été chargées dans R, vous pouvez commencer à effectuer des opérations et des analyses pour les vérifier et y déceler des problèmes potentiels.

## 4. Vérifier les données

R est un outil beaucoup plus flexible que Excel pour le travail avec des données. Nous allons aborder les fonctions R de base pour la vérification d'un jeu de données.

Assumons que le fichier texte suivant, stocké sous le nom *sample.csv*, est composé de huit rangées et de cinq colonnes.

```
1,4.1,3.5,setosa,A
2,14.9,3,set0sa,B
3,5,3.6,setosa,C
4,NA,3.9,setosa,A
5,5.8,2.7,virginica,A
6,7.1,3,virginica,B
7,6.3,NA,virginica,C
8,8,7,virginica,C
```

Prenons maintenant la commande suivante pour l'importation des données. Puisque le jeu de données ne contient pas d'en-tête (c'est-à-dire, la première rangée ne fait pas la liste des noms de colonnes), vous devez préciser « header=FALSE ». Si vous voulez définir manuellement le nom des colonnes, vous ajoutez l'argument *col.names*. Dans la commande ci-dessous, nous demandons à *read.csv* de définir le nom des colonnes à *ID*, *Longueur*, *Largeur*, *Espece* et *Site* (à noter que les accents sont éliminés des mots en français afin d'éviter l'ajout de caractères qui risquent d'être mal interprétés lors des manipulations). En utilisant *colClasses*, vous pouvez préciser le type de données (nombres, caractères, etc.) que vous vous attendez à retrouver dans les colonnes de données. Dans cet exemple, nous avons précisé à *read.csv* que la première colonne ainsi que les deux dernières (variables catégoriques) doivent être traitées comme des données de type facteur (*factor*) tandis que les deux colonnes du milieu comme des données numériques (*numeric*).

Inscrivez la commande suivante dans votre script et exécutez-la :

```
> mydata_csv <- read.csv("sample.csv", header = F, col.names =
c("ID","Longueur", "Largeur", "Espece", "Site"),
colClasses=c("factor","numeric","numeric","factor", "factor"))
```

Les données ont maintenant été chargées dans *mydata_csv*. Pour afficher les données chargées, exécutez la ligne suivante :

```
> mydata_csv
```

R fera un renvoi des données à partir du fichier qu'il a lu.

```
   ID Longueur Largeur    Espece Site

1   1     4.1      3.5    setosa    A
2   2    14.9      3.0    set0sa    B
3   3     5.0      3.6    setosa    C
4   4      NA      3.9    setosa    A
5   5     5.8      2.7 virginica    A
6   6     7.1      3.0 virginica    B
7   7     6.3       NA virginica    C
8   8     8.0      7.0 virginica    C
```

Les sorties ci-dessus illustrent cinq colonnes de données. La première colonne, qui spécifie le numéro de la rangée, est créée de façon automatique par R lorsque les données sont chargées. La première rangée affiche le nom des colonnes que nous avons précisées.

Une des premières commandes à exécuter après le chargement du jeu de données est la commande dim, qui imprime les dimensions des données chargées par rangée et colonne. Cette commande vous permet de vérifier que toutes les entrées ont été lues correctement par R. Dans ce cas-ci, l'échantillon du jeu de données devrait comporter huit entrées avec cinq colonnes. Exécutons *dim* pour vérifier que toutes les données sont chargées.

```
> dim(mydata_csv)
```

Une fois la commande ci-dessus exécutée, R produira ce qui suit :

```
[1] 8 5
```

Cette sortie nous indique qu'il y a huit rangées et cinq colonnes dans les données chargées, ce qui correspond à nos attentes. Toutes les données ont donc été chargées.

Vous pouvez aussi exécuter la commande *summary*, qui vous donne des informations de base sur chacune des colonnes du jeu de données. Cette commande renvoie les valeurs maximales et minimales, les **quartiles** supérieurs et inférieurs (le quartile inférieur correspond à la valeur en dessous de laquelle se retrouvent 25% des données d'un jeu de données, tandis que le quartile supérieur correspond à la valeur au-dessus de laquelle se retrouvent 75% des données du jeu de données) ainsi que la médiane pour les colonnes de données numériques et la fréquence pour les colonnes de données catégoriques (le nombre de fois que chaque valeur apparaît dans une colonne).

```
> summary(mydata_csv)
```

```
      ID         Longueur        Largeur          Espece      Site
 1      :1   Min.   : 4.100   Min.   :2.700   set0sa   :1   A:3
 2      :1   1st Qu.: 5.400   1st Qu.:3.000   setosa   :3   B:2
 3      :1   Median : 6.300   Median :3.500   virginica:4   C:3
 4      :1   Mean   : 7.314   Mean   :3.814
 5      :1   3rd Qu.: 7.550   3rd Qu.:3.750
 6      :1   Max.   :14.900   Max.   :7.000
 (Other):2   NA's   :1        NA's   :1
```

Nous obtenons un résultat de cinq colonnes. Puisque nous avons demandé à R de lire les colonnes *Longueur* et *Largeur* comme des données numériques, il a calculé et affiché les informations sommaires en lien avec ces nombres dans leur colonne respective, notamment le minimum, le maximum, la moyenne et les quartiles. L'information sur chacune des colonnes est affichée dans les rangées sous le nom de la colonne. Par exemple, dans la colonne *Longueur*, la valeur minimale est de 4,1 et la valeur maximale, de 14,9. *1st Qu.* affiche le quartile inférieur qui est de 5,4 et *3rd Qu.* affiche le quartile supérieur, qui est de 7,55.

La rangée qui affiche « NA » nous indique s'il y a des valeurs manquantes. Dans R, les valeurs manquantes sont représentées par le symbole NA (*not available*). Le sommaire demandé fait apparaître deux valeurs manquantes : une dans la colonne *Longueur* et l'autre dans *Largeur*.

Dans la colonne *Espece*, où les données sont lues comme des données catégoriques (ou facteur), chacune des rangées affiche la fréquence à laquelle chacune des valeurs apparaît dans la colonne. Le sommaire montre trois instances de *setosa*, quatre de *virginica* et une de *set0sa*.

Il est possible ici d'identifier des erreurs d'enregistrement. Une des espèces de fleurs a été saisie de façon erronée; *set0sa* plutôt que setosa. Ce type d'erreur de frappe est très fréquent lors de l'enregistrement des données, mais il est souvent difficile à repérer, car le zéro et la lettre « o » se ressemblent beaucoup dans la plupart des polices.

R utilise une fonction de base – *is.na* – pour vérifier et énumérer les valeurs de données qui pourraient être manquantes. Cette fonction renvoie une valeur de vrai et faux pour chaque valeur d'un jeu de données. Si la valeur est manquante, la fonction *is.na* renvoie la valeur *TRUE*. Autrement, il renvoie la valeur *FALSE*.

```
> is.na(mydata_csv$Longueur)
[1] FALSE FALSE FALSE TRUE FALSE FALSE FALSE FALSE
```

À noter que le signe de dollar ($) est utilisé pour préciser les colonnes. Dans ce cas-ci, nous voulons vérifier si

la colonne *Longueur* dans le jeu de données CSV contient des valeurs manquantes. Le sommaire indique qu'une des valeurs est manquante; la fonction renvoie donc l'énoncé *TRUE* pour cette entrée de la colonne.

Les **valeurs aberrantes** sont des points de données qui diffèrent de façon importante des autres points du jeu de données; elles peuvent entraîner des problèmes avec certains types de modèles ou d'analyses de données. Par exemple, une valeur aberrante peut affecter la moyenne en étant anormalement petite ou grande. Certes, les valeurs aberrantes peuvent affecter les résultats d'une analyse, mais il faut faire preuve de prudence avant de les éliminer. Éliminez seulement une valeur aberrante que si vous êtes en mesure de prouver qu'elle est erronée (p. ex., si elle est attribuable à une erreur évidente dans la saisie de données). Un moyen simple d'identifier les valeurs aberrantes est de visualiser la distribution des données. Par exemple, inscrivez la commande suivante, qui demandera à R de générer une **boîte à moustache** (*box plot*).

```
> boxplot(mydata_csv$Longueur)
```

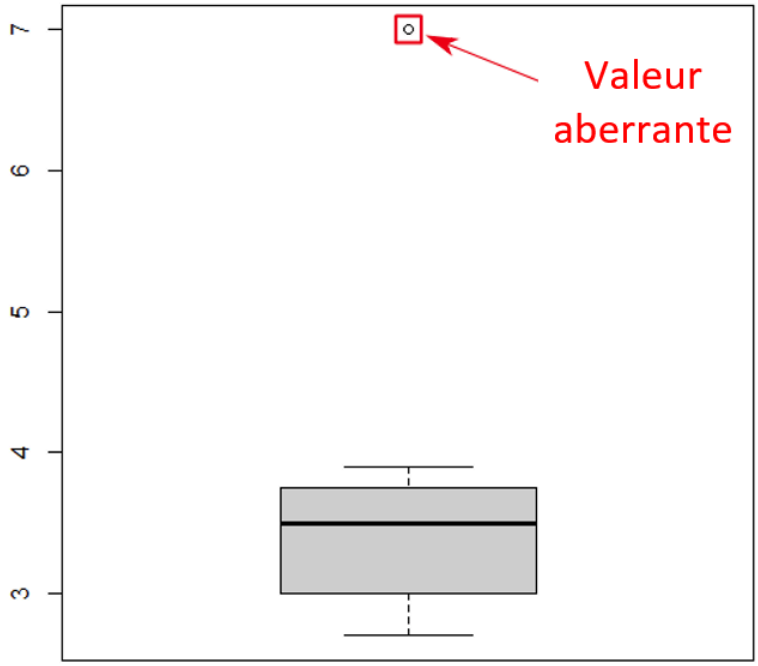



Figure 9. La valeur aberrante en relation à la boîte à moustache.

Les boîtes à moustache sont utiles pour détecter des valeurs aberrantes potentielles (voir figure 9). Une boîte à moustache aide à visualiser une colonne de données quantitatives en affichant le résumé des cinq emplacements courants : le minimum, la médiane, les premier et troisième quartiles (Q1 et Q3) et le maximum. Elle affiche également toute observation qui pourrait être qualifiée de valeur aberrante potentielle en utilisant le critère de l'écart interquartile (https://statsandr.com/blog/descriptive-statistics-in-r/#interquartile-range), c'est-à-dire l'écart ou la différence entre le premier et troisième quartile (voir figure 10). Une valeur

aberrante se définit comme étant un point de données situé à l'extérieur des moustaches de la boîte à moustache. Dans la boîte à moustache de la figure 9, un cercle au haut de la figure représente un point de données particulièrement éloigné des autres; la plupart des données se retrouvent à l'intérieur de la boîte du diagramme.

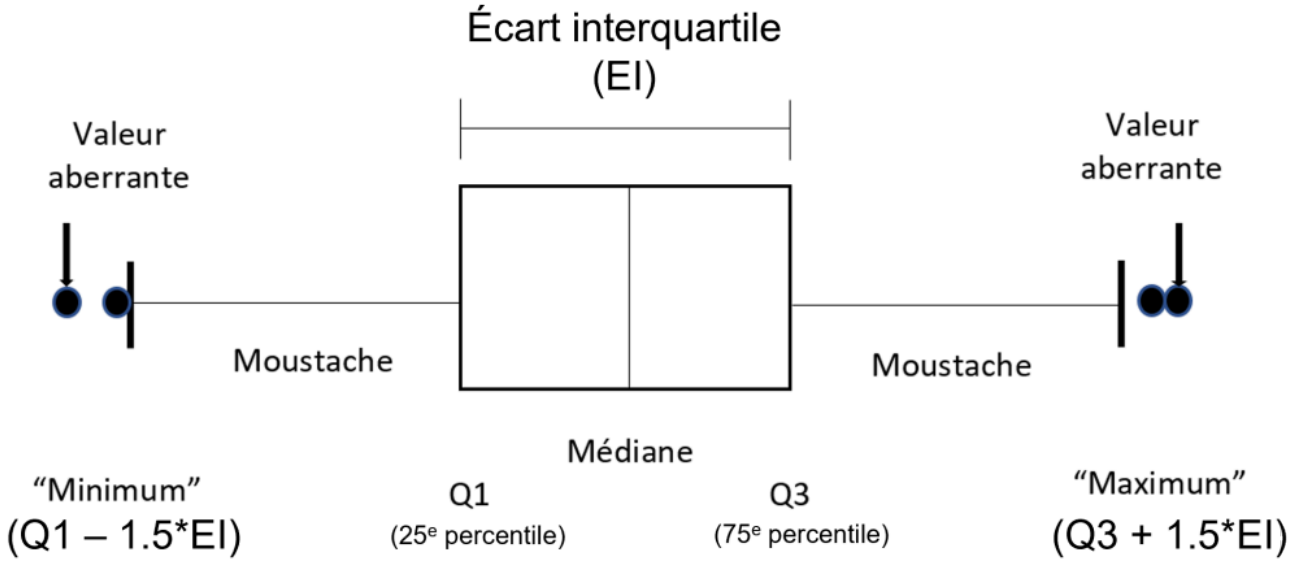



Figure 10. Comment interpréter une boîte à moustache.

Un autre moyen courant de détecter les valeurs aberrantes est de dessiner un histogramme des données (https://statsandr.com/blog/descriptive-statistics-in-r/#histogram). Un **histogramme** illustre la distribution des différentes valeurs des données. Selon l'histogramme ci-dessous, une observation est plus élevée que toutes les autres (la barre à la droite), ce qui correspond à ce que démontre la boîte à moustache. La commande suivante peut générer un histogramme :

```
> hist(mydata_csv$Longueur)
```

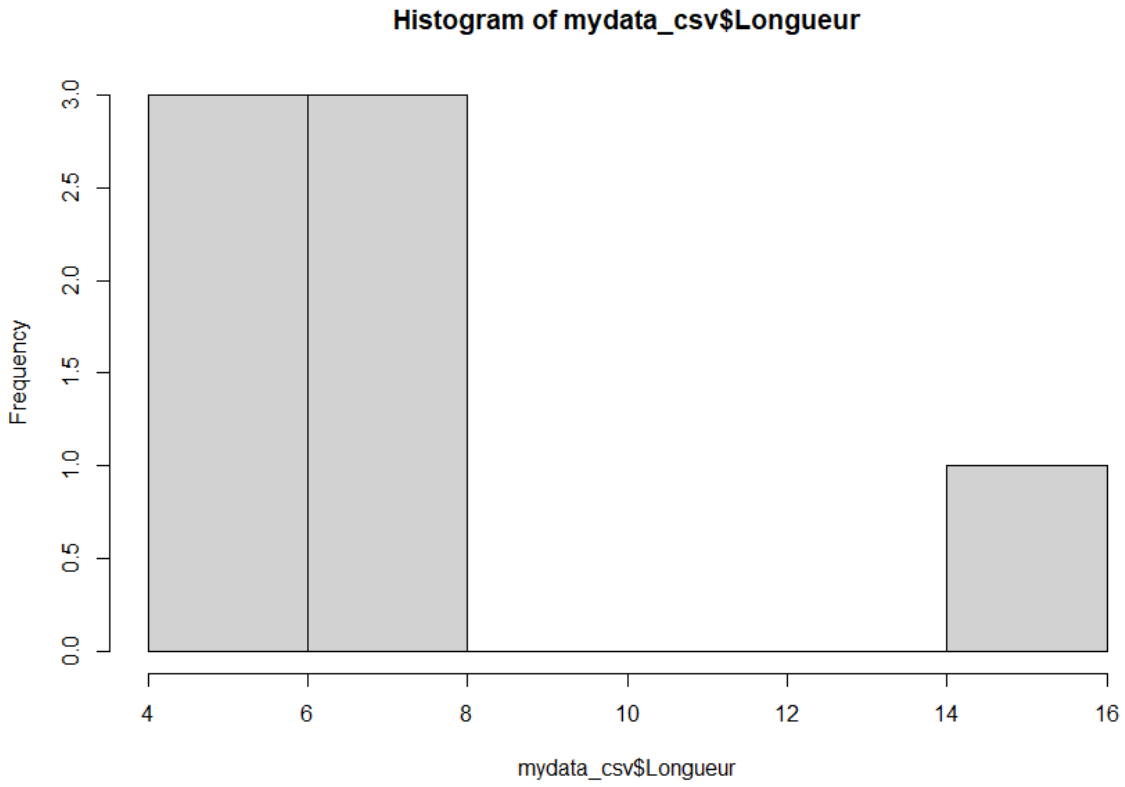



Figure 11. Histogramme de Longueur.

```
> summary(mydata_csv$Longueur)

Min.   1st Qu.  Median  Mean   3rd Qu.  Max.    NA's
4.100  5.400    6.300   7.314  7.550    14.900  1
```

À partir du sommaire, une des valeurs pour la longueur – 14,90 – semble anormalement élevée, bien qu'elle reste dans le domaine du possible. Un examen plus approfondi est donc nécessaire. Ce type de valeur aberrante peut avoir un impact important sur l'analyse des données; il faut donc bien comprendre sa validité. L'élimination des valeurs aberrantes doit être faite de façon judicieuse puisqu'elles peuvent représenter des observations réelles et importantes plutôt que des erreurs d'enregistrement.

Une fois que l'inspection préliminaire des données brutes est terminée, supposons que les données brutes contiennent quelques problèmes à corriger. Ces problèmes sont les suivants :

- un dédoublement de la colonne *Site*
- une erreur de frappe dans la colonne *Espece*
- des valeurs manquantes dans les colonnes *Longueur* et *Largeur*
- une valeur aberrante dans la colonne *Longueur*

Avec ces problèmes en tête, nous pouvons maintenant passer à la prochaine étape et commencer le nettoyage des données.

## 5. Nettoyer les données

Commençons par éliminer la colonne de trop. En utilisant les données ci-dessus, supposons que nous voulons éliminer la colonne *Site*. Tel que nous l'avons vu dans la sortie de la commande de sommaire, cette colonne est la cinquième du jeu de données. Pour l'éliminer, nous pouvons exécuter la commande suivante :

```
mydata_csv <- mydata_csv[-5]
```

La commande ci-dessus utilise les crochets pour préciser les colonnes de données originales. En utilisant un numéro négatif, nous indiquons à R de récupérer toutes les colonnes à l'exception de la colonne spécifiée. Dans ce cas-ci, la colonne *Site* est la cinquième colonne. Puisque nous voulons éliminer la cinquième colonne, mais conserver toutes les autres, nous inscrivons « -5 » entre crochets pour indiquer à R de récupérer toutes les colonnes sauf la cinquième. Ensuite, en réaffectant les nouvelles données récupérées à *mydata_csv*, nous réussissons à éliminer la cinquième colonne.

Pour confirmer que la colonne a bel et bien été éliminée, nous pouvons utiliser la commande dim, comme nous l'avons vu précédemment.

```
> dim(mydata_csv)

[1] 8 4
```

Le résultat montre que les données comptent désormais quatre (plutôt que cinq) colonnes.

Ensuite, il faut nettoyer les fautes de frappe. Dans ce cas-ci, nous savons que « set0sa » devrait être corrigée à « setosa ». Toutes les cellules concernées peuvent être remplacées en utilisant la commande suivante :

```
> mydata_csv[mydata_csv=="set0sa"] = "setosa"
> summary(mydata_csv)

       ID        Longueur          Largeur          Espece
 1      :1   Min.   : 4.100   Min.   :2.700   set0sa   :0
 2      :1   1st Qu.: 5.400   1st Qu.:3.000   setosa   :4
 3      :1   Median : 6.300   Median :3.500   virginica:4
 4      :1   Mean   : 7.314   Mean   :3.814
```

```
5       :1   3rd Qu.: 7.550   3rd Qu.:3.750
6       :1   Max.   :14.900   Max.   :7.000
(Other):2   NA's   :1         NA's   :1
```

À noter que l'opérateur d'égalité – == – sélectionne toutes les instances de *set0sa* (une seule dans ce cas-ci) et le symbole = lui attribue la valeur *setosa*. L'utilisation de deux signes d'égalité pour tester l'égalité et d'un seul signe d'égalité pour rendre quelque chose égal à quelque chose d'autre est une convention courante en programmation.

Il y a maintenant zéro entrée dans la colonne *Espece* qui porte le nom *set0sa* dans la sortie *summary()*. Les données ont donc été nettoyées de cette faute de frappe.

Il existe plusieurs moyens de gérer les données manquantes. Une option implique d'exclure la valeur manquante de l'analyse. Avant d'éliminer « NA » de la colonne *Longueur*, la fonction *mean()* renvoie « NA » comme suit :

```
> mean(mydata_csv$Longueur)

[1] NA
```

En effet, il est impossible d'utiliser « NA » dans une analyse numérique. L'utilisation de *na.rm* pour éliminer la valeur manquante « NA » donne une moyenne de 7,314286 :

```
> mean(mydata_csv$Longueur, na.rm = T)

[1] 7.314286
```

**Exercice 2**

Vérifiez s'il y a des valeurs aberrantes dans la colonne Largeur de sample.csv en utilisant une boîte à moustache. Calculez ensuite la moyenne de la longueur en éliminant les valeurs aberrantes.

Voir le solutionnaire pour les réponses.

# Conclusion

Les procédures de nettoyage des données sont d'une importance capitale pour une analyse des données réussie et elles devraient être appliquées avant de procéder à l'analyse. Dans ce chapitre, nous avons fait un survol rapide des problèmes et des solutions liées au nettoyage des données auxquels les chercheuses et chercheurs sont confrontés en utilisant Excel/Google Sheets et le langage R. Il existe des bibliothèques complètes de fonctions de manipulation des données et elles fournissent une panoplie de fonctionnalités pour vous aider dans votre processus de nettoyage des données. Vous pouvez trouver des informations supplémentaires sur le langage R sur les sites suivants (en anglais uniquement) : https://cran.r-project.org/manuals.html (https://cran.r-project.org/manuals.html) et https://cran.r-project.org/web/packages/available_packages_by_name.html (https://cran.r-project.org/web/packages/available_packages_by_name.html).

**Éléments clés à retenir**

- Les procédures générales pour se préparer au nettoyage des données sont de faire une copie de sauvegarde, de comprendre les données, de planifier le processus de nettoyage et de choisir les outils qui conviennent.
- Les fonctionnalités de Excel peuvent servir à effectuer de nombreuses tâches de base de nettoyage des données.
- Le langage de programmation R est un progiciel utile et gratuit qui peut servir pour des procédures de nettoyage plus avancées.

# Remerciements

Les autrices aimeraient remercier Kristi Thompson et d'autres éditrices dont les commentaires constructifs ont permis d'améliorer ce chapitre.

## À propos des auteurs

Dr. Rong Luo

Rong Luo est spécialiste de l'apprentissage au Academic Data Center de la bibliothèque Leddy de l'Université de Windsor. Ses recherches portent sur la modélisation statistique, l'imputation des données manquantes, l'analyse des données d'enquêtes sociales et l'évaluation des compétences informationnelles. Elle utilise des modèles quantitatifs et qualitatifs pour ses projets de recherche.

## Berenica Vejvoda

Berenica Vejvoda est bibliothécaire chargée des données de recherche à la bibliothèque Leddy de l'Université de Windsor où elle est responsable de la coordination et de la gestion des services de données de recherche. Elle est également responsable de l'orientation stratégique et de la mise en œuvre des services de gestion des données de recherche pour l'Université de Windsor dans le cadre d'une initiative à l'échelle du campus. Berenica est également directrice académique de la branche de l'Université de Windsor du centre de données de recherche de Statistique Canada. Les recherches de Berenica portent sur les déterminants sociaux de la santé des populations marginalisées dans une optique intersectionnelle ainsi que sur les principes d'inclusion des données appliqués à la gestion des données de recherche.

9.

# UN APERÇU DU FASCINANT MONDE DES FORMATS DE FICHIERS ET DES MÉTADONNÉES

Émilie Fortin

**Objectifs d'apprentissage**

À la fin de ce chapitre, vous pourrez :

1. Comprendre ce qu'est un format pérenne.
2. Choisir adéquatement un format selon vos besoins.
3. Comprendre l'utilité des métadonnées.
4. Distinguer les différents types de métadonnées.

## Introduction

Le cycle de vie des données de recherche inclut toujours une étape de préservation, qui porte parfois le nom de conservation ou d'archivage. Cette étape est liée à celle de la réutilisation des données, car personne ne peut remployer des données endommagées ou inaccessibles. Le chapitre « La préservation numérique des données de recherche » aborde la question de la préservation numérique et le présent chapitre se concentre sur deux éléments qui permettront de repérer et de réutiliser des données : les formats de fichiers et les métadonnées.

# Les formats de fichiers

**Évaluation préliminaire**

Répondez le plus honnêtement possible aux questions suivantes (Oui, Non) :

- Avez-vous des difficultés à ouvrir certains fichiers que vous avez créés il y a plus de dix ans?
- Pensez-vous que, dans une dizaine d'années, vous aurez des difficultés à ouvrir les fichiers que vous créez cette année?
- Pensez-vous qu'un fichier PDF est un parfait format de préservation?
- Est-ce que vous vous réveillez la nuit en vous demandant si vos arrière-petits-enfants vont encore avoir des photos numériques de vous?
- Est-ce que vous adorez les applications interactives et vous voudriez que tous vos projets soient le plus connectés possible?

Si vous avez répondu oui à plus de deux questions, cette section devrait vous être utile.

## Qu'est-ce qu'un format?

Les formats de fichiers numériques sont conçus selon des principes structurels et organisationnels prédéfinis. Ces principes sont généralement listés dans un document de spécifications qui fournit des détails sur les subdivisions, l'encodage et les relations internes qui permettent de construire et de valider un format. Une spécification de format indique les frontières entre les **séquences de bits**. Celles-ci peuvent représenter un caractère, une opération à effectuer (instruction-machine), une sélection de couleur, etc.

En résumé, il s'agit d'une série de 1 et de 0 spécifique et conventionnée utilisée pour reconnaître un format.

À partir du moment où vous utilisez un support informatique, peu importe l'usage que vous en faites, gardez à l'esprit que vous utilisez, créez ou modifiez des formats.

## Qu'est-ce qu'un format pérenne?

Aucun format n'est vraiment pérenne. Ceux qui sont jugés acceptables pour la préservation à long terme sont

des formats qui restent accessibles dans le temps malgré les évolutions technologiques. Un bon format aujourd'hui peut devenir désuet dans deux, cinq ou dix ans.

Voici certains critères qui permettent de juger de la pérennité d'un format.

- Complexité
- Rétrocompatibilité
- Encodage
- Dépendance
- Ouverture
- Métadonnées
- Propriété
- Utilisation
- Évolution
- Protections

**Complexité.** Le format doit offrir de bonnes capacités, mais éviter d'être trop complexe, sinon il sera difficile à préserver dans le temps avec toutes ses fonctionnalités. La complexité d'un format peut se définir par sa lisibilité par les humains, sa compression et la variété de ses fonctionnalités. Plus il faut déployer d'efforts pour déchiffrer un format, plus il y a de chance qu'il ne soit pas parfaitement compris.

**Rétrocompatibilité.** Le format est-il reconnu pour sa **rétrocompatibilité**? Lorsqu'une nouvelle version d'un logiciel est produite, à quel point est-ce possible d'ouvrir les formats créés avec les anciennes versions du logiciel? Les générations d'un même format sont-elles très différentes entre elles?

> **Fait intéressant :** saviez-vous qu'Adobe assure la rétrocompatibilité des formats PDF jusqu'à la version 1.3 (sortie en 1999) uniquement?

**Encodage.** Dans un environnement occidental, le format utilisera probablement un encodage standard de type **ASCII** ou **Unicode**. Si vous utilisez des symboles ou des langues non latines, l'encodage est important, car vous désirez que la lettre ou le symbole que vous utilisez s'affiche correctement, peu importe qui ouvrira votre fichier.

**Dépendance.** Il est ici question de dépendance du format envers son logiciel, mais également envers une technologie spécifique, envers d'autres fichiers ou envers son environnement. Le format peut-il être ouvert seulement par un logiciel spécifique? Le format est-il une sorte de contenant dans lequel on retrouve d'autres

formats (p. ex., format de compression de type ZIP, vidéo intégrée dans un fichier texte, fichier vidéo avec une bande-son)? Le format doit-il se connecter à votre environnement pour fonctionner (p. ex., livre interactif qui est connecté à la caméra de votre téléphone)?

Les ressources externes à votre fichier peuvent être perdues avec le temps, donc plus le format a de dépendances, plus il sera difficile à préserver dans sa forme actuelle.

**Ouverture.** Un **format ouvert** est préférable.

Exemples de formats ouverts : fichiers Office avec X (p. ex., XLSX, DOCX), PDF, TXT, JPG, PNG, CSV.

> **Fait intéressant :** certaines **extensions** cachent parfois des fichiers aux formats ouverts. Par exemple, un fichier de scripts peut avoir des extensions comme HTML, XML, SC, mais il s'agit en réalité de formats texte.

> **Fait intéressant :** certains formats ouverts sont devenus au fil du temps des normes, par exemple PDF et PDF/a sont des normes ISO.

**Métadonnées.** Il s'agit ici des métadonnées internes au fichier. Pensez aux propriétés du fichier auxquelles vous pouvez accéder dans les logiciels et par le biais de votre système d'exploitation.

Identifier un format est une première étape, mais documenter le plus possible le contenu et le contenant à même le format est également très utile. Plus un objet numérique est documenté, mieux il pourra être compris dans les années à venir. Un format qui est un bon support aux métadonnées est avantageux, car si le fichier ne s'ouvre plus, il est parfois possible d'obtenir de l'information précieuse grâce aux métadonnées (p. ex., titre, créateur, logiciel utilisé pour enregistrer le format). Pour plus de détails à ce sujet, veuillez consulter la section sur les métadonnées.

**Propriété.** Un format propriétaire appartient à une personne morale. Il peut être ouvert ou non. Son évolution est contrôlée par son propriétaire. Ces formats sont généralement rattachés à des logiciels particuliers. Lorsque les formats sont **non propriétaires**, leur évolution est contrôlée par une communauté d'utilisateurs et ils sont en grande majorité ouverts.

- Exemples de formats non propriétaires : MKV, TXT, XML, CSV, PNG;

- Exemples de formats propriétaires, mais ouverts : fichiers Office avec X (p. ex., DOCX, XLSX), PDF, RAR;
- Exemples de formats propriétaires : AutoCAD, PSD, WMA.

**Utilisation.** Si uniquement une dizaine de personnes utilisent un format, même si celui-ci est ouvert et non propriétaire, il va disparaître. À l'opposé, un format propriétaire extrêmement populaire est très peu à risque de s'éteindre dans les prochaines années.

Si un format propriétaire fermé est adopté comme norme par une bibliothèque, un centre d'archives ou une communauté de recherche, il est fort possible que le format soit pérenne grâce à sa popularité. Toutefois, son évolution doit être surveillée de près.

**Évolution.** Le format doit suivre un cycle d'amélioration en continu tout en évitant les abus. Les systèmes changent, donc les logiciels et les formats doivent évoluer; un format statique n'est pas nécessairement meilleur qu'un format qui se développe. Toutefois, lancer une série de nouvelles versions d'un format dans un intervalle de temps limité peut être considéré comme abusif, car les changements fréquents menacent l'accessibilité à long terme.

**Protections.** Il existe plusieurs mesures techniques de protection de fichiers. Par exemple, le cryptage et l'utilisation d'un mot de passe sont de bonnes méthodes pour protéger des données sensibles, mais elles ne sont pas compatibles avec la préservation à long terme. Imaginons simplement l'impact qui peut avoir la perte d'un mot de passe!

De la même manière, certaines mesures permettant de protéger la propriété intellectuelle d'un fichier, comme les verrous sur les livres électroniques, risquent de compromettre l'accès au contenu.

> **Fait intéressant :** certaines plateformes permettent de restreindre l'accès aux fichiers en appliquant un contrôle de permissions. Cette méthode est de loin préférable au verrouillage des fichiers eux-mêmes.

## Comment choisir un format pour un projet de recherche?

Les critères qui définissent un format pérenne sont importants, mais il est fondamental de bien les appliquer aux besoins de votre projet. Il n'est pas nécessaire de se conformer à tous les critères. De plus, si votre domaine de recherche vous oblige à utiliser un format qui ne répond à aucun critère de format pérenne, vous ne devez

pas vous empêcher de l'utiliser, simplement rester conscient qu'il y aura un impact sur la préservation des données.

Voici quelques questions que vous pouvez vous poser pour vous aider à choisir le meilleur format :

- Avez-vous besoin de préserver vos données à long terme? Si vous prévoyez supprimer l'ensemble de vos données dans cinq ans et ne pas les partager, ne pensez qu'à vos propres besoins d'utilisation.
- Si vous utilisez des appareils/instruments de recherche, avez-vous un choix de format? Si oui, tentez d'opter pour un format pérenne si cette option n'a aucun impact sur votre recherche.
- Est-ce que l'aspect ou la mise en forme des données est important ou seulement les données elles-mêmes? Si l'aspect des données n'est pas important, vous pouvez opter pour un format plus simple. Par exemple, un document textuel conservé en tant que PDF permet de préserver l'aspect et la mise en forme d'un document, mais la réutilisation du contenu est complexe. Cependant, si le document textuel est converti en format TXT, la mise en forme est perdue, mais le contenu pourra facilement être réutilisé.
- Est-ce que les données sont indépendantes ou connectées à d'autres données? Si vos données sont rattachées à des équations ou à d'autres fichiers, vous devez conserver ces liens.
- Est-ce que vous devez contrôler le poids de vos fichiers? Si vous êtes limité en espace, vous n'aurez peut-être pas le choix d'opter pour une compression. Essayer d'utiliser une **compression sans perte**.
- Dans votre discipline, existe-t-il un format qui est utilisé par la majorité de vos collègues et qui est incontournable?

Dans certains cas, il est envisageable de garder des données à la fois dans leur format d'origine et dans un format pérenne, mais cette préservation en double doit avoir un objectif. Par exemple, vos données peuvent desservir deux communautés très différentes qui n'utilisent pas le même niveau de technologie. Toutefois, vous devez éviter au maximum la confusion que peuvent apporter deux versions d'un même jeu de données.

Une autre option pourrait être de garder uniquement le format original et de générer au besoin des copies moins lourdes. Cette option est risquée dans le sens qu'elle implique une dépendance aux logiciels qui sont capables de lire le format original.

Vous devez garder en tête que des données illisibles dans dix ans ne seront plus utiles à personne, y compris pour vous-mêmes.

La plupart des bibliothèques nationales publient une liste de formats recommandés (voir la section Lectures et ressources supplémentaires); il peut être utile de les consulter. Vous trouverez ici quelques-uns des formats qui font généralement consensus en 2023.

## Bases de données

Une base de données implique des valeurs, mais également une structure et des relations entre les valeurs. Les bases de données les plus couramment utilisées au moment d'écrire ces lignes sont Microsoft Access, Oracle, MySQL et PostgresSQL. Lorsque vient le temps de se pencher sur la préservation à long terme d'une base de données, il faut évaluer les besoins futurs : est-ce que la base de données est encore utilisée? Est-ce que la préservation des valeurs seules sera suffisante ou faut-il aussi documenter la structure et les relations?

Les bases de données sont complexes à préserver vu leur structure et l'évolution de leur contenu. Il est important de circonscrire les besoins avant de choisir un format de préservation.

Quelques formats recommandés :

- Formats avec séparateurs de valeurs (CSV, TSV, TXT) : préserve les données, mais pas les relations ni les formules. Surtout utile pour les bases de données simples et de petites tailles;
- Format de préservation de base de données (SIARD 1.0 et 2.0) : format ouvert établi pour la préservation de bases de données, mais n'est utilisable que pour certains types de bases de données;
- Format léger de base de données relationnelles (SQLITE) : format simple utilisé pour les bases de données relationnelles.

## Données tabulaires

Des **données tabulaires** sont des données disposées sous la forme de tables ou de tableaux, c'est-à-dire en ligne et en colonnes.

Le principal défi de ces formats est de composer avec les formules, les macros et le contenu intégré. Il faut aussi retenir que d'exporter un fichier tabulé vers un logiciel infonuagique, ou l'inverse, peut occasionner des pertes ou des erreurs.

Notez que le format SAV de SPSS est parfois recommandé, bien que sa documentation ne soit pas officielle et que sa rétrocompatibilité ne soit pas garantie.

Quelques formats recommandés :

- Données avec séparateurs (CSV, TXT, TSV) : fichiers simples, mais perte des formules et des relations entre les cellules;
- Microsoft Excel (XLSX) : format documenté et ouvert, mais non recommandé par certains dépôts, car il s'agit d'un format propriétaire complexe. Dans certains cas, il reste incontournable. Si utilisé, s'assurer de créer un fichier avec Office 2013 ou une version plus récente;

- OpenDocument (ODS, FODS) : généralement associé à LibreOffice, une suite logicielle développée comme équivalent ouvert des logiciels Microsoft. Structure basée sur le XML. La version 1.2 est certifiée en tant que norme ISO; la version 1.3 a obtenu le statut de standard.

## Texte

Un document textuel peut être très simple, mais il peut également poser certains défis. Par exemple, l'utilisation d'un logiciel de traitement de texte dans le nuage facilite grandement la collaboration, mais l'extraction de ces documents pour les enregistrer localement peut affecter leur mise en forme et parfois la fonctionnalité des hyperliens. Vous devez aussi vous demander quelles versions garder, car il n'est pas pertinent de préserver toutes les modifications et les commentaires d'un texte. Ce peut être uniquement certaines versions intermédiaires avec la mouture finale.

Si le document textuel contient des objets intégrés, par exemple une image ou un tableau, le format sélectionné peut varier. Le choix de la police peut également affecter la préservation d'un document textuel.

Pour la compréhension du contenu, le texte peut également faire référence à d'autres documents. Ces relations sont importantes et doivent être maintenues.

Le format le plus approprié est celui qui conservera les fonctionnalités du document d'origine tout en permettant sa consultation à long terme.

Quelques formats recommandés :

- OpenDocument (ODT, OTT) : généralement associé à LibreOffice, une suite logicielle développée comme équivalent ouvert des logiciels Microsoft. Structure basée sur le XML. La version 1.2 est certifiée en tant que norme ISO; la version 1.3 a obtenu le statut de standard;
- Plein texte (TXT) : pas de mise en page, mais accessible facilement, ne dépend d'aucun programme, c'est d'ailleurs pourquoi il est très recommandé pour les fichiers **LISEZ-MOI**;
- PDF et PDF/A : format commun, souvent utilisé pour la préservation à long terme. Idéalement, s'assurer de ne garder que des versions 1.3 et suivantes;
- Publication électronique (EPUB) : format ouvert, très utilisé pour la publication numérique.

**Fait intéressant :** Les fichiers EPUB commerciaux peuvent contenir des protections intégrées visant à protéger la propriété intellectuelle qui empêchent la copie et le partage. Ces verrous numériques sont incompatibles avec la préservation à long terme.

## Images

La plupart des institutions de préservation numérique s'entendent sur les formats d'image les plus sécuritaires à utiliser. Les formats mentionnés ci-dessous sont matriciels, c'est-à-dire qu'ils se composent d'une série de points appelés pixels.

La qualité d'un format peut varier selon plusieurs facteurs comme la résolution (la plus connue), mais également l'espace colorimétrique ou la profondeur des couleurs. Souvent, plus une image est de qualité, plus le fichier est lourd.

Les formats propriétaires RAW ne sont pas recommandés pour la préservation à long terme. À l'opposé, une image créée avec un format compressé (p. ex., GIF, JPG, BMP) pourrait être préservée telle quelle. Avant de choisir un format d'image, les besoins et les moyens technologiques, humains et financiers doivent être évalués.

Quelques formats recommandés :

- Tagged Image File Format (TIFF) : format le plus utilisé pour la préservation d'images, mais lourd;
- Joint Photographic Experts Group 2000 (JP2) : plus léger que le TIFF, mais moins largement utilisé;
- Joint Photographic Expert Group (JPG) : très utilisé, mais l'image est compressée;
- Portable Network Graphics (PNG) : utilise une compression sans perte. Assez couramment utilisé, mais il n'est pas toujours pris en charge par les logiciels.

## Audio

Un format audio est un contenant avec un ou plusieurs flux de données audio.

Un fichier audio comporte plusieurs caractéristiques à considérer qui influenceront le rendu et l'authenticité du son (p. ex., canaux, compression, nombre de bits par échantillon, nombre d'échantillons par seconde). Si le fichier d'origine est déjà compressé (p. ex., MP3, AAC), il n'est peut-être pas pertinent de le migrer vers un autre format.

Notez que le format MP3 est un format compressé généralement non recommandé pour la préservation à long terme, mais son adoption généralisée en fait un format assez fiable si le fichier d'origine a été créé ainsi.

Quelques formats recommandés :

- Free Lossless Audio Codec (FLAC) : fichier avec une compression sans perte, format plus léger que les WAVE;

- PCM WAVE (WAV) : format de qualité utilisé par plusieurs bibliothèques nationales lors de la numérisation;
- Broadcast WAVE (BWF) : permet l'ajout de métadonnées dans les fichiers;
- Ogg Vorbis (OGG) : format ouvert avec une meilleure compression que le MP3, mais moins populaire.

## Vidéo

Les formats vidéo sont complexes, en constante évolution, et aucun ne fait consensus dans la communauté de préservation numérique.

Les formats vidéo sont généralement des contenants avec des images ou des flux de données vidéo et du son. Plusieurs caractéristiques (p. ex., couleur, compression, son) peuvent influencer leur préservation à long terme. Plus d'un format peut être utilisé pour un projet selon les besoins de création, de transformation, de diffusion ou autre.

Le défi le plus important est de trouver l'équilibre entre le poids du fichier et sa qualité.

Quelques formats recommandés :

- MP4 avec H.264 : format compressé surtout utilisé pour la diffusion, très largement répandu;
- QuickTime (MOV) ou Audio Video Interleaved (AVI) non compressé 4:2:2 : formats très lourds, mais de bonne qualité;
- Matroska avec codec FFV1 (MKV) : format standardisé pas trop compressé;
- Material Exchange Format avec JPG 2000 (MXF) : recommandé par certaines bibliothèques nationales, bien documenté, mais peu utilisé dans le public;
- Digital Picture Exchange (DPX) : format très lourd utilisé lors de la numérisation de pellicules filmiques.

## Données géospatiales

Les données géospatiales sont également abordées dans le chapitre « Les données de recherche géospatiales au Canada: un survol des projets régionaux. » Ces données consistent généralement en une série de fichiers qui se complètent. Elles peuvent être intrinsèquement liées au système d'information géographique qui les exploite. Les métadonnées, les systèmes de référencement des coordonnées et la précision des coordonnées, c'est-à-dire à quel point une valeur observée et enregistrée est proche de la valeur réelle, doivent être préservées avec les données.

Lister des formats recommandés pour la préservation à long terme des données géospatiales est presque impossible vu leur complexité (p. ex., plusieurs types de structures différentes, beaucoup de formats

propriétaires). Il n'y a aucun consensus à ce sujet, conserver le format d'origine peut s'avérer la meilleure solution.

Quelques formats recommandés :

- Geospatial Tagged Image File Format (GEOTIFF) : format ouvert qui permet d'ajouter des coordonnées géographiques à une image;
- Geographic Markup Language (GML) : format ouvert basé sur une norme, mais il est complexe;
- Keyhole Markup Language (KML, KMZ) : langage XML qui peut être associé à plusieurs autres fichiers qui doivent aussi être archivés (évitez d'utiliser des hyperliens). Format ouvert et largement utilisé;
- ESRI Shapefile (SHP SHX, DBF, PRJ, SBX, SBN) : format propriétaire, mais ouvert et très utilisé.

## Aller plus loin : comment identifier un format?

Pour identifier un format de fichier, il suffit la plupart du temps de regarder sa section finale, c'est-à-dire son extension. Par exemple, le fichier « mes-notes.xlsx » est un fichier Excel alors que « ma-photo.jpg » est une image. Cette méthode a ses limites puisqu'une extension peut être modifiée, volontairement ou par erreur, ou être complètement inconnue. Certains systèmes d'exploitation sont même configurés par défaut pour cacher l'extension des fichiers, ce qui peut compliquer la tâche.

Le meilleur moyen d'identifier un format est d'utiliser sa **signature**. La signature d'un fichier correspond à une série de bits qui s'enchaînent de façon prévisible au début, à la fin ou aux deux extrémités d'un fichier.

Un outil comme PRONOM, très utilisé dans la communauté de préservation numérique, enregistre les signatures de début de fichiers (BOF pour Beginning of File) et de fin de fichiers (EOF pour End of File) et permet de récupérer l'identifiant unique d'un format. Par exemple, la signature x-fmt/398 identifie les JPG version 2.0. Connaître un format permettra aux personnes qui voudront consulter les jeux de données de savoir comment les ouvrir.

Quelques outils d'identification :

- PRONOM : http://www.nationalarchives.gov.uk/pronom/ (http://www.nationalarchives.gov.uk/pronom/);
- Siegfried : https://www.itforarchivists.com/siegfried (https://www.itforarchivists.com/siegfried);
- FIDO : https://github.com/openpreserve/fido (https://github.com/openpreserve/fido) ou https://fido-js.glitch.me/ (https://fido-js.glitch.me/).

Des outils qui permettent de visualiser les fichiers en code hexadécimal :

- HexEd.it : https://hexed.it/ (https://hexed.it/);
- Literate-binary : https://github.com/marhop/literate-binary (https://github.com/marhop/literate-binary).

# Les métadonnées

**Évaluation préliminaire**

Répondez le plus honnêtement possible aux questions suivantes (Oui, Non) :

- Comprenez-vous ce que signifient « des données à propos de données »?
- Savez-vous qu'il existe plus d'un type de métadonnées?
- Savez-vous que certaines métadonnées s'inscrivent automatiquement dans vos fichiers?
- Savez-vous que votre beau-frère pourrait apparaître comme auteur d'un fichier que vous avez créé si vous avez utilisé son ordinateur?
- Réalisez-vous le pouvoir des métadonnées?

Si vous avez répondu non à plus de deux questions, cette section devrait vous être utile.

## Une introduction aux métadonnées

Les métadonnées sont des éléments d'information utilisés pour décrire le contenu ou le contenant d'une ressource. Elles peuvent être structurées ou non.

Afin de mieux comprendre les métadonnées, commençons avec un exemple de données brutes :

CCTTTATCTAATCTTTGGAGCATGAGCTGGCATAGTTGGAACCGCCCTCAGCCTCCT
CATCCGTGCAGAACTTGGACAACCTGGAACTCTTCTAGGAGACGACCAAATTTACAA
TGTAATCGTCACTGCCCACGCCTTCGTAATAATTTTCTTTATAGTAATACCAATCATG
ATCGGTGGTTTCGGAAACTGACTAGTCCCACTCATAATCGGCGCCCCCGACATAGCA
TTCCCCCCGTATAAACAACATAAGCTTCTGACTACTTCCCCCATCATTTCTTTTACTTC
TAGCATCCTCCACAGTAGAAGCTGGAGCAGGAACAGGGTGAACAGTATATCCCCCTC
TCGCTGGTAACCTAGCCCATGCCGGTGCTTCAGTAGACCTAGCCATCTTCTCCCTCC

ACTTAGCAGGTGTTTCCTCTATCCTAGGTGCTATTAACTTTATTACAACCGCCATCAA
CATAAAACCCCCAACCCTCTCCCAATACCAAACCCCCCTATTCGTATGATCAGTCCT
TATTACCGCCGTCCTTCTCCTACTCTCTCTCCCAGTCCTCGCTGCTGGCATTACTAT
ACTACTAACAGACCGAAACCTAAACACTACGTTCTTTGACCCAGCTGGAGGAGGAG
ACCCAGTCCTGTACCAACACCTCTTCTGATTCTTCGGCCATCCAGAAGTCTATATCC
TCATTTTAC

Les données brutes provenant de la recherche, dépourvues de métadonnées, sont intéressantes, mais peu parlantes pour la majorité des gens. Il est facile de se rendre compte qu'il y a un long chemin entre les données brutes extraites au cours d'un projet de recherche et leur signification utilisable par l'humain.

Si une généticienne décrit les données brutes ci-dessus, elle pourrait ajouter un premier niveau de métadonnées :

- >Seq1 [organism=*Carpodacus mexicanus*] *C. mexicanus* clone 6b actin (act) mRNA, partial cds

Un deuxième niveau de métadonnées serait la description du jeu de données dont fait partie cette séquence : le séquençage génétique, dans ce cas-ci de Carpodacus mexicanus, une espèce d'oiseau.

- Il s'agit d'une séquence nucléotidique de *Carpodacus mexicanus* (clone 6 b). (A = Adénine, G = Guanine, C = Cytosine, T = Thymine : bases d'acide nucléique).

Un troisième niveau de métadonnées permettrait de mieux caractériser les métadonnées précédentes en normalisant la nomenclature utilisée, ce qui facilitera le repérage et la relance dans d'autres corpus documentaires, tels les répertoires d'articles ou les dépôts institutionnels :

- Roselin familier – Génétique
- Séquence nucléotidique

Un quatrième niveau lierait ces métadonnées à d'autres informations pertinentes, comme une image.

*Carpodacus mexicanus QC (https://commons.wikimedia.org/wiki/File:Carpodacus_mexicanus_QC.jpg)*, Simon Pierre Barrette, CC BY-SA 3.0 (https://creativecommons.org/licenses/by-sa/3.0/).

Le principal rôle des métadonnées est de décrire et de favoriser le repérage. Toutes les métadonnées présentes devraient répondre aux tâches que réalisent les personnes utilisant des moteurs de recherche généraux ou académiques.

- Trouver – c'est-à-dire trouver des ressources correspondant aux critères de recherche;
- Identifier – établir le contexte des données et confirmer que la ressource décrite correspond à la ressource recherchée, ou bien établir une distinction entre deux ou plusieurs ressources possédant des caractéristiques semblables;
- Sélectionner – c'est-à-dire sélectionner une ressource pertinente pour les besoins de la personne qui cherche.

Les métadonnées nécessaires à la préservation sont celles qui assurent l'authenticité et l'accessibilité à long terme des ressources numériques et qui permettent la restitution des fichiers dans une forme accessible, lisible et intelligible. Vous devez être en mesure de gérer et de découvrir les métadonnées indépendamment des systèmes avec lesquels les ressources ont été créées.

## Normalisation des métadonnées

Certaines métadonnées peuvent être normalisées, tels les noms des personnes responsables de la recherche, les méthodes de collecte et d'analyse des données, les titres des variables, les sujets abordés par la recherche ainsi que les couvertures temporelles ou géographiques. D'autres types de métadonnées obéiront tout simplement

à des règles de description plus ou moins précises visant à en uniformiser la présentation, par exemple le titre attribué à un projet de recherche ou un résumé décrivant un jeu de données.

Plus les métadonnées sont normalisées, plus elles contribuent aux **principes FAIR** (pour plus de détails, voir le chapitre 2, « Les principes FAIR et la gestion des données de recherche » ) et plus elles permettent la repérabilité, l'accessibilité, l'interopérabilité et la réutilisation des ressources qu'elles représentent. Au moment de décrire une ressource, que ce soit une donnée ou un jeu de données, il faut cibler les métadonnées qui seront les plus utiles, car l'investissement en temps et en argent doit être rentable.

Plusieurs moyens peuvent être utilisés pour normaliser des métadonnées et il y a souvent une confusion terminologique, car certains termes sont utilisés pour décrire de façon erronée des réalités différentes.

## Schémas de métadonnées

Pour bien comprendre ce qu'est un **schéma de métadonnées**, imaginez un formulaire en ligne avec des boîtes à remplir. Le schéma se cache derrière, il s'agit de ce qui va donner un sens aux renseignements que vous inscrirez dans les boîtes.

Certains schémas spécifient avec quelle syntaxe les éléments doivent être encodés alors que d'autres, tels **Dublin Core** et **Data Documentation Initiative (DDI)**, ne procurent que des champs pour stocker l'information, sans donner d'indications sur la formulation du contenu ou sa syntaxe.

Prenons l'exemple du roselin familier. Un ornithologue amateur désire entrer une observation de l'oiseau dans un dépôt qui utilise le schéma Darwin Core. Il devra remplir les boîtes suivantes :

| **Boîtes à remplir** | **Éléments du Darwin Core qui se cachent derrière** |
|---|---|
| Moment de l'observation | eventDate |
| Observateur | identifiedBy |
| Nom scientifique | scientificName |
| Règne | kingdom |
| Classe | class |
| Ordre | order |
| Famille | family |
| Genre | genus |

Il existe un grand nombre de schémas, certains généralistes, d'autres disciplinaires. Un schéma normé et utilisé

à large échelle peut être **compris par les machines**, ce qui augmente la visibilité et les possibilités de réutilisation des données décrites. Ces avantages sont perdus en créant un schéma de métadonnées maison.

En résumé, un schéma de métadonnées sert de structure et de contenant aux renseignements sur les jeux de données et, dans une certaine mesure, ajoute à leur signification.

## Règles de descriptions

Les règles de description permettent de standardiser, normaliser et structurer de l'information portant sur les jeux de données. Ces règles vont prescrire la transcription des renseignements, l'utilisation des majuscules, l'ordre ou la syntaxe des éléments. Les règles sont indépendantes des schémas et sont utilisables dans n'importe quel dépôt de données.

Pour illustrer, utilisons l'exemple de l'ornithologue amateur de roselins. Il cherche à savoir si cette espèce a été aperçue dans sa région à une date spécifique. Il consulte trois dépôts qui utilisent le schéma Darwin Core. En cherchant avec la date du 10 octobre 2021, il ne trouve de résultats que dans un seul des dépôts. Pourquoi? Parce que les dépôts utilisent des règles de descriptions différentes pour les dates. L'un n'a aucune exigence, soit le dépôt où l'entrée du 10 octobre 2021 est repérée; l'autre demande la norme ISO 8601, soit AAAA-MM-JJTHH:MM:SSZ et où la date est indiquée comme 2021-10-10; et le dernier veut la forme JJMMAAAA et où l'entrée souhaitée est représentée par 10102021.

Des règles de descriptions claires sont également très utiles au niveau des noms de personnes, particulièrement dans le cas de noms communs. Il faut éviter d'utiliser des initiales, des homonymes ou des pseudonymes. Déposer des données permet de donner de la visibilité aux chercheurs et chercheuses, mais pour ce faire, il faut pouvoir identifier sans ambiguïté la personne responsable des données!

Le nom n'est parfois pas suffisant pour faire la distinction entre les gens et c'est pourquoi il est recommandé d'utiliser également des **identifiants uniques pérennes** comme l'**ORCiD**.

## Vocabulaires contrôlés

Les **vocabulaires contrôlés** normalisent l'indexation et facilitent la recherche et le repérage d'informations. Il s'agit d'un ensemble de termes reconnus, normalisés et validés par un groupe ou une communauté de pratiques utilisés pour indexer ou analyser le contenu d'une ressource.

Si plusieurs termes désignent un même concept, un seul d'entre eux sera choisi et identifié comme le « terme préféré », les autres, considérés comme de possibles synonymes, seront mentionnés comme « termes rejetés ».

Revenons à l'ornithologue amateur qui, cette fois-ci, cherche de l'information sur le roselin dans un dépôt de

données anglophone. Les données de ce dépôt sont indexées avec du vocabulaire libre, mais également avec le FAST (*Faceted Application of Subject Terminology*). Afin de s'assurer de récupérer l'ensemble de l'information sur l'espèce, l'ornithologue cherche le terme « roselin » dans le RVMFAST, un vocabulaire qui fait les équivalences entre les termes français et anglais. Il découvre que « *House finch* » est le terme choisi par le FAST. Il effectue donc sa recherche dans le dépôt avec succès et récupère toutes les données disponibles!

Les thésaurus et les répertoires de vedettes-matière sont les exemples les plus répandus et les plus connus de vocabulaires contrôlés. Il existe des vocabulaires encyclopédiques, mais également des vocabulaires spécialisés propres à certaines disciplines, par exemple ERIC, un thésaurus spécialisé en éducation ou WORMS, un catalogue des noms d'organismes marins.

Plusieurs de ces vocabulaires sont multilingues, ou gèrent des équivalents linguistiques, ce qui est d'un apport précieux pour l'**interopérabilité**.

### Aller plus loin : ontologies

Une ontologie est une représentation théorique d'un domaine de connaissances dont les concepts sont liés par des relations sémantiques et logiques. Une ontologie comprend des vocabulaires, des définitions et une indication de la manière dont les concepts sont interdépendants entre eux. L'ontologie permet d'établir un ensemble de relations et de décrire des situations spécifiques dans un domaine donné. Une ontologie impose une structure sur le domaine et limite les possibles interprétations des termes. Plus simplement, l'ontologie permet d'offrir un langage commun à des blocs d'information liés entre eux. Elle est aux métadonnées ce que la grammaire est au langage.

Un des principaux avantages de l'utilisation d'une ontologie est l'interopérabilité, la réutilisation et le partage des métadonnées. La principale différence entre une ontologie et un vocabulaire contrôlé est que le vocabulaire contrôlé propose des relations sémantiques entre les éléments qui le composent, alors que l'ontologie proposera des relations fonctionnelles permettant de décrire précisément des situations.

Par exemple, dans un vocabulaire contrôlé, « roselin familier » est le terme préféré. Il est lié à « *Carpodacus* », qui est le terme général, ainsi qu'à « roselin du Mexique » et « *Carpodacus mexicanus* » qui sont deux termes rejetés. Dans une ontologie, « roselin familier » pourrait être lié grâce à la relation « habitat » aux termes « banlieue » et « semi-désert ». L'ontologie pourra également pointer vers la relation « alimentation » pour faire un lien entre le roselin et d'autres « granivore » et « insectivore ».

## Types de métadonnées

La séparation des métadonnées en différentes catégories est variable selon les sources. Les regroupements

suivants seront utilisés ici : métadonnées descriptives, structurelles, techniques, d'accès et de préservation. Les trois derniers types de métadonnées étant plus complexes, ils sont suggérés comme formation avancée.

Au-delà de ces catégories, les métadonnées peuvent également être classées par leur source (interne, externe), leur mode de création (manuel, automatique), leur statut (statique, dynamique), leur structure (structuré ou non) et d'autres caractéristiques. Pour plus d'information à ce sujet, veuillez consulter les ressources du présent chapitre.

## Métadonnées descriptives

Comme leur nom l'indique, les métadonnées descriptives servent à décrire une ressource afin d'en connaître le contenu et d'en assurer le repérage, que ce soit par l'humain ou par la machine. Le titre d'une œuvre, le nom du créateur d'une ressource ainsi que sa date de création sont des exemples de métadonnées descriptives qu'on retrouve dans des dépôts de données, des catalogues de bibliothèques ou des bases de données.

Dans le cas de données de recherche, les métadonnées descriptives font généralement référence aux champs à remplir dans les dépôts de données. Outre les métadonnées précédemment nommées, si les données ne sont pas versées dans un dépôt, un fichier texte, tel un LISEZ-MOI, peut servir de support aux métadonnées descriptives.

Les métadonnées de projet décrivent le « qui, quoi, où, quand et pourquoi » du jeu de données, ce qui fournit un contexte pour comprendre le but de la collecte, la méthodologie et de l'utilisation des données.

Les métadonnées de l'ensemble des données sont plus granulaires. Elles décrivent et contextualisent les données avec plus de détails, par exemple les variables, les unités de mesure, les observations. Ces renseignements peuvent également être présents avec les données elles-mêmes.

Les règles à respecter pour les métadonnées descriptives ne sont pas anodines. Mieux un jeu de données est décrit, plus il sera repérable et plus il sera simple d'attribuer le crédit aux bonnes personnes. En ce sens, l'utilisation d'identifiants uniques comme les DOI et les ORCiD ainsi que de vocabulaires contrôlés tels que le FAST et son équivalent francophone, le RVMFAST, permet de désambiguïser les gens et les objets numériques. La normalisation des métadonnées soutient également l'interopérabilité entre les systèmes.

Le meilleur moyen de profiter du pouvoir des métadonnées descriptives consiste à :

- utiliser des identifiants uniques lorsque possible;
- utiliser des schémas de métadonnées déjà existants, bien établis dans votre communauté de recherche;
- normaliser les métadonnées qui peuvent l'être (p. ex., noms, sujets, coordonnées géospatiales, date), idéalement avec des vocabulaires contrôlés; et

- respecter les conseils suggérés par les dépôts pour remplir leurs champs de métadonnées, soit les champs obligatoires, les champs recommandés et les champs facultatifs.

Chaque discipline utilise des métadonnées, des schémas, des ontologies et des vocabulaires contrôlés qui leur sont propres. Pour avoir quelques exemples de ces particularités, consultez les chapitres « La gestion des données quantitatives en sciences sociales » et « La gestion des données de recherche qualitatives ».

> **Fait intéressant :** plusieurs fichiers ont des métadonnées descriptives intégrées dans leur format. Avez-vous déjà regardé les propriétés des fichiers à partir d'un logiciel ou de votre système d'exploitation? Vous pourriez être surpris! Parfois, les logiciels remplissent automatiquement l'information « auteur » avec le nom du propriétaire du logiciel ou insèrent les coordonnées géographiques au fichier d'une photo prise avec un téléphone!

## Métadonnées structurelles

Les métadonnées structurelles permettent d'établir des liens entre les fichiers et à l'intérieur de ceux-ci. Il est autant question de la structure physique d'un fichier (les liens entre différentes parties de contenu) que de la structure logique d'un document (les liens entre des fichiers). Par exemple, vous pourriez avoir un article en PDF et les graphiques associés dans un fichier différent, en DOCX. Vous pourriez également avoir de l'information qui indique à quel endroit se situent le texte et les images dans une page ainsi que de l'information sur l'ordre des pages.

Certaines de ces métadonnées se génèrent automatiquement, d'autres doivent être entrées manuellement. Elles peuvent vous être utiles si vous êtes obligé de passer d'un format complexe à un format simple et que cela implique d'éclater vos données. Vous pourriez avoir à décrire les liens entre vos fichiers afin de représenter le format original. L'information peut être notée dans un fichier texte ou en utilisant du code.

Si vos fichiers ne sont pas indépendants ou qu'ils réfèrent à d'autres fichiers, ayez une pensée pour les métadonnées structurelles, car elles permettront la pleine compréhension de vos données.

## Aller plus loin : autres métadonnées

Les métadonnées descriptives et structurelles sont assez faciles à circonscrire, bien que leurs limites puissent être discutables. Les frontières sont plus floues lorsqu'il est question des métadonnées techniques, d'accès et

de préservation. Parfois, celles-ci sont regroupées sous le terme de « métadonnées administratives ». Les séparations ci-dessous sont utilisées à des fins d'explications uniquement.

La plupart des métadonnées ci-dessous se créent automatiquement à l'intérieur des fichiers et il n'est pas essentiel de les connaître. Vous pouvez modifier quelques-unes de ces métadonnées internes et certains logiciels permettent de les extraire pour les conserver séparément, mais il est recommandé de bien connaître les formats et les métadonnées avant de s'adonner à cette opération.

Comme indiqué précédemment, un changement de format peut être positif pour la préservation à long terme de vos fichiers. Une telle conversion peut avoir un impact sur les métadonnées internes du fichier. Les extraire du format original et les garder en accompagnement de l'objet numérique permet de documenter la provenance et l'authenticité de vos fichiers.

## Métadonnées techniques

Les métadonnées techniques sont très liées aux formats et la plupart sont intégrées à l'intérieur des fichiers. Elles documentent la création du fichier (p. ex., logiciel utilisé, version, système d'exploitation, date de création et de dernière modification) et des caractéristiques sur les objets numériques qui varient selon le type de format.

Exemples de métadonnées techniques :

- Pour le texte : l'encodage, la structure éventuelle en XML...
- Pour l'image : la résolution, le profil colorimétrique, la profondeur d'encodage...
- Pour le son : le débit, le codec, la fréquence d'échantillonnage...
- Pour la vidéo : le nombre d'images par seconde, le profil colorimétrique, la durée...
- Pour des contenus Web : le format déclaré dans l'en-tête, la réponse du serveur collecté...

L'extraction des métadonnées techniques aide à prouver qu'un format est bien ce qu'il prétend être. Elle permet aussi de se renseigner sur un objet numérique inconnu ou corrompu.

## Métadonnées d'accès et d'utilisation

Les métadonnées d'accès et d'utilisation comprennent de l'information qui permet à la communauté de recherche de télécharger des données et de les réutiliser en toute légalité.

Afin d'éviter les violations de droits, les métadonnées informent sur la provenance, les possibilités d'accès (p. ex., libre accès, embargo, formulaire de confidentialité) et d'utilisation (p. ex., libre, avec citation, consultation

uniquement). Vous pouvez également y trouver des **signatures numériques**. Du côté de l'administration des dépôts, ce sont ces métadonnées qui donnent la possibilité d'effectuer des actions de préservation en toute légalité.

### Métadonnées de préservation

Les métadonnées de préservation sont généralement liées à des schémas spécifiques comme METS ou PREMIS et aux actions effectuées sur les fichiers pour les préserver.

Elles regroupent tout ce qui touche à l'intégrité et à l'authenticité d'un objet numérique (voir le chapitre sur la préservation numérique). Minimalement, une **somme de contrôle** devrait être calculée. Avec les métadonnées de préservation, vous pouvez retracer toutes les modifications apportées à un fichier comme les changements de format, les vérifications des sommes de contrôle, les déplacements de supports physiques, etc. ainsi que ceux et celles qui ont effectué les changements.

## Conclusion

Le titre de ce chapitre renvoie à un monde fascinant pour de bonnes raisons. Les formats de fichiers et les métadonnées sont de vastes sujets dont nous n'avons fait qu'entrouvrir la porte. Il n'est toutefois pas essentiel de maîtriser les secrets de tous les formats de fichiers et de tous les vocabulaires contrôlés pour s'en sortir de façon respectable et avoir des données accessibles et utilisables des années après la fin d'un projet de recherche.

**Questions de réflexion**

*Un élément interactif H5P a été exclu de cette version du texte. Vous pouvez le consulter en ligne ici : https://ecampusontario.pressbooks.pub/gdrcanada/?p=47#h5p-3 (https://ecampusontario.pressbooks.pub/gdrcanada/?p=47#h5p-3)*

**Éléments clés à retenir**

- Le choix d'un format dépend de plusieurs facteurs, mais principalement des besoins et des capacités de ceux et celles qui les utilisent.
- Les meilleures données de recherche ne pourront être retrouvées et comprises y compris par ceux et celles qui les ont créées, sans métadonnées de qualité. La qualité est à privilégier sur la quantité.
- Faites des formats et des métadonnées vos alliés et non des obstacles, vous allez trouver en eux des amis un peu névrosés, mais fiables!

# Lectures et ressources supplémentaires

Corti, L., Van den Eynden, E., Bishop, L., Woollard, M., Haaker, M., et Summers, S. (2019). *Managing and sharing research data : a guide to good practice* (2e éd.). Sage.

## Formats

### Ressources canadiennes

Bibliothèque et Archives nationales du Québec. (2020, mars) *Guide concernant les formats recommandés par BAnQ.* https://numerique.banq.qc.ca/patrimoine/details/52327/4076856 (https://numerique.banq.qc.ca/patrimoine/details/52327/4076856)

Bieman, E. et Vinh-Doyle. W. (2019). *Stratégie de numérisation du patrimoine documentaire (SNPD) : Recommandations relatives aux formats de fichier pour la préservation numérique*. Gouvernement du Canada, Réseau canadien d'information sur le patrimoine. https://www.canada.ca/fr/reseau-information-patrimoine/services/preservation-numerique/recommandations-formats-fichier-preservation-numerique.html (https://www.canada.ca/fr/reseau-information-patrimoine/services/preservation-numerique/recommandations-formats-fichier-preservation-numerique.html)

Bibliothèque et Archives Canada. (2022). *Lignes directrices sur les formats de fichier à utiliser pour transférer des ressources documentaires*. https://bibliotheque-archives.canada.ca/fra/services/gouvernement-canada/

information-disposition/lignes-directrices-information/pages/lignes-directrices-formats-fichier-ressources-documentaires.aspx (https://bibliotheque-archives.canada.ca/fra/services/gouvernement-canada/information-disposition/lignes-directrices-information/pages/lignes-directrices-formats-fichier-ressources-documentaires.aspx)

Library and Archives Canada. (s.d.). *File Format Guidelines for Preservation and Long-term Access Version 1.0.* https://www.councilofnsarchives.ca/sites/default/files/LAC%20File%20Format%20Guidelines%20for%20Preservation%20and%20Long-term%20v1_2010-12_0.pdf (https://www.councilofnsarchives.ca/sites/default/files/LAC%20File%20Format%20Guidelines%20for%20Preservation%20and%20Long-term%20v1_2010-12_0.pdf)

## Autres ressources

Bibliothèque nationale de France. (s.d.). *Fiches formats*. https://github.com/hackathonBnF/FichesFormat/wiki (https://github.com/hackathonBnF/FichesFormat/wiki)

Caplan, P. (2008). What Is Digital Preservation? *Library Technology Reports*, *58*(2). https://journals.ala.org/index.php/ltr/article/view/4224/4809/ (https://journals.ala.org/index.php/ltr/article/view/4224/4809/).

Caplan, P. (dir.). (2010). Digital Preservation [Special issue]. *Information Standards Quarterly, 22(*2). https://www.niso.org/sites/default/files/2019-07/ISQ%20Spring%202010.pdf (https://www.niso.org/sites/default/files/2019-07/ISQ%20Spring%202010.pdf)

Centre de coordination pour l'archivage à long terme de document électroniques. (s.d.). *Catalogue des formats de fichiers pour l'archivage*. https://kost-ceco.ch/cms/kad_main_fr.html (https://kost-ceco.ch/cms/kad_main_fr.html)

Dappert, A. (2016). *Digital Preservation Metadata and Improvements to PREMIS in Version 3.0* [Présentation PowerPoint]. https://www.loc.gov/standards/premis/v3/tutorialslides.pdf (https://www.loc.gov/standards/premis/v3/tutorialslides.pdf)

Digital Preservation Coalition. (2015). *Digital Preservation Handbook* (2e éd.). https://www.dpconline.org/handbook (https://www.dpconline.org/handbook)

Digital Preservation Coalition. (s.d.). *Technology Watch Publications*. https://www.dpconline.org/digipres/discover-good-practice/tech-watch-reports (https://www.dpconline.org/digipres/discover-good-practice/tech-watch-reports)

Digital Preservation Coalition et Artefactual System. (2021). *Preserving Audio*. http://doi.org/10.7207/twgn21-11 (http://doi.org/10.7207/twgn21-11)

Digital Preservation Coalition et Artefactual System. (2021). *Preserving Databases*. http://doi.org/10.7207/twgn21-06 (http://doi.org/10.7207/twgn21-06)

Digital Preservation Coalition et Artefactual System. (2021). *Preserving Documents*. http://doi.org/10.7207/twgn21-07 (http://doi.org/10.7207/twgn21-07)

Digital Preservation Coalition et Artefactual System. (2021). *Preserving GIS*. http://doi.org/10.7207/twgn21-16 (http://doi.org/10.7207/twgn21-16)

Digital Preservation Coalition et Artefactual System. (2021). *Preserving Moving Images*. http://doi.org/10.7207/twgn21-12 (http://doi.org/10.7207/twgn21-12)

Digital Preservation Coalition et Artefactual System. (2021). *Preserving Raster Images*. http://doi.org/10.7207/twgn21-13 (http://doi.org/10.7207/twgn21-13)

Digital Preservation Coalition et Artefactual System. (2021) *Preserving Spreadsheets*. http://doi.org/10.7207/twgn21-09 (http://doi.org/10.7207/twgn21-09)

Federal Agencies Digital Guidelines Initiative. (s.d.). *Guidelines, File Format Comparison Projects*. https://www.digitizationguidelines.gov/guidelines/File_format_compare.html (https://www.digitizationguidelines.gov/guidelines/File_format_compare.html)

Federal Records Management. (s.d.). *Appendix A: Tables of File Formats*. National Archives and Records Administration. https://www.archives.gov/records-mgmt/policy/transfer-guidance-tables.html (https://www.archives.gov/records-mgmt/policy/transfer-guidance-tables.html)

Library of Congress. (s.d.). *Recommended Formats Statement*. https://www.loc.gov/preservation/resources/rfs/ (https://www.loc.gov/preservation/resources/rfs/)

Loftus, C. (2019, 23 août). File format identification: A student project at the University of Sheffield Library. *Digital Preservation Coalition*. https://www.dpconline.org/blog/file-format-identification-sheffi-uni (https://www.dpconline.org/blog/file-format-identification-sheffi-uni)

McLellan, E. P. (2007) *General Study 11 Final Report: Selecting Digital File Formats for Long-Term Preservation*. InterPARES 2 Project. http://www.interpares.org/display_file.cfm?doc=ip2_file_formats(complete).pdf (http://www.interpares.org/display_file.cfm?doc=ip2_file_formats(complete).pdf)

UK Data Service. (s.d.). *Recommended formats*. https://ukdataservice.ac.uk/learning-hub/research-data-management/format-your-data/recommended-formats/ (https://ukdataservice.ac.uk/learning-hub/research-data-management/format-your-data/recommended-formats/)

Vitam. (2020). *Identification des formats de fichier*. https://www.programmevitam.fr/ressources/DocCourante/autres/fonctionnel/20200131_NP_Vitam_preservation-identification-format-v2.0.pdf (https://www.programmevitam.fr/ressources/DocCourante/autres/fonctionnel/20200131_NP_Vitam_preservation-identification-format-v2.0.pdf)

## Jeux sur les formats

Archives & Records Association. (2022). *File Format or Fake?* https://www.exploreyourarchive.org/archives/digital-preservation/ (https://www.exploreyourarchive.org/archives/digital-preservation/).

Fortin, É. et Ruest, J.-F. (2022). *Mille Formats*. Bibliothèque de l'Université Laval. https://www5.bibl.ulaval.ca/formations/tutoriels-en-ligne/autres-tutoriels/mille-formats (https://www5.bibl.ulaval.ca/formations/tutoriels-en-ligne/autres-tutoriels/mille-formats).

# Métadonnées

Alliance de recherche numérique du Canada. (2021). *RDM and Metadata for Discovery: What's in it for researchers?* [Vidéo]. YouTube. https://youtu.be/4fjPBSKMPlw (https://youtu.be/4fjPBSKMPlw)

Baca, M. (dir.). (2016) *Introduction to Metadata* (3e éd.). Getty Publications. http://www.getty.edu/publications/intrometadata/ (http://www.getty.edu/publications/intrometadata/).

Bascik, T., Boisvert, P., Cooper, A., Gagnon, M., Goodwin, M., Huck, J., Leahey, A., Stathis, K. et Steeleworthy, M. (2021). *Guide des pratiques exemplaires sur les métadonnées de Dataverse Nord v 3.0* (Version 3). Zenodo. https://doi.org/10.5281/zenodo.5668962 (https://doi.org/10.5281/zenodo.5668962)

Bibliothèque Université Laval. (s.d.). *RVMFAST*. https://rvmweb.bibl.ulaval.ca/rvmfast/rechercheSimple.do (https://rvmweb.bibl.ulaval.ca/rvmfast/rechercheSimple.do).

Canning, E., Brown, S., Roger, S. et Martin, K. (2022). The Power to Structure: Making Meaning from Metadata Through Ontologies. *KULA: Knowledge Creation, Dissemination, and Preservation Studies, 6*(3). https://doi.org/10.18357/kula.169 (https://doi.org/10.18357/kula.169)

DoRANum. (s.d.). *Métadonnées, standards, formats : comment décrire les données?* https://doranum.fr/metadonnees-standards-formats/ (https://doranum.fr/metadonnees-standards-formats/).

Dublin Core. https://www.dublincore.org/ (https://www.dublincore.org/).

ERIC. https://eric.ed.gov/ (https://eric.ed.gov/).

Guenther, R. (2017). *Metadata for Digitization and Preservation. Part 1: Metadata schemes* [Présentation PowerPoint]. Lyrasis.

Lacroix, C. (2017). *Meilleures pratiques de gestion des métadonnées décrivant les données de recherches* [Présentation]. Bureau de Coopération Interuniversitaire. https://libguides.pbuq.ca/ld.php?content_id=36275448 (https://libguides.pbuq.ca/ld.php?content_id=36275448)

OCLC FAST. https://fast.oclc.org/ (https://fast.oclc.org/)

ORCiD. https://orcid.org/ (https://orcid.org/)

Research Data Management Service Group. (s.d.). *Guide to writing "readme" style metadata*. Cornell University. https://data.research.cornell.edu/content/readme (https://data.research.cornell.edu/content/readme)

RDA. (2017, 24 juillet). *Supporting public procurement in Europe – 4 RDA Recommendations for open data sharing now published as ICT Technical specifications*. https://www.rd-alliance.org/node/57123 (https://www.rd-alliance.org/node/57123)

UK Data Archives. (s.d.). *Standards and procedures*. https://www.data-archive.ac.uk/managing-data/standards-and-procedures/ (https://www.data-archive.ac.uk/managing-data/standards-and-procedures/)

WORMS: World Register of Marine Species. https://www.marinespecies.org/ (https://www.marinespecies.org/)

## À propos de l'auteur

### Émilie Fortin

Émilie Fortin est bibliothécaire à la gestion des données de recherche et à la préservation numérique à l'Université Laval depuis 2021. Auparavant, elle occupait le poste de responsable de la production numérique, préservation et conservation des collections. Elle a complété sa maîtrise en science de l'information de l'Université de Montréal en passant une année à la Haute école de gestion de Genève. Impliquée dans le Groupe d'experts de l'Alliance de recherche numérique sur la planification de la gestion des données, dans le groupe de travail sur la gestion des données de recherche du Partenariat des bibliothèques universitaires du Québec (PBUQ), elle participe également régulièrement aux conférences de l'iPRES sur la préservation numérique. ORCID: 0000-0002-9717-6840 (https://orcid.org/0000-0002-9717-6840)

10.

# SOUTENIR LA RECHERCHE REPRODUCTIBLE AVEC LA CURATION ACTIVE DE DONNÉES

Sandra Sawchuk; Louise Gillis; et Lachlan MacLeod

**Objectifs d'apprentissage**

À la fin de ce chapitre, vous pourrez:

- Comprendre le rôle de la curation active de données dans le domaine plus large de la gestion des données de recherche.
- Identifier les caractéristiques principales des outils de gestion active des données, tels que le versionnage, la rédaction de scripts, les conteneurs informatiques et les machines virtuelles.
- Évaluer un exemple d'un jeu de données reproductible dans un conteneur.

## Introduction

Ce chapitre mettra l'accent sur les aspects **interopérables** et réutilisables du modèle FAIR (facile à trouver, accessible, interopérable et réutilisable) qui ont été présentés dans le chapitre 2, « Les principes FAIR et la gestion des données de recherche. » Ainsi, vous obtiendrez la confiance et les compétences nécessaires pour entreprendre la curation active des données.

La curation active des données appliquée en cours de recherche produit des données qui sont FAIR : faciles à trouver, accessibles, interopérables et réutilisables (Johnston *et al*., 2017a; Wilkinson *et al*., 2016). Le terme « active » décrit les pratiques de curation appliquées pendant les étapes de la collecte, de l'analyse et de la diffusion des données d'une recherche. La curation des données implique la gestion des données de recherche

sélectionnées ou qui doivent être déposées pour le stockage et la préservation à long terme. (Krier et Strasser, 2014). Règle générale, la curation est abordée vers la fin d'un projet, souvent une fois l'analyse complétée. D'excellentes ressources, telles que le *Guide pour la curation dans Dataverse (https://doi.org/10.5281/zenodo.5579827)* ainsi que le flux de travail *CURATED* (https://datacurationnetwork.org/outputs/workflows/) (en anglais uniquement) du Data Curation Network, fournissent de précieux conseils pour la curation lorsque la phase active du projet est terminée (Cooper *et al*., 2021; Johnston *et al*., 2017b). Il y a des avantages à travailler sur la curation tout au long du projet. Cela permet de repérer les erreurs avant qu'elles ne deviennent catastrophiques et les données peuvent aussi être mieux décrites et contextualisées (Sawchuk et Khair, 2021).

Ce chapitre vous informera sur les outils et les techniques qui facilitent la curation des données de recherche tout au long des phases actives d'une recherche. Comme Cooper *et al*. (2021), nous savons que la capacité d'appui à la curation peut varier d'un établissement canadien à l'autre et que les bibliothèques ont souvent le rôle d'éduquer et de sensibiliser sur les meilleures pratiques. Quant à la gestion concrète et quotidienne de la recherche et des données associées au projet, elle tombe sous la responsabilité des chercheuses et chercheurs qui effectuent le travail.

Nous discuterons des stratégies pour la mise en place de bonnes pratiques de gestion des données en misant sur les activités qui contribuent à l'amélioration de l'interopérabilité et de la reproductibilité des données. Nous examinerons également les meilleures pratiques en matière de curation des données de recherche, y compris les outils pour la communication et la collaboration. Même si les outils présentés dans ce chapitre sont généralement utilisés pour appuyer la recherche computationnelle, les principes de reproductibilité que nous présenterons pourront servir à toutes les disciplines.

## Les plateformes

Le choix d'une plateforme de stockage des données n'est pas considéré comme de la curation. Toutefois, le choix d'une plateforme de stockage plutôt qu'une autre peut avoir des conséquences importantes sur la curation.

Les options de stockage sont abordées plus en détail dans le chapitre 5, « Partage et réutilisation des données de recherche au Canada, » mais voici un survol rapide. Votre choix de plateforme s'inscrit dans l'une ou l'autre des trois catégories suivantes :

- Le stockage local est intégré ou se connecte directement à vos appareils. Cela inclut les disques durs et les clés USB;
- Les unités de stockage en réseau (NAS ou *Network Attached Storage*) relient les appareils d'un même réseau local. Des exemples comprennent les serveurs départementaux, de facultés ou universitaires;

- Le stockage infonuagique est un service en ligne fourni par une tierce partie. Des exemples comprennent Dropbox, Google Drive, OSF et OneDrive.

Le tableau 1 énumère les avantages et désavantages de chacun de ces types de plateformes. Chacun a son utilité propre, mais somme toute, l'infonuagique semble offrir des fonctionnalités bien adaptées à la curation.

**Tableau 1. Comparaison des plateformes de stockage**

| | **Avantages** | **Désavantages** |
|---|---|---|
| Local | 1. Aucune connexion Internet nécessaire<br>2. Peu coûteux<br>3. Protection contre les accès non autorisés | 1. Potentiel de perte, de corruption ou de dommages attribuables aux défaillances matérielles, désastres naturels (incendie et inondation) et au vol<br>2. Ne facilite pas la collaboration ou le partage de fichiers |
| Réseau | 1. Espace de travail collaboratif<br>2. Accès à distance<br>3. Sauvegardes automatisées<br>4. Bonne sécurité | 1. Nécessite un accès Internet<br>2. N'est pas accessible aux partenaires externes<br>3. Coûteux |
| Infonuagique | 1. Le contrôle des versions et la récupération des fichiers sont automatisés<br>2. Sauvegardes automatisées<br>3. Espace de travail collaboratif<br>4. Accès à distance | 1. Les politiques de confidentialité varient selon les fournisseurs<br>2. Absence de contrôle sur l'emplacement du stockage des données<br>3. Danger de piratage, logiciels malveillants et hameçonnage |

Les données personnelles de santé sont assujetties à des législations qui empêchent leur

stockage à l'extérieur du Canada. Les données personnelles pouvant identifier des personnes participantes ne doivent pas être stockées sur des plateformes infonuagiques dont la prise en charge n'est pas institutionnelle.

## Lignes directrices pour le stockage des données

1. Lorsque possible, privilégiez l'utilisation d'une plateforme infonuagique et faites la sauvegarde de vos données sur un réseau institutionnel. La plupart des plateformes infonuagiques comportent des fonctionnalités qui gèrent le versionnage de façon automatique. L'automatisation implique moins de travail pour vous et moins de risques d'erreurs humaines. Les fichiers importants peuvent être copiés sur le réseau de l'établissement, régulièrement sauvegardé, protégeant davantage contre la perte de données qui pourrait survenir sur des disques locaux.

   À chaque fois que vous faites une modification dans un environnement infonuagique, une nouvelle version de votre fichier est sauvegardée avec l'information sur la provenance du fichier:

   - qui a fait la modification;
   - quand la modification a été faite;
   - qu'est-ce qui a été modifié.

2. Choisissez une solution soutenue par votre établissement. Vous aurez ainsi accès au soutien technique, à la formation et à l'assurance qu'elle a été évaluée. Le choix d'une solution bien supportée permet d'augmenter les probabilités que vos données soient accessibles et utilisables à long terme. Au Canada, Microsoft Office 365 peut s'avérer un bon choix puisque de nombreuses universités utilisent cette suite de logiciels.

3. Utilisez un **cahier de laboratoire électronique** ou un outil de gestion de projet. Les cahiers de laboratoire électroniques sont des outils en ligne conçus selon le design et l'utilisation de leurs équivalents papier. Dans leur forme la plus simple, ils fournissent un espace où consigner les protocoles de recherches, les observations, les notes et d'autres données reliées au projet. Leur format électronique permet une bonne gestion des données évitant des problématiques telles qu'une écriture manuscrite difficile à déchiffrer et des pertes de données liées à des dommages matériels. Les cahiers de laboratoire électroniques assurent également la sécurité des données et permettent la collaboration. Ils peuvent être

particulièrement utiles si vous travaillez dans le secteur privé ou dans des contextes où les membres de l'équipe de recherche proviennent de différents établissements. Au-delà des solutions institutionnelles, vous pouvez vous tourner vers des outils de collaboration comme Open Science Framework (OSF), une application en code **source libre** dont l'utilisation est gratuite et qui fournit des détails sur la provenance des fichiers. OSF peut être utilisé en tant qu'espace collaboratif de partage des données, ou en tant que cahier de laboratoire électronique.

# La sécurité des données

Identifiez les risques potentiels dans votre plan de **gestion des données de recherche** (GDR) et assurez-vous que la mise en œuvre des mesures élaborées est réalisable et répond bien au risque associé à vos données. Si vous travaillez avec des données personnelles de santé par exemple, vous devez faire davantage attention que si vous travaillez avec du code source libre. Des considérations similaires doivent être gardées à l'esprit si vous travaillez avec des données issues de groupes marginalisés ou racisés. Votre choix de plateforme de stockage est également important. Les données stockées sur des clés USB peuvent être perdues ou endommagées, tandis que celles sur un stockage infonuagique peuvent être la proie de piratage, de logiciels malveillants et d'hameçonnage.

## Lignes directrices dans le traitement de la sécurité des données

1. Évitez l'utilisation de disques externes ou de stockage local.
2. Sécurisez votre ordinateur et vos réseaux en installant les mises à jour logicielles et les protections antivirus, en activant les pare-feux et en verrouillant votre ordinateur et autres appareils quand vous ne les utilisez pas.
3. Utilisez des mots de passe robustes. Les mots de passe robustes sont uniques et complexes (de longues chaînes de caractères qui utilisent une combinaison de symboles, de chiffres et de lettres majuscules et minuscules). Malheureusement, ils sont aussi difficiles à mémoriser. Une solution est d'utiliser un **gestionnaire de mots de passe**, tel que KeePassX (https://www.keepassx.org) ou 1Password (https://1password.com), qui enregistre vos noms d'utilisateurs et mots de passe dans un même endroit. Modifiez régulièrement vos mots de passe!
4. Chiffrez vos fichiers et disques si vous travaillez avec des **données sensibles** ou propriétaires. Avec des ordinateurs Mac, vous pouvez utiliser Firevault (https://support.apple.com/en-us/HT204837) et avec des PC, Bitlocker (https://docs.microsoft.com/en-us/windows/security/information-protection/bitlocker/bitlocker-overview).
5. Si vous travaillez sur une plateforme infonuagique, utilisez l'**authentification multifactorielle** pour

accéder aux fichiers.

6. Lorsque vous transférez des données, utilisez le chiffrement. OneDrive est un exemple de plateforme de stockage qui permet l'envoi et la réception de fichiers chiffrés. Le transfert de fichiers par Globus (https://www.globus.org/data-transfer) est une bonne option pour des fichiers lourds et de nombreux établissements de recherche utilisent Globus pour les données de recherche sensibles.

# La curation active des données

La curation active des données implique l'organisation, la description ainsi que la gestion de vos fichiers de recherche et de leur documentation. La façon dont vous organisez vos fichiers est un choix personnel. Il n'y a pas une seule et unique façon de faire et la meilleure façon de travailler sera celle qui conviendra le mieux à vous et à votre équipe. Mettez vos décisions par écrit, communiquez-les à toutes les personnes concernées et réévaluez régulièrement vos choix. Si une stratégie ne s'avère plus efficace, modifiez-la et passez à autre chose.

> Vous n'avez pas à développer de toutes pièces votre structure organisationnelle ! Des ressources telles que le protocole TIER (https://www.projecttier.org/tier-protocol/protocol-4-0) (en anglais uniquement) peuvent servir comme bon point de départ.

## Lignes directrices pour la curation active des données

### 1. L'organisation des fichiers de recherche

- Identifiez une personne qui prendra en charge l'organisation et le nommage des fichiers. Cette personne peut mener des vérifications ponctuelles pour s'assurer que la documentation, le nommage et les chemins d'accès aux fichiers sont uniformément établis. Elle peut aussi servir de contact principal pour tous les membres de l'équipe de recherche qui auraient des questions sur les pratiques organisationnelles ou voudraient signaler des erreurs dans les données.
- Gardez votre plan organisationnel, la structure et les conventions de nommage des fichiers dans un seul document que vous pourrez imprimer et conserver près de votre ordinateur de travail ou dans un fichier de documentation avec votre projet. Si les documents restent à la portée, ils seront plus facilement consultés que s'ils sont difficiles à trouver.
- Mettez en place des processus de travail clairs pour vous assurer que le travail ne soit pas modifié ou

détruit. « Protégez vos données originales en verrouillant vos fichiers (avec un mot de passe) ou en définissant un accès en lecture seule » (Groupe d'experts sur la formation, 2020) pour ensuite les compresser. Créez des espaces de travail distincts pour les différentes équipes de travail avec une personne centrale qui sera responsable de la coordination et de regrouper les différents éléments. Autre option : quand le projet et le calendrier le permettent, faites travailler les équipes selon un horaire régulier, sans chevauchement. Utilisez un diagramme de Gantt ou un modèle semblable pour établir un calendrier de projet et gérer les tâches.

- Organisez avec simplicité. Limitez la quantité de dossiers utilisés. Ainsi, vous trouverez vos données plus facilement et réduirez le temps de traitement pour les sauvegardes et la fusion ou l'analyse d'importants jeux de données.

Saviez-vous que? Les dates dans le format ISO 8601 (https://www.iso.org/fr/iso-8601-date-and-time-format.html) sont lisibles par machine et peuvent être triées de façon chronologique.

## 2. La description des fichiers

- Utilisez une convention de nommage qui s'applique à tous les fichiers et créez un document qui explique cette convention. Vous pourrez ainsi prévenir les erreurs et réduire le temps de formation des membres de l'équipe de recherche. Le document servira aussi de base à votre dictionnaire de données (décrit ci-dessous). Il est utile d'y inclure les abréviations ou acronymes des noms de projets, les organismes subventionnaires, les numéros des subventions, le type de contenu, etc. Ajoutez-y des dates (nous recommandons le format AAAA-MM-JJ) et de brèves descriptions. Comme séparateur, utilisez la **notation chameau** (NotationChameau) ou le trait de soulignement (trait_de_soulignement). Les systèmes informatiques ne reconnaissent pas toujours les espaces et caractères spéciaux.
- Le **versionnage** devrait être clair et effectué de façon judicieuse. Il n'est pas nécessaire de créer de nouvelles versions pour chaque modification; la mise à jour des numéros de version ne s'impose que lorsque d'importants changements sont apportés au fichier. Utilisez V01, 02, etc. pour rendre l'historique des versions clair et facile à suivre, ou utilisez un système automatisé de contrôle des versions.
- Les fichiers de syntaxe sont des fichiers de code qui contiennent les séquences d'actions effectuées par le logiciel d'analyse statistique; ils peuvent être générés par le logiciel ou codés par les analystes. Effectuez ou enregistrez toutes vos actions en utilisant un fichier de syntaxe qui fait la liste des actions prises par le logiciel d'analyse statistique. Selon le logiciel utilisé, les fichiers de syntaxe peuvent aussi s'appeler des fichiers de programme, des fichiers script ou une autre désignation semblable. La plupart des éditeurs de

syntaxe possèdent une fonctionnalité intégrée de notation (ou de commentaire) qui vous aide à vous rappeler ce que vous avez fait et à communiquer ce procédé à vos partenaires de recherche. Ajoutez des descriptions de ce que vous avez fait dans les fichiers de syntaxe et nettoyez la syntaxe au fur et à mesure. Le procédé peut aussi être utile si votre code est utilisé pour de futurs projets ou s'il est publié dans un dépôt de données de recherche.

- Si vous utilisez des logiciels spécialisés pour l'exploration et l'analyse des données, déterminez si la documentation sur le traitement des fichiers de données est générée de façon automatique et complétez-la au besoin. Inscrivez tous les détails dont vous aurez besoin pour reproduire votre flux de travail. Si vous prévoyez de revoir vos données ultérieurement, vous serez reconnaissant de l'effort que vous y aurez mis!

Créez votre propre convention de nommage des fichiers. La feuille de travail (https://resolver.caltech.edu/CaltechAUTHORS:20200601-161923247) (en anglais uniquement) de Krista Briney pour l'établissement de conventions de nommage des fichiers vous guidera tout au long du processus de création d'un plan concret.

## 3. La création de guides de codification et de dictionnaires de données

Un **guide de codification** est un document qui fait la description d'un jeu de données, y compris les détails sur son contenu et sa conception. Un **dictionnaire de données** est un document semblable au guide de codification, lisible et souvent exploitable par une machine, qui contient généralement des informations détaillées sur la structure technique d'un jeu de données en plus de ses contenus (Buchanan *et al*., 2021). Toutefois, les deux termes sont souvent utilisés de façon interchangeable. Le guide de codification peut être généré de façon automatique par le logiciel de statistiques que vous utilisez ou vous pouvez avoir à le créer vous-même. Développer le guide de codification au fur et à mesure est une bonne pratique qui permet de standardiser les données. Documentez toute modification au code ou toute autre modification aux données. Même si le guide de codification est généré par le logiciel, vous devrez probablement y ajouter des informations supplémentaires. Idéalement, votre guide de codification sera simple, inclura le nom des variables et de brèves descriptions. Toutefois, selon le Inter-university Consortium for Political and Social Research (ICPSR, 2023), les informations contenues dans les guides de codification peuvent varier d'un projet à l'autre et d'un domaine à l'autre.

Vous devriez inclure un guide de codification dans la section méthodologique de l'étude. Comme point de

départ, documentez toute analyse effectuée sous forme de notation dans le fichier syntaxe. Un fichier syntaxe bien annoté peut devenir la base de votre guide de codification ou même de la section méthodes d'un rapport, d'une thèse ou d'une publication. Les descriptions méthodologiques varieront considérablement d'un domaine d'étude à l'autre, mais certains éléments clés peuvent toujours être inclus :

- Valeurs et étiquettes de tous les champs
  - Inclure une description de la façon dont les valeurs nulles ont été traitées au cours de l'analyse;
- Descriptions ou distributions sommaires des résultats;
- Variables omises ou éliminées;
- Rapport entre les variables, y compris le **chaînage** (quand du texte est inséré de façon automatique par le logiciel de sondage selon les réponses précédentes) ou des expériences subséquentes.

La figure 1 montre l'extrait d'un guide de codification publié par Statistiques Canada pour l'Enquête nationale sur la santé de la population. Dans cet exemple, le guide de codification contient le nom de la variable, la question du sondage et ses réponses ainsi qu'une note sur l'âge des répondants. Ce guide de codification contient aussi la position et la longueur de la variable; cette information serait également incluse dans le dictionnaire des données.

**QUESTIONS SUPPLÉMENTAIRES**
**à l'enquête nationale sur la santé de la population**

septembre 1996 Page 1.8

*Variable:* **UT_Q1** *Position:* 33 *Longeur:* 1

Dans les 12 derniers mois, avez-vous passé la nuit comme patient à l'hôpital, etc.?

| | | FREQ | POND |
|---|---|---|---|
| 1 | OUI | 1,423 | 2,269,617 |
| 2 | NON | 11,976 | 21,677,619 |
| 6 | SANS OBJET | 0 | 0 |
| 9 | NON DÉCLARÉ | 1 | 1,367 |

*Nota:* Répondants agés de 12 ans ou plus

Figure 1. Manuel des codes A – Enquête nationale sur la santé de la population (ENSP) – 1994-1995 – Questions supplémentaires (Statistiques Canada, 1996).

# Pour aller plus loin

Peu importe le logiciel que vous choisissez d'utiliser, une documentation complète constitue la clé d'une

gestion et d'une curation efficaces des données. Cette section présentera des concepts importants dont il faut tenir compte pour la curation active de la **recherche computationnelle**, y compris le versionnage des fichiers, la rédaction de scripts et les conteneurs informatiques.

Nous pouvons appliquer les leçons de la curation active des données à la recherche computationnelle. Les ordinateurs sont devenus tellement plus simples et conviviaux à utiliser qu'il est facile d'oublier leur complexité. Les chercheuses et chercheurs ont le choix d'une variété de logiciels libres ou propriétaires pour effectuer des tâches à chacune des étapes de leurs projets, allant de la collecte à la visualisation des données.

Les logiciels propriétaires, tels que SPSS ou Microsoft Excel, sont souvent comparés à des boîtes noires où les données entrent et ressortent avec peu d'indications sur ce qui s'est passé à l'intérieur (Morin *et al*., 2012). Selon les conditions d'utilisation, l'inspection du code peut être interdite ou impossible. Les logiciels propriétaires sont souvent plus faciles à utiliser que les logiciels libres et ils sont parfois gratuits, parfois payants (Singh *et al*., 2015). Les logiciels libres sont souvent gratuits, mais ils peuvent être plus complexes à utiliser (Cox, 2019). Cette complexité est contrebalancée par la possibilité d'inspecter le code source et de modifier le programme lui-même selon ce qu'autorisent les licences des logiciels (Singh *et al*., 2015).

> Un logiciel est un ensemble d'instructions textuelles qui exécute ou qui fonctionne avec l'aide d'un ordinateur. Les instructions sont assujetties à des règles articulées par le langage de programmation particulier dans lequel le logiciel a été écrit et l'exécution de ce code dépend de l'environnement informatique, qui regroupe des éléments tels que le matériel informatique et le système d'exploitation (Possati, 2020).

## Le contrôle programmatique des versions de fichiers

Tel que discuté plus tôt dans le chapitre, la curation active des données implique beaucoup plus que la création d'une arborescence de dossiers et l'utilisation de pratiques uniformisées de nommage de fichiers. Vous devez également gérer le contenu des fichiers de façon systématique et avec transparence en gardant en tête leur réutilisation. Vous pouvez y arriver par le biais de la programmation en utilisant des fonctionnalités automatisées de **gestion des versions** disponibles dans de nombreux gestionnaires de documents infonuagiques tels que Office 365 et Google Docs. L'activité d'évaluation à la fin du chapitre est hébergée sur une plateforme de gestion des versions appelée GitHub qui est couramment utilisée par celles et ceux qui écrivent et développent du code.

La gestion des versions, ou versionnage implique le suivi des modifications apportées à un fichier, peu importe l'ampleur des modifications. Lorsque les fichiers sont sauvegardés par le contrôle automatique des versions, tant le contenu que les révisions sont automatiquement sauvegardés, permettant ainsi aux utilisateurs de revenir à toutes les sauvegardes antérieures du fichier (Vuorre et Curley, 2018). Dès que vous sauvegardez un fichier, chacune des modifications au fichier est enregistrée;  le fichier est sauvegardé sous une nouvelle version sans avoir à lui attribuer un nouveau nom de fichier. Vous pourrez donc « revenir dans le temps » pour voir de quelle façon le fichier s'est développé puisque toutes ses modifications seront identifiées.

Des dépôts de données, tels que Dataverse et Zenodo, incluent des informations sur les versions dans les citations qu'ils génèrent, permettant ainsi à tout un chacun d'identifier laquelle des versions d'un jeu de données ou d'un manuscrit a été utilisée.

Le présent chapitre s'est surtout penché sur les projets dont les données sont créées par les chercheuses et chercheurs eux-mêmes. Pour les projets qui utilisent des données secondaires, il est essentiel d'accorder une attention particulière à la provenance. Arguillas *et al.* (2022) ont publié un excellent guide sur la curation et la reproductibilité qui comprend une discussion sur cet enjeu important.

## La rédaction de scripts: pour rendre l'analyse reproductible et automatiser les processus de gestion des données

Automatiser les processus de travail de recherche, tels que l'importation, le nettoyage et la visualisation des données, vous permet de mettre en œuvre vos expériences computationnelles avec un minimum d'intervention manuelle. L'automatisation dépend des scripts, c'est-à-dire des ensembles de routines informatiques transcrites en code (Alston et Rick, 2021; Rokem *et al*., 2017). Les scripts devraient être accompagnés d'une documentation détaillée qui fait la description de chacune des étapes de la routine afin que la **provenance** d'une expérience puisse être comprise. La provenance en recherche computationnelle a la même signification que la provenance en archivistique; il s'agit d'un enregistrement de la source, de l'historique et de la propriété d'un artéfact. Dans le contexte présent, l'artéfact est de nature informatique.

L'automatisation et le suivi de la provenance facilitent la reproductibilité et la réutilisation pour les chercheuses et chercheurs et les évaluatrices et évaluateurs externes, mais les plus grands bénéficiaires seront toujours les membres de l'équipe de recherche originale (Rokem *et al*., 2017; Sawchuk et Khair, 2021). Une documentation détaillée permet d'identifier les erreurs et fournit de précieuses informations en matière de

contexte pour la formation de nouveaux membres de l'équipe. L'automatisation permet de mener l'expérience à une ou plusieurs reprises avec un minimum d'effort, ce qui peut être particulièrement utile quand les jeux de données ont été modifiés ou mis à jour.

Dans certains cas, l'automatisation et la provenance peuvent avoir lieu au même endroit. Tel que déjà mentionné, les fichiers de syntaxe comprennent les commandes utilisées pour la manipulation, l'analyse et la visualisation des données; ces fichiers peuvent aussi être modifiés pour inclure des commentaires qui décrivent le raisonnement et l'analyse du projet. Les fichiers de syntaxe peuvent ensuite être regroupés avec les données et les fichiers de sortie, permettant à d'autres d'évaluer et de réutiliser le projet dans son ensemble.

Des cahiers électroniques de code sont d'autres outils qui intègrent l'automatisation et le suivi de la provenance à l'intérieur d'un seul document linéaire. Un cahier de codes, tel que Jupyter Notebook (https://jupyter.org), est une interface qui encourage la pratique de la **programmation lettrée**, là où le code, les commentaires et les sorties s'affichent ensemble de façon linéaire, telle une œuvre de littérature (Hunt et Gagnon-Bartsch, 2021; Kery *et al*., 2018).

Une bonne documentation est essentielle pour la **reproductibilité de la recherche**, peu importe la personne qui réutilise les données (Benureau et Rougier, 2018). Une bonne pratique consiste à inclure des annotations descriptives avec toutes les ressources informatiques d'un projet afin de fournir un contexte précieux à toutes les étapes du cycle de vie de la recherche.

## Le partage du code: les cahiers électroniques et les conteneurs informatiques

Le fonctionnement du code n'est pas garanti d'un ordinateur à l'autre. Des différences au niveau du matériel informatique, des systèmes d'exploitation, des logiciels installés et des privilèges administratifs créent des obstacles au fonctionnement ou à la lecture du code utilisé pour mener l'analyse des données. Des chercheuses ou chercheurs peuvent utiliser des formats de fichiers propriétaires qui ne sont accessibles qu'en achetant des logiciels particuliers ou en s'y abonnant. De plus, la **littératie en matière de codage** varie souvent chez les personnes qui mènent et gèrent les projets de recherche, entraînant des incohérences dans la documentation et l'inclusion d'erreurs (Hunt et Gagnon-Bartsh, 2021). Bien qu'il soit recommandé de partager les données de recherche et le code dans un dépôt qui facilite le versionnage, vous devriez également prendre des mesures concrètes au cours de la phase active d'un projet de recherche pour encourager la reproductibilité et la réutilisation.

Il existe plusieurs solutions techniques qui facilitent le partage du code et qui varient en complexité sur un spectre allant du statique au dynamique. L'approche statique pour le partage du code est simplement de téléverser le code brut vers un dépôt avec un fichier **LISEZ-MOI** bien documenté ainsi qu'une liste des

dépendances ou des exigences au niveau de l'environnement informatique. Dans l'approche dynamique, les données, le code et les dépendances sont rassemblés dans un format autonome appelé un conteneur (Hunt et Gagnon-Bartsh, 2021; Vuorre et Crump, 2021).

Un **conteneur informatique** est comme un ordinateur autonome virtuel à l'intérieur d'un ordinateur. Les conteneurs peuvent être hébergés sur un service Web tel que Docker (https://www.docker.com/resources/what-container/) ou sur une clé USB. Ils comprennent tout ce qui est nécessaire pour faire fonctionner un logiciel (y compris le système d'exploitation), sans avoir à télécharger et installer des programmes ou des données. La mise en conteneur facilite la reproductibilité informatique, ce qui survient quand les aspects informatiques d'un projet de recherche peuvent être indépendamment reproduits par un tiers (Benureau et Rougier, 2018). Pour qu'un projet puisse être complètement reproductible, tous les produits de la recherche – en partant des données jusqu'au code et à l'analyse – doivent être inclus. Voilà pourquoi les conteneurs informatiques comprennent des informations détaillées sur l'environnement numérique utilisé pour mener la recherche (Hunt et Gagnon-Bartsch, 2021). Il s'agit notamment d'informations sur le type d'ordinateur et le système d'exploitation (p. ex., Mac OS Monterey v12.3, Windows v11, Linux Ubuntu v21.10); le nom et la version de tout logiciel commercial utilisé pour la collecte ou l'analyse des données; ou inversement, le langage de programmation utilisé pour créer le logiciel ainsi que le nom et le numéro de version de toute dépendance prise en charge par le logiciel.

Une **dépendance** est une bibliothèque de logiciels supplémentaires qui peut être téléchargée à partir d'Internet et utilisée pour certaines tâches précises de programmation. Par exemple, les personnes qui utilisent le langage de programmation Python peuvent aller en ligne et télécharger des ensembles complets de code déjà écrit pour certaines opérations spécifiques telles que la création de graphiques mathématiques ou l'analyse de texte. Les dépendances sont écrites et revues par des gens extérieurs au projet, ce qui signifie que les versions peuvent être mises à jour, soit régulièrement, soit pas du tout. Certaines dépendances sont plus largement utilisées et viennent avec beaucoup de documentation, tandis que d'autres ne le sont pas. Il est de la responsabilité de la chercheuse ou du chercheur de vérifier si le code fait bien ce qu'il doit faire et s'il n'y a pas d'erreurs ou de bogues qui pourraient avoir un impact sur les données ou sur les résultats d'analyse (Cox, 2019). Il est essentiel de bien documenter les dépendances (et leurs versions) d'un projet pour la recherche reproductible puisque même de petits changements entre les versions peuvent créer des ruptures dans le fonctionnement du code, ou pire, produire des résultats erronés.

Un des moyens les plus courants d'écrire du code pour des conteneurs informatiques est d'utiliser des cahiers de code électroniques. La mise en conteneur d'un cahier de code permet d'analyser et de modifier le code pour tester les sorties et les analyses. Les utilisatrices ou utilisateurs finaux peuvent faire des essais avec le code sans avoir à se soucier de causer des ruptures de fonctionnement ou des modifications permanentes. De plus, pas besoin de se soucier des questions de sécurité liées à l'installation des logiciels.

# Conclusion

La curation active des données de recherche permet d'améliorer la recherche, car elle fait gagner du temps et réduit le potentiel d'erreurs. L'utilisation de processus de travail standardisés, l'application uniforme de méthodes d'organisation et d'étiquetage des produits de recherche et la création d'une documentation exhaustive facilitent la réutilisation, tant pour l'équipe de recherche primaire que pour des utilisatrices et utilisateurs secondaires. La normalisation améliore la découvrabilité des données dans les dépôts, ce qui permet l'ajout de jeux de données dans les revues systématiques et les méta-analyses augmentant ainsi la fréquence de citations et le profil de l'équipe de recherche.

Bien que les suggestions présentées dans ce chapitre soient considérées comme de meilleures pratiques, la meilleure gestion des données de recherche est de faire une gestion des données. Chacun des projets comportera ses défis particuliers, mais en veillant à la curation active des données, vous vous assurerez que la documentation est suffisante pour le dépôt et la découverte des données.

**Questions de réflexion**

Voir l'Annexe 3 pour une série d'exercices.

**Éléments clés à retenir**

- La curation active des données aide les chercheuses et chercheurs à s'assurer que leurs données sont précises, fiables et accessibles aux personnes qui en ont besoin. Quand les données de recherche sont gérées de façon appropriée et mises à jour, elles demeurent utiles et accessibles dans le temps.
- Les pratiques de gestion des données, telles que le versionnage et la rédaction de scripts, contribuent à améliorer l'exactitude et la sécurité des données. L'automatisation des descriptions, de l'organisation et du stockage des données de recherche permet de gagner du

temps et limite les erreurs.
- Les outils qui favorisent la reproductibilité des calculs et de l'analyse – tels que les cahiers électroniques de laboratoire et les conteneurs informatiques – permettent aux recherches d'être reproduites et vérifiées. En ouvrant l'accès aux données et aux méthodes d'analyses, les chercheuses et chercheurs peuvent prouver la rigueur et la fiabilité de leur recherche ainsi que permettre à d'autres d'examiner leur travail.

# Bibliographie

Alston, J. M. et Rick, J. A. (2021). A beginner's guide to conducting reproducible research. *The Bulletin of the Ecological Society of America, 102*(2), 1–14. https://doi.org/10.1002/bes2.1801 (https://doi.org/10.1002/bes2.1801)

Arguillas, F., Christian, T.-M., Gooch, M., Honeyman, T., Peer, L. et CURE-FAIR WG. (2022). *10 things for curating reproducible and FAIR research* (1.1). Zenodo. https://doi.org/10.15497/RDA00074 (https://doi.org/10.15497/RDA00074)

Benureau, F. C. Y. et Rougier, N. P. (2018). Re-run, repeat, reproduce, reuse, replicate: Transforming code into scientific contributions. *Frontiers in Neuroinformatics, 11*. https://doi.org/10/ggb79t (https://doi.org/10/ggb79t)

Buchanan, E. M., Crain, S. E., Cunningham, A. L., Johnson, H., Stash, H. R., Papadatou-Pastou, M., Isager, P. I., Carlsson, R. et Aczel, B. (2021). Getting started creating data dictionaries: How to create a shareable data set. *Advances in Methods and Practices in Psychological Science, 4*(1), 1-10. https://doi.org/10.1177/2515245920928007 (https://doi.org/10.1177/2515245920928007)

Cooper, A., Steeleworthy, M., Paquette-Bigras, È., Clary, E., MacPherson, E., Gillis, L. et Brodeur, J. (2021). Creating guidance for Canadian Dataverse curators: Portage Network's Dataverse curation guide. *Journal of EScience Librarianship, 10*(3), 1-26. https://doi.org/10/gmgks4 (https://doi.org/10/gmgks4)

Cox, R. (2019). Surviving software dependencies: Software reuse is finally here but comes with risks. *ACMQueue, 17*(2), 24-47. https://doi.org/10.1145/3329781.3344149 (https://doi.org/10.1145/3329781.3344149)

Groupe d'experts sur la formation. (2020, 2 septembre). *Guide Éclair: Gestion des données de recherche*. Zenodo. https://doi.org/10.5281/zenodo.4012530 (https://doi.org/10.5281/zenodo.4012530)

Hunt, G. J. et Gagnon-Bartsch, J. A. (2021). *A review of containerization for interactive and reproducible analysis*. ArXiv. https://arxiv.org/abs/2103.16004 (https://arxiv.org/abs/2103.16004)

ICPSR Institute for Social Research. (2023). *Glossary of social science terms. National Addiction and HIV Data Archive Program*. https://www.icpsr.umich.edu/web/NAHDAP/cms/2042 (https://www.icpsr.umich.edu/web/NAHDAP/cms/2042)

Johnston, L., Carlson, J., Hudson-Vitale, C., Imker, H., Kozlowski, W., Olendorf, R. et Stewart, C. (2017a). *Data Curation Network: A cross-institutional staffing model for curating research data*. https://conservancy.umn.edu/bitstream/handle/11299/188654/DataCurationNetworkModelReport_July2017_V1.pdf (https://conservancy.umn.edu/bitstream/handle/11299/188654/DataCurationNetworkModelReport_July2017_V1.pdf)

Johnston, L., Carlson, J. R., Kozlowski, W., Imker, H., Olendorf, R. et Hudson-Vitale, C. (2017b). *Checklist of DCN CURATE steps*. IASSIST & DCN – Data Curation Workshop 4. https://openscholarship.wustl.edu/data-curation-workshop-2017/schedule/Schedule/4 (https://openscholarship.wustl.edu/data-curation-workshop-2017/schedule/Schedule/4)

Kery, M. B., Radensky, M., Arya, M., John, B. E. et Myers, B. A. (2018). The story in the notebook: Exploratory data science using a literate programming tool. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–11. https://doi.org/10.1145/3173574.3173748 (https://doi.org/10.1145/3173574.3173748)

Krier, L. et Strasser, C. A. (2014). *Data management for libraries: A LITA guide*. American Library Association.

Morin, A., Urban, J., Adams, P. D., Foster, I., Sali, A., Baker, D. et Sliz, P. (2012). Shining light into black boxes. *Science, 336*(6078), 159–160. https://doi.org/10/m5t (https://doi.org/10/m5t)

Possati, L. M. (2020). Towards a hermeneutic definition of software. *Humanities and Social Sciences Communications, 7*(1), 1–11. https://doi.org/10.1057/s41599-020-00565-0 (https://doi.org/10.1057/s41599-020-00565-0)

Rokem, A., Marwick, B. et Staneva, V. (2017). Assessing reproducibility. Dans J. Kitzes, D. Turek et F. Deniz (dir.), *The practice of reproducible research: Case studies and lessons from the data-intensive sciences*. University of California Press. http://www.practicereproducibleresearch.org/core-chapters/2-assessment.html#

Sawchuk, S. L. et Khair, S. (2021). Computational reproducibility: A practical framework for data curators. *Journal of EScience Librarianship, 10*(3), 1-16. https://doi.org/10/gmgkth (https://doi.org/10/gmgkth)

Singh, A., Bansal, R. et Jha, N. (2015). Open source software vs proprietary software. *International Journal of Computer Applications, 114*(18), 26-31. https://doi.org/10/gh4jxn (https://doi.org/10/gh4jxn)

Statistiques Canada. (1996). *Manuel des codes A – Enquête nationale sur la santé de la population (ENSP) 1994-95 – Questions supplémentaires*. https://www.statcan.gc.ca/fra/programmes-statistiques/document/3225_DLI_D2_T22_V1-fra.pdf (https://www.statcan.gc.ca/fra/programmes-statistiques/document/3225_DLI_D2_T22_V1-fra.pdf)

Vuorre, M. et Crump, M. J. C. (2021). Sharing and organizing research products as R packages. *Behavior Research Methods, 53*(2), 792–802. https://doi.org/10/gg9w4c (https://doi.org/10/gg9w4c)

Vuorre, M. et Curley, J. P. (2018). Curating research assets: A tutorial on the Git Version Control System. *Advances in Methods and Practices in Psychological Science, 1*(2), 219–236. https://doi.org/10/gdj7ch (https://doi.org/10/gdj7ch)

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., ... Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data, 3*. https://doi.org/10/bdd4 (https://doi.org/10/bdd4)

## À propos des auteurs

### Sandra Sawchuk

Sandra Sawchuk est bibliothécaire des services de données et de l'expérience utilisation à la bibliothèque et aux archives de l'Université Mount Saint Vincent. Elle a une formation universitaire en humanités numériques et ses recherches portent sur le sauvetage et la réutilisation des données. Elle a récemment rédigé conjointement un article sur la reproductibilité informatique et participe actuellement à une subvention de partenariat de deux ans du CRSH visant à améliorer l'accès aux données historiques du recensement du Canada. ORCID : 0000-0001-5894-0183 (https://orcid.org/0000-0001-5894-0183)

### Louise Gillis

Louise Gillis est bibliothécaire à la gestion des données de recherche aux bibliothèques de l'Université de

Dalhousie. Dans le cadre de ses fonctions, Louise facilite les bonnes pratiques en matière de gestion des données de recherche en soutenant des outils tels que l'Assistant PGD et Dataverse. Elle est actuellement membre du comité de préservation et d'intendance numérique du Conseil des bibliothèques de l'Atlantique ainsi que du groupe de travail sur le dépôt et le stockage des données de l'Alliance de recherche numérique. En tant qu'ancienne membre du groupe de travail de Portage sur le guide de curation dans Dataverse, elle a corédigé un guide de curation pour le dépôt Dataverse de Scholars Portal. ORCID : 0000-0001-8250-5886 (h ttps://orcid.org/0000-0001-8250-5886)

## Lachlan MacLeod

Lachlan MacLeod est coordinateur des services sur le droit d'auteur et la gestion des données de recherche aux bibliothèques de l'Université de Dalhousie. Lachlan a travaillé dans les bibliothèques de Dalhousie où il a apporté son soutien aux services de droits d'auteur, aux données et aux statistiques, à la gestion des données de recherche et à l'évaluation des bibliothèques. Il était auparavant employé par le Centre de données de recherche de l'Atlantique (Statistique Canada). Il a une formation et une expérience dans les méthodes de recherche en sciences sociales, la gestion des données de recherche et le soutien des chercheuses et chercheurs avec leurs données. ORCID : 0000-0002-2702-9810 (https://orcid.org/0000-0002-2702-9810)

11.

# LA PRÉSERVATION NUMÉRIQUE DES DONNÉES DE RECHERCHE

Grant Hurley et Steve Marks

**Objectifs d'apprentissage**

À la fin de ce chapitre, vous pourrez :

1. Identifier les menaces à l'accès à long terme des données de recherche numériques.
2. Développer un plan pour la préservation d'un jeu de données particulier dans le contexte d'une Communauté d'utilisateurs cible (*Designated community*) et son cas d'utilisation anticipé.
3. Déterminer si certaines actions de préservation potentielles peuvent contribuer de façon positive à l'accès à long terme d'un jeu de données particulier.

## Introduction

La **préservation numérique** est couramment définie comme étant « série d'activités gérées nécessaires pour garantir un accès continu aux **objets numériques** aussi longtemps que nécessaire » (Digital Preservation Coalition, 2015, p. 282). Que le matériel soit d'abord sous forme numérique ou qu'il ait été numérisé à partir d'une autre source, l'objectif reste le même. La préservation numérique est un domaine relativement nouveau (du moins, en comparaison avec la préservation physique!), mais la préservation des **données de recherche** a fait partie du champ d'étude dès le départ. D'ailleurs, un des documents formatifs de la plupart des approches récentes à la préservation numérique – le modèle *Open Archival Information System (https://public.ccsds.org/Pubs/650x0m2%28F%29.pdf)* (Système ouvert d'archivage d'information ou **OAIS**) – a été développé par un

consortium d'agences spatiales pour résoudre le problème de l'accès aux données historiques des missions spatiales..

L'objectif de ce chapitre est de présenter certains des concepts de base de la préservation numérique avec un accent sur les approches pratiques aux problèmes courants que vous pourrez avoir à affronter, ainsi que leurs solutions, lors de vos démarches vers la préservation à long terme des données de recherche.

## Les menaces aux objets au fil du temps

Pour mieux comprendre les risques liés aux objets numériques (y compris aux données de recherche) au fil du temps, imaginons le scénario suivant : nous trouvons une pile de vieilles disquettes de 5,25 pouces. Nous croyons qu'elles pourraient contenir une forme quelconque de données utiles; des journaux de recherche d'un de nos prédécesseurs ou des données historiques dans notre domaine d'étude, etc. Tout ce qui nous importe, c'est de récupérer ce qui se trouve sur ces disquettes!

Toutefois, les lecteurs de ce type de disquettes ne font plus partie des modèles courants d'ordinateurs. En fait, ils peuvent être difficiles à trouver en bon état. Voilà donc notre première menace : l'**obsolescence des médias**. Notre média de stockage – dans ce cas-ci, la disquette – nécessite une configuration particulière de matériel informatique et de logiciels pour pouvoir être lu. Lorsque le matériel nécessaire n'est plus disponible (ou difficile à trouver), le média ne peut donc plus être utilisé et il est considéré comme obsolète.

Pour les fins de ce module, supposons que nous avons eu de la chance de mettre la main sur un lecteur de disquettes de 5,25 pouces. Nous insérons la disquette dans le lecteur et nous double-cliquons sur Finder ou Windows Explorer et... qu'est-ce qui se passe? L'ordinateur affiche que la disquette ne contient aucune donnée. Pourquoi? Il y a deux raisons potentielles. Il est en effet possible que la disquette ne contienne aucune donnée, mais il est aussi possible que nous soyons victimes d'une deuxième menace : la **dégradation du média**. Il s'agit de la détérioration dans le temps du média ou de l'information qu'il contient. La plupart des médias numériques ont une durée de vie limitée et, une fois disparues, les données peuvent être difficiles, voire impossibles à récupérer.

Toutefois, peut-être que les données sont toujours présentes, mais qu'elles ne peuvent pas être lues. Elles ont probablement été générées sur des modèles d'ordinateurs plus anciens et il est possible que le système d'origine ait inscrit les données sur la disquette d'une manière différente de celle des ordinateurs plus modernes. Sans logiciel pour aider l'ordinateur moderne à lire la disquette, il est difficile de déterminer la présence de fichiers, le nom de ces fichiers et où un fichier commence et l'autre finit. Toutes ces fonctions font partie d'une structure de données appelée système de fichiers.

Supposons maintenant que nous avons réussi à naviguer le système de fichiers de la disquette, soit parce que

l'ordinateur a été capable de le lire, soit parce que nous avons installé un élément qui en a permis la lecture. Il pourrait y avoir un autre problème à affronter : les fichiers eux-mêmes peuvent être inintelligibles aux applications couramment utilisées dans l'environnement informatique actuel. Les fichiers ont peut-être été créés avec un ancien programme de base de données ou ont été encodés dans un format qui ne peut être accessible qu'avec un programme de visualisation propriétaire qui n'est plus disponible. Ces problèmes de systèmes de fichiers sont des exemples d'**obsolescence des formats**.

Pour terminer, si nous réussissons à accéder à la disquette et aux fichiers qu'elle contient, à lire ces fichiers et à comprendre la façon dont ils sont décodés, il y a toujours un risque que certaines informations cruciales sur les données soient manquantes. S'il s'agit de données d'observation, certaines informations sur le quand, comment et où les données ont été recueillies peuvent être manquantes. S'il s'agit de données sous forme d'images, il est possible qu'il nous manque des informations sur ce que les images sont censées représenter. Pour tout type de données, il pourrait y avoir des informations manquantes, telles que qui les a créées et s'il y a des restrictions de propriété intellectuelle qui s'appliquent. Selon le cas d'usage, ces questions peuvent n'être d'aucun intérêt, mais si nous désirons effectuer un travail académique rigoureux, nous devons nous soucier de ce dernier problème du scénario : la **perte de la provenance**.

Ressentez-vous une certaine angoisse? La bonne nouvelle, c'est que plusieurs personnes ont déjà été confrontées à ce genre de problème. Le domaine plus global de la préservation numérique est d'ailleurs consacré à identifier, éviter et corriger la majorité de ces problèmes. Avant de discuter de la façon de les traiter, regardons d'abord quelques-uns des principes fondamentaux.

## Les objectifs de la préservation numérique

La Digital Preservation Coalition (DPC) définit la préservation numérique comme étant la « série d'activités gérées nécessaires pour garantir un accès continu aux objets numériques aussi longtemps que nécessaire » (2015, p. 282). Passons à travers les éléments qui composent cette définition pour expliquer les objectifs plus larges de la préservation numérique.

Commençons par « objets numériques » puisqu'ils sont la raison d'être des activités de préservation numérique. Que sont les objets numériques? Le mot « objet » suggère une forme physique; les objets numériques ont toujours une incarnation physique *quelque part*, qu'il soit stocké sur une disquette de 5,25 pouces, un serveur, un disque dur externe, une clé USB ou un CD. Chacune de ces méthodes de stockage encode les informations d'une manière quelconque, soit par fluctuation magnétique (serveurs, disquettes et plusieurs disques durs externes), par charge dans les cellules (lecteurs flash) ou par alvéoles (CD). La première couche de médiation est suivie par plusieurs autres couches. Les documents analogiques n'ont pas cette complexité. Prenons, par exemple, un document textuel tel un mémorandum. En format papier, il y a deux niveaux immédiats de médiation : la feuille de papier physique (est-elle intacte et complète ou endommagée?)

et le texte écrit dessus (est-il lisible ou effacé? Quelle est la langue du texte?). L'équivalent numérique du mémorandum en format DOCX de Microsoft Word doit d'abord être extrait d'un médium de stockage comme une suite d'octets qui, regroupés, forme une séquence de bits (aussi connu comme un train de bits) avec un début et une fin distincts. Règle générale, plusieurs séquences de bits sont nécessaires pour composer un fichier. C'est le cas du format DOCX qui est composé de plusieurs dossiers et de fichiers texte XML regroupés dans une archive ZIP. Il est facile d'oublier qu'un fichier numérique s'appelle *fichier* parce qu'il regroupe plusieurs petits éléments d'informations, tout comme un fichier papier pourrait contenir plusieurs documents. Dans d'autres cas, des fichiers individuels multiples peuvent être accessible, mais ils doivent être exploités ensemble pour obtenir le résultat escompté, tel que des scripts utilisés pour traiter des saisies de données ou une collection de fichiers texte en format HTML, CSS et JavaScript, ainsi que des images et des PDF qui constituent ensemble un site Web. Dans sa plus simple expression, une seule séquence de bits constitue l'entièreté d'un fichier texte.

Dans l'un ou l'autre des cas, les séquences de bits doivent être interprétées selon une structure particulière : le **format de fichier**. Un format de fichier est « une convention qui établit les règles sur la façon de structurer et de stocker l'information dans un fichier[1] » [traduction] (Owens, 2018, p.47). Les formats de fichiers relient les séquences de bits et les systèmes de fichiers aux logiciels. Face à un format de fichier particulier, les systèmes d'exploitation permettent l'installation d'éléments particuliers d'un logiciel qui pourront lire, interagir et sauvegarder les fichiers dans ce format. Les formats ont aussi l'avantage de soutenir les échanges; puisque chaque fichier dans son format particulier est structuré de la même façon, il devient lisible pour différentes applications ou divers systèmes qui souhaitent ouvrir un fichier dans ce format. Mais le format de fichier est une construction humaine : « toute conversation autour des formats doit partir du principe qu'ils sont des conventions qui établissent la façon dont les fichiers devraient être structurés, il ne s'agit pas d'une vérité essentielle[2] » [traduction] (Owens, 2018, p.120). Certains formats de fichiers, surtout ceux liés à un logiciel, ne sont pas accessibles sans ce logiciel et contraignent les personnes qui l'utilisent à un produit commercial particulier. Les formats de fichier peuvent aussi changer au fil du temps pour répondre aux exigences des logiciels et des personnes qui les utilisent; le logiciel sous une version peut être incompatible avec un format de fichier d'une version antérieure. Les logiciels spécialisés (utilisés dans les domaines de recherche comme les sciences humaines, les sciences sociales ou la biologie) peuvent utiliser des formats de fichier uniques ou être exploités par différentes versions mal documentées ou soutenues d'un logiciel.

Un logiciel nécessite un ordinateur physique pour être exploité; cet ordinateur est composé de pièces de matériel informatique telles que la mémoire, les processeurs et l'espace de stockage. Un système d'exploitation

1. "*a convention that establishes the rules for how information is structured and stored in a file*."
2. "*all conversations about formats need to start from the understanding that they are conventions for how files are supposed to be structured, not essential truths*."

tel que Windows, Mac OSX ou Linux est un morceau de logiciel qui contrôle tous ces composants, en plus de certains autres comme les périphériques d'entrée (clavier, souris), les périphériques de sortie (afficheur, imprimante), le stockage et le réseau. Les systèmes d'exploitation contrôlent également l'accès au système de fichiers de l'ordinateur; c'est ce qui établit les règles sur la façon de stocker et de récupérer les données ainsi que l'emplacement du médium de stockage. En raison des mises en œuvre particulières de chacun des systèmes d'exploitation, certains logiciels ne peuvent fonctionner que sur des systèmes d'exploitation spécifiques ou être limités à certaines de ses versions.

Examinons maintenant cette notion de « l'accès continu ». Cet accès est affecté par le niveau d'ouverture, c'est-à-dire si l'objet numérique est disponible gratuitement en ligne, sur demande ou s'il est limité à des individus précis ou communautés particulières en fonction des coûts, de la confidentialité, de droits d'auteur ou d'autres restrictions. L'accès continu peut être menacé par des questions telles que la perte entraînée par une annulation d'abonnement ou par l'arrêt des activités d'un fournisseur de service. À cet effet, les responsables de la préservation numérique doivent s'assurer de conserver au fil du temps les informations sur la **provenance** ainsi que les droits à l'accès des objets numériques. La norme de facto pour cette information est celle du **PREMIS pour les métadonnées**, gérée par la Library of Congress; elle fournit la structure pour l'enregistrement d'informations détaillées sur les actions prises pour conserver le matériel numérique au fil du temps.

Pour terminer, la définition du DPC reconnaît que les objets numériques ne pourront pas tous être conservés à jamais; « aussi longtemps que nécessaire » est donc plus réaliste. Certains objets ont une valeur immédiate, mais cette valeur peut disparaître au fil du temps. D'autres objets doivent être supprimés, conformément aux législations sur la confidentialité ou aux règles d'éthique de la recherche. Idéalement, les responsables de la préservation numérique confient le travail de maintenance à d'autres qui le poursuivent. C'est la deuxième signification de « gestion », telle que décrite plus haut; le travail de préservation numérique doit se faire à l'intérieur d'une structure – institutionnelle ou autre – pour qu'il puisse durer plus longtemps qu'une prise en charge par des individus particuliers.

## La préservation numérique versus la curation

Si la préservation numérique est une série de procédés de gestion dont l'intention est de conserver un accès dans le temps, il faut donc se poser la question suivante : compte tenu de toutes les exigences en matière de ressources humaines et techniques, que faut-il préserver? Le sujet de l'établissement des priorités de préservation – ce qui permet d'identifier les objets auxquels un établissement choisit ou ne choisit pas d'accorder des ressources pour la préservation – s'inscrit dans le domaine plus large de la curation numérique et, plus particulièrement, dans la partie évaluation du processus de curation. L'évaluation, telle que précisée dans le *Guide d'évaluation pour la préservation des données de recherche (https://zenodo.org/record/6283886)*,

rédigé par Jonathan Dorey, Grant Hurley et Beth Knazook, implique la détermination d'une valeur. Dans le cas des données de recherche, qui sont généralement déposées par une créatrice ou un créateur en relation avec un établissement, la question devient donc : est-ce que ce jeu de données possède une valeur éventuelle suffisante pour justifier son acquisition et sa préservation? Si votre établissement a comme mission de faire de la préservation à long terme, vous aurez besoin d'accéder aux sujets ou domaines de connaissances appropriés pour pouvoir établir la valeur. Vous pouvez également consulter les stratégies ou politiques en matière de développement de collections pour déterminer si un jeu de données candidat s'aligne avec les priorités de votre établissement. De plus, une expertise particulière en préservation numérique peut être nécessaire pour identifier à quel point les objets numériques peuvent être préservés, les types d'interventions de préservation nécessaires et les ressources nécessaires pour effectuer le travail. Ce procédé est une *évaluation technique*. Une fois la valeur d'un jeu de données établie, les activités subséquentes de curation peuvent miser sur son amélioration par le biais de contrôles de qualité, de tests sur le code et d'amélioration de la documentation et des **métadonnées**. Vous pourrez aussi avoir à identifier des fichiers individuels d'un jeu de données qui ne devraient pas être retenus ou inversement, des fichiers manquants qui doivent être recueillis. Une liste exhaustive de ces types d'activités est offerte par le flux de travail *CURATE(D)* (https://datacurationnetwork.org/outputs/workflows/) (en anglais uniquement) du Data Curation Network ainsi que le *Guide pour la curation de Dataverse (https://zenodo.org/record/5579827)*, préparé par l'Alliance de recherche numérique du Canada.

Conformément à la définition de préservation numérique du DCP qui parle de la notion du « aussi longtemps que nécessaire », le choix de conserver un jeu de données n'est pas permanent; les jeux de données peuvent être revus par le biais d'un processus de réévaluation pour déterminer s'ils maintiennent toujours une certaine valeur pour l'établissement et sa communauté.

## Les Communautés d'utilisateurs cibles

Les possibilités d'interventions de préservation pour une série particulière d'objets numériques qu'un établissement décide de conserver sont nombreuses, alors les responsables de la préservation peuvent se demander quels critères doivent être utilisés pour faire leur choix. La norme *Open Archival Information System* (Système ouvert d'archivage d'information ou OAIS) contient un concept utile pour aider dans ce travail : la notion des **Communautés d'utilisateurs cibles**. La norme OAIS les définit ainsi:

> Un groupe identifié d'Utilisateurs potentiels, susceptibles de comprendre un ensemble donné d'informations. La Communauté d'utilisateurs cible peut être constituée de plusieurs communautés d'utilisateurs. La Communauté d'utilisateurs cible est définie par l'Archive et sa définition peut évoluer au cours du temps (CCSDS, 2012, p. 1-9).

Plusieurs bibliothécaires et archivistes ont trouvé ce concept difficile à assimiler; le fait de restreindre leurs

activités autour d'un groupe particulier peut sembler contradictoire à leur mission professionnelle qui est d'assurer un accès large et ouvert au grand public (Bettivia, 2016, p.5). Identifier une Communauté d'utilisateurs cible n'exclut pas de préserver des objets numériques pour la population en général. Les responsables de la préservation doivent tenir compte des besoins des utilisatrices et utilisateurs dans la prise de décisions pour la préservation, y compris les résultats des interventions de préservation, les métadonnées rendues disponibles et la série de services courants qui permet l'accès (Marks, 2015, p.16). Autrement dit, « faire de la préservation pour quelqu'un plutôt que de préserver quelque chose[3] » [traduction] (Bettivia, 2016, p.3). Plusieurs établissements ont implicitement des Communautés d'utilisateurs cible, notamment les membres du corps professoral, la communauté étudiante et le personnel dans un établissement universitaire, la population d'un village ou d'un territoire ou le personnel d'un organisme privé, et ce, même s'ils détiennent un mandat auprès du grand public. Désigner une Communauté d'utilisateurs cible oblige à rendre ces prises en charge explicites. Des communautés primaires, secondaires et tertiaires peuvent être désignées, chaque groupe ayant des spécificités moins restreintes, regroupant ainsi un large ensemble de membres sans avoir à faire d'impossibles promesses de préserver tous les objets numériques au nom du monde entier.

En envisageant la préservation en fonction de communautés ciblées, l'information préservée doit rester immédiatement compréhensible aux membres de cette Communauté d'utilisateurs cible. La norme OAIS définit « immédiatement compréhensible » comme étant « la qualité propre d'une information suffisamment documentée pour être interprétée, comprise et utilisée par la Communauté d'utilisateurs cible sans recourir à des ressources particulières difficiles d'accès comme par exemple des personnes physiques » (CCSDS, 2012, p.1-11). Autrement dit, le matériel devrait être utilisable par les membres de la communauté sans aide externe. En tant que curatrice ou curateur, vous devez être au courant des connaissances des membres de la Communauté d'utilisateurs cible et vous devez leur fournir des objets numériques qui leur seront accessibles. Dans le contexte de la **gestion des données de recherche** (GDR), il est courant de présumer du niveau d'expertise liée au domaine ou à la discipline dans laquelle les données ont été créées. Par exemple, un dépôt de données en sciences sociales peut présumer que les membres de sa communauté principale (les chercheuses et chercheurs en sciences sociales) savent utiliser des logiciels d'analyse statistique; il est donc suffisant de préserver et fournir des données tabulaires en format brut qui peuvent être utilisées dans R ou un autre logiciel. Si le dépôt souhaite s'ouvrir aux gens qui n'ont pas cette expertise, il pourrait être nécessaire de fournir d'autres options d'accès telles qu'une interface visuelle interactive pour faire des requêtes de données tabulaires. De cette façon, à certains niveaux de l'infrastructure de préservation et d'accès, « il y a de multiples services et, à un moment donné, la spécificité d'un sujet peut nécessiter la mise en place d'une approche différente pour servir les différentes Communautés d'utilisateurs cible[4] » [traduction] (Bettivia,

3. "*preservation for someone rather than preservation of something.*"
4. "*there is a commonality of services, and at some point subject-specificity may dictate a need for different approaches to serve different Designated Communities.*"

2016, p.6). En fin de compte, comme l'observe McGovern dans *Digital Preservation Management Model Document* (2016), « une archive numérique peut être sombre, peu lumineuse ou éclairée, mais la preuve absolue d'une préservation réussie est dans la capacité de fournir un accès significatif à long terme[5] » [traduction]. Autrement dit, si le matériel numérique ne peut être utilisé, c'est qu'il n'a pas été préservé de manière utile.

## Les propriétés significatives

Après avoir établi le concept de Communauté d'utilisateurs cible, nous pouvons maintenant aborder un autre concept important qui découle directement de la Communauté d'utilisateurs cible et de ses besoins : les propriétés significatives. Le glossaire de la Digital Preservation Coalition définit les *propriétés significatives* comme les « caractéristiques des objets numériques et intellectuels qui doivent être préservées dans le temps afin de garantir l'accessibilité, la facilité d'utilisation et la signification des objets et leur capacité à être acceptés comme (preuve de) ce qu'ils prétendent être » (2015, p. 283).

Les propriétés significatives sont importantes parce qu'elles découlent des perspectives et besoins des Communautés d'utilisateurs cible. Plus précisément, elles sont les propriétés d'un **objet numérique** précis qui répond aux besoins de la Communauté d'utilisateur cible. Ces propriétés significatives peuvent varier selon l'objet numérique et même dans le cadre d'un seul objet, elles peuvent être aussi différentes que les Communautés d'utilisateurs cible qui pourront y accéder. Cela étant dit, il existe plusieurs propriétés significatives importantes qui s'appliquent à presque tous les cas.

Une de ces propriétés significatives clés est le format. Comme mentionné plus tôt, les objets numériques ont souvent besoin de morceaux particuliers de logiciels pour pouvoir être accessibles et la capacité du logiciel à s'exécuter dépend de sa capacité à interpréter le sens des données codifiées dans le fichier – le format de fichier. Les différents types de données de recherche, telles que les données tabulaires, les documents textes, les images et les enregistrements audio ou vidéo, peuvent utiliser différents formats de fichiers pour stocker les informations de façon précise et efficace.

Les métadonnées sont une autre des propriétés significative des données de recherche; elles peuvent inclure des informations sur l'autorat des données, la méthodologie, la couverture et autres détails pertinents. Des métadonnées précises et complètes sont essentielles pour bien comprendre le contexte et la signification des données, en plus de permettre la citation et l'attribution exactes. Dans le domaine des données de recherche, les métadonnées peuvent être très spécialisées, au même titre que les données. Par exemple, les données d'une enquête historique utilisées pour appuyer une recherche en sciences sociales peuvent être décrites avec la

5. "*A digital archive may be dark, dim, or lit, but the absolute proof of preservation is in the capability to provide meaningful long-term access.*"

norme DDI pour les métadonnées (https://ddialliance.org/) qui prône une description robuste de tout détail susceptible d'être pertinent tel que la population étudiée, la méthodologie d'échantillonnage, etc. Un jeu de données recueilli dans le contexte d'un projet d'astronomie n'aura vraisemblablement pas besoin de ces types de balises, mais en nécessitera plutôt une série d'autres liées notamment à l'orientation du télescope, aux conditions météorologiques et autres. Pour plus d'informations sur les métadonnées et sur les considérations importantes dont il faut tenir compte en choisissant des formats de fichiers pour une longue durée de vie, veuillez consulter le chapitre 9, « Un aperçu du fascinant monde des formats de fichiers et des métadonnées. »

En plus de ces propriétés techniques, les données de recherche peuvent avoir d'autres propriétés significatives reliées à leur contenu ou au contexte. Par exemple, les données peuvent faire partie d'une étude ou d'un projet de recherche plus large ou peuvent être liées à d'autres jeux de données ou objets numériques. Il est important de tenir compte de ces liens et rapports en préservant les données de recherche pour assurer que les données puissent être comprises et utilisées dans le contexte où elles ont été créées. Il est difficile de se prononcer sur la façon courante de stocker ces propriétés significatives parce que les moyens dépendent du contexte dans lequel se trouve la chercheuse, le chercheur ou le groupe ayant recueilli les données ou du dépôt dans lequel les données se retrouvent. Vous pouvez vous poser quelques-unes des questions suivantes quand vous évaluez les propriétés dont il faut tenir compte :

- Le jeu de données fait-il partie d'une série?
- Le jeu de données a-t-il d'autres versions?
- Ces données appuient-elles une publication particulière?
- Ces données représentent-elles un sous-ensemble d'un jeu de données plus large?

Les propriétés significatives peuvent, dans un premier temps, être complexes à identifier, mais l'élément le plus important à se rappeler, c'est qu'elles sont l'expression des besoins de la Communauté d'utilisateurs cible. En cas de doute, vous n'avez qu'à consulter un membre de la communauté ou à réfléchir à quels aspects des données sont essentiels pour assurer leur utilisation par cette communauté.

# La préservation numérique dans le contexte des données de recherche

## Les actions de préservation

Dans cette section, nous passons des cadres conceptuels à la pratique au quotidien de la préservation numérique par l'entremise de l'identification, de la performance et de l'évaluation des actions de préservation.

Quatre grandes catégories d'actions courantes de préservation sont abordées ci-dessous :

- Les **sommes de contrôle** et la **préservation au niveau des bits** établissent l'intégrité et une assurance de base que le matériel demeure intact et complet au fil du temps. La préservation au niveau des bits nécessite que les organismes identifient des stratégies robustes pour le stockage de préservation; elle est associée à la prévention des problèmes liés à l'obsolescence et à la dégradation des médias.
- Les métadonnées techniques sont généralement extraites de fichiers ou de séquences de bits, ce qui peut fournir des renseignements sur la façon de gérer les fichiers et les séquences de bits au fil du temps. L'information la plus couramment extraite à cette fin est l'identification des formats de fichiers. L'extraction des métadonnées techniques aide à réduire les risques associés à l'obsolescence des formats et à la perte de la provenance.
- La validation des formats de fichiers prend l'information obtenue du processus d'identification et, dans le cas de certains formats, évalue si le fichier en question répond aux normes de base en matière de structure et de qualité qui ont été définies pour ce format. Ce processus est lié à l'obsolescence des formats, mais peut aussi aider à identifier des dégradations potentielles au niveau des supports.
- Pour terminer, les actions de **normalisation** et de migration peuvent être prises pour assurer que les données ne soient pas enfermées dans un format oublié ou propriétaire. Une fois de plus, l'action répond au problème de l'obsolescence des formats.

Bien que cette liste ne comprenne pas toutes les activités possibles de préservation numérique, ces fonctions comptent parmi les plus couramment mises en place sur une base régulière en utilisant des outils et des procédés particuliers. Elles représentent le travail pratique de la préservation numérique, qu'elles soient exécutées manuellement, ou plus souvent, par le biais de scripts ou de logiciels de traitement pour la préservation. Quand vient le temps d'évaluer les capacités d'un dépôt à préserver des objets numériques, il est primordial d'identifier la présence (ou l'absence) de ces fonctions.

## Les sommes de contrôle, la préservation au niveau des bits et le stockage de préservation

La préservation au niveau des bits est généralement considérée comme étant la base en matière d'actions que peut prendre un organisme pour appuyer la préservation à long terme. Cette approche vise à garantir que les fichiers conservent leur **intégrité** (c'est-à-dire qu'ils demeurent intacts et sans modification dans l'ordre de la séquence des bits) et que les fichiers soient stockés dans de multiples emplacements pour les protéger contre la perte, la modification ou la corruption accidentelle. La préservation au niveau des bits ne garantit pas que le contenu ou le format des fichiers puisse être utilisé/accessible à l'avenir, elle assure simplement que les fichiers soient intacts. Les actions de préservation de base comprennent : pendant le traitement ou le stockage des données en vue de la préservation, la personne responsable de la préservation exécute un algorithme de

somme de contrôle aux fichiers téléversé dans le système et enregistre les résultats. Selon un horaire variable, le responsable exécute une nouvelle vérification de la somme de contrôle à une date ultérieure. La deuxième vérification (et toutes celles qui suivent) est appelée le contrôle d'intégrité. Si les résultats de la deuxième vérification correspondent à celles de la première, les objets numériques conservent leur intégrité. Chaque fois qu'un contrôle de l'intégrité est effectué, il est recommandé de stocker les résultats accompagnés de la date et de l'heure dans une base de données ou à un autre emplacement.

Les sommes de contrôle désignent des chaînes numériques ou alphanumériques uniques de longueurs variées produites par un algorithme générateur de sommes de contrôle tel que CRC, MD5, SHA1 et SHA256; le résultat de l'algorithme dépend du contenu du fichier. Quand les contenus d'un fichier sont modifiés de quelque façon que ce soit, la valeur de la somme de contrôle variera, indiquant que l'intégrité du fichier est atteinte et qu'il doit être remplacé par une autre copie. CRC, MD5 et SHA1 ne sont pas considérés comme des algorithmes sécurisés pour la cryptographie, mais ils sont quand même couramment utilisés pour détecter des problèmes d'intégrité. Pour une discussion sur le sujet, vous pouvez consulter le guide de Matthew Addis, *Which checksum algorithm should I use? (https://www.dpconline.org/docs/technology-watch-reports/2399-twgn-checksums-addis/file)* En effet, les sommes de contrôle constituent la composante centrale de nombreuses infrastructures informatiques. La clé, c'est d'identifier quand et comment elles sont exécutées. Les fichiers sont plus susceptibles de perdre leur intégrité en cours de transfert d'un système à l'autre, comme lors du téléversement des fichiers à des stockages de préservation à distance sur le Web. Idéalement, une somme de contrôle est effectuée localement sur l'ordinateur et celle-ci est ensuite comparée aux résultats de la vérification de l'intégrité une fois à destination. Gardez en tête qu'il existe une variété d'outils qui peuvent automatiser ce travail à votre place; parmi des exemples courants, des fichiers stockés sous format BagIt utilisent des outils mis en œuvre par la bibliothèque Python BagIt.

La deuxième composante importante d'une stratégie de préservation au niveau des bits est la création de multiples copies. Si vous identifiez un problème d'intégrité, la solution idéale est de remplacer la « mauvaise » copie avec sa version intacte. Le fait d'avoir plusieurs copies, idéalement dans différents emplacements, vous permet donc de limiter les problèmes d'intégrité qui peuvent survenir. Les différentes méthodes de stockage pour la préservation varient beaucoup selon les ressources accessibles aux organismes qui préservent. Les seules méthodes possibles pour certains organismes sont la création de copies distinctes stockées sur des disques durs externes, ou dans un système RAID, ou un réseau de stockage local (avec idéalement, une copie de sauvegarde). Les organismes qui font de la préservation à plus grande échelle peuvent utiliser des systèmes de sauvegarde sur bande magnétique.  Et les services tiers – tels que le stockage infonuagique ou autre réseau de stockage dupliqué – peuvent répondre aux besoins des institutions de mémoire. Dans le document *Levels of Digital Preservation Matrix (https://osf.io/36xfy)* du National Digital Stewardship Alliance (NDSA), la section sur le stockage est utile pour aider à déterminer la quantité nécessaire de copies et pour identifier un emplacement de stockage; les options varient de deux copies conservées dans des endroits séparés (mais dans un même lieu géographique) à « au moins trois copies dans des lieux géographiques soumis à des menaces de

nature différentes» (Ledoux *et al.*, 2019). Il est important de noter que les niveaux NDSA n'ont pas besoin d'être appliqués uniformément à tous les objets numériques; plusieurs établissements appliquent différentes stratégies de stockage pour préserver différentes catégories ou genres d'objets numériques. Vous pouvez aussi consulter *Digital Preservation Storage Criteria* de Schaefer *et al*. (2018) qui propose une structure pour l'évaluation des différentes options de stockage pour la préservation.

## L'identification des formats de fichiers

L'identification des formats de fichiers est généralement la première étape dans le travail du responsable de la préservation numérique, une fois que l'intégrité et le stockage sécuritaire de l'objet numérique à préserver sont confirmés. Connaître le format (et parfois la version particulière de ce format) peut vous aider à déterminer les façons d'accéder au fichier et de le conserver dans le temps. Par conséquent, la communauté de GDR se soucie particulièrement de bien comprendre les formats de fichiers. Les chercheuses et chercheurs sont encouragés à exporter les fichiers de données finaux dans des formats non propriétaires et des établissements, tels que le Data Archiving and Network Services (DANS) (https://dans.knaw.nl/en/) des Pays-Bas, ont établi des préférences de formats de fichiers pour l'inclusion dans leur dépôt (2022).

Comme il est nécessaire de pouvoir identifier de façon fiable les formats de fichiers, des procédés ont été établis pour vous aider. Vous pouvez généralement déduire le format d'un fichier par son extension. Toutefois, les formats propriétaires, obsolètes ou spécialisés peuvent être plus difficiles à identifier et souvent, les systèmes permettent de changer l'extension d'un fichier sans modifier son contenu. La clé, c'est de trouver un outil qui peut identifier un format de fichier par sa signature. La **signature** est une séquence d'octets qui survient de façon prévisible au début ou souvent à la fin d'un fichier. Pour que cette série d'octets soit un marqueur fiable, chaque instance d'un format de fichier particulier devrait inclure la signature. Certains formats de fichier – tels que les fichiers de texte brut – n'ont pas de signature de sorte que le format du texte doit être déduit à partir du contenu et de la structure du fichier.

Les outils qui identifient les signatures de format de fichier interrogent généralement la base de données PRONOM (https://www.nationalarchives.gov.uk/PRONOM/), gérée par les archives nationales du Royaume-Uni. Celle-ci comprend une liste exhaustive des signatures associées aux différents formats et leurs versions.De nouveaux formats sont régulièrement ajoutés à PRONOM. Les identifications de type MIME, qui sont couramment utilisées par les navigateurs Web, les logiciels de courrier électronique et autres logiciels qui identifient les types de fichiers, peuvent utiliser des signatures, mais peuvent aussi avoir recours aux extensions. Les types MIME n'identifient pas les versions particulières des formats de fichiers, mais peuvent être utiles lorsque l'identification par la signature, plus exigeante, échoue. Les outils d'identification des formats de fichiers par la signature comprennent Siegfried (https://www.itforarchivists.com/siegfried) (administré par Richard Lehane) et FIDO (https://openpreservation.org/tools/fido/) (administré par la Open Preservation Foundation).

## La validation du format de fichier

Une fois le format de fichier identifié, des actions subséquentes peuvent être prises. La validation du format de fichier est un processus de vérification pour confirmer que le format de fichier répond aux spécifications qui ont été conçues pour ce format. Les formats de fichiers n'ont pas tous des spécifications publiées, mais lorsqu'il y en a, il est possible de vérifier si l'instance d'un fichier constitue une représentation juste de ce format. Dans le jargon de la préservation, deux questions se posent : le fichier est-il bien formé et est-il valide?

Un fichier bien formé implique qu'il obéit aux règles *syntaxiques* de son format, c'est-à-dire qu'il suit les règles de structure de base établies par les spécifications du format de fichier. Ensuite, pour qu'un fichier soit valide, il doit être bien formé et répondre aux règles *sémantiques*. Ces dernières sont plus exigeantes en matière de qualité minimale pour un format de fichier, par exemple un fichier de format TIFF doit contenir une quantité minimale de données d'image. Comme l'observe Owens (2018), « de nombreuses applications logicielles courantes créent des fichiers qui sont, à des degrés différents, invalides selon les spécifications[6] » [traduction] (p.120).

Dans le contexte des données de recherche, le fait de valider ou non le format dépendra du format en question et des problèmes identifiés : des spécifications ont-elles été établies pour ce format? Existe-t-il un outil qui peut vérifier le fichier par rapport à sa spécification? Plus important encore, si le fichier est invalide ou valide, mais mal formé, quelles actions doivent alors être prises? Si le fichier est entièrement corrompu ou que les problèmes identifiés ont un impact important sur son utilisabilité, il peut être souhaitable de revenir vers la personne qui l'a créé pour lui demander de corriger le problème. Dans d'autres cas, les responsables de la préservation enregistrent l'information sur la validation dans les métadonnées, mais n'agissent pas en conséquence. Wheatley (2018) a documenté une série de questions utiles qui aident à évaluer les erreurs de validation : le fichier est-il chiffré? Dépend-il de composantes externes que vous n'avez pas? Est-il très endommagé? Le fichier est-il dans le format que vous pensiez? La validation peut aider à identifier les problèmes à différents stades. Certaines de ces questions peuvent être répondues au cours de l'étape de la phase de curation lorsque la personne responsable de la curation des données vérifie activement la qualité, l'exhaustivité et la facilité d'utilisation des fichiers. Ensuite, un flux de travail de préservation pourra simplement enregistrer la validité des formats dans les métadonnées, ce qui pourra servir à une éventuelle nouvelle vérification de la validité. Les outils (en anglais uniquement) pour la validation des formats de fichier comprennent JHOVE (https://jhove.openpreservation.org/) (administré par la Open Preservation Foundation) pour une variété de formats et veraPDF (https://verapdf.org/) pour les fichiers PDF/A.

## La conversion des formats de fichiers : la normalisation et la

6. "*many everyday software applications create files ... that are to varying degrees invalid according to the specifications*."

## migration

La conversion des formats de fichiers est probablement le plus actif des processus discutés dans ce chapitre. Plutôt que de recueillir des informations sur les fichiers, la conversion des fichiers en d'autres formats affecte directement le contenu des fichiers eux-mêmes. Tel que mentionné précédemment, cette action peut être effectuée avant de transférer les fichiers dans un dépôt, par exemple lorsque les chercheuses et chercheurs sont encouragés d'exporter leurs fichiers dans des formats non propriétaires particuliers ou des formats plus favorables à être préservés. Selon les résultats de l'identification des formats de fichiers, la conversion peut aussi se faire au cours du traitement des fichiers en vue de leur préservation à long terme. La conversion des formats de fichiers peut avoir un impact important sur les propriétés ou sur le contenu des informations du fichier et elle doit être entreprise en s'assurant que le fichier dans son nouveau format répondra toujours aux besoins de la Communauté d'utilisateurs cible. Il est donc important de multiplier les tests et de régulièrement valider les sorties de conversion avec une variété d'échantillons de fichiers.

La normalisation et la migration sont deux processus différents qui aboutissent au même résultat. La normalisation est le processus de conversion des fichiers vers une série de formats normalisés tels que définis par l'archive ou le dépôt au moment de la réception ou de l'ingestion. Le dépôt n'a donc qu'à prendre en charge la gestion d'un sous-ensemble de formats de fichiers pour l'avenir. Il est question de migration quand un dépôt doit convertir ses fichiers en un format secondaire, généralement à grande échelle, en réponse à un risque identifié, notamment un format qui n'est plus supporté. Pour un processus comme pour l'autre, une nouvelle copie du fichier est créée dans un format différent, qui sera aussi géré par le dépôt. La copie originale est généralement conservée pour éviter la perte accidentelle d'information survenue en cours de conversion. Par le passé, la normalisation pour la préservation était un procédé par défaut pour de nombreux dépôts. Désormais, les dépôts évaluent plus attentivement à quel moment la normalisation devrait se faire pour assurer qu'ils réduisent les impacts environnementaux et financiers liés à la création de copies trop nombreuses.

La normalisation et la migration à des fins de préservation doivent être distinctes de ces mêmes actions à des fins d'accès. La normalisation ou la migration pour favoriser l'accès est appliquée pour permettre aux Communautés d'utilisateurs cibles d'accéder aux fichiers en fonction de leurs besoins. Par exemple, un fichier TIFF important qui contient une carte géographique peut être normalisé en un fichier JPEG pour faciliter l'accès en ligne.

Les outils pour la conversion des formats de fichiers sont nombreux et varient en fonction des formats en question. Par exemple, les outils couramment utilisés pour les processus de travail automatisés comprennent ImageMagick (https://imagemagick.org/) pour les images et les FFmpeg (https://www.ffmpeg.org/) pour les fichiers audios et vidéos.

# Évaluer les actions de préservation

## Au niveau du fichier et de la collection

Évaluer les résultats des actions de préservation pour des fichiers individuels ou des collections à différents niveaux d'agrégation implique d'exécuter une action – telle que l'identification ou la normalisation des formats de fichiers – et d'examiner les résultats. En règle générale, ces actions sont effectuées en tant que tests jusqu'à ce que les résultats soient considérés comme acceptables. À ce moment-là, des approches plus automatisées et flexibles prennent le relais pour créer la version finale. Pour l'identification et la validation des formats de fichiers, il faut déterminer si le résultat correspond aux attentes. Par exemple, les fichiers NVP qui sont produits par le logiciel NVivo pour l'analyse des **données qualitatives** ne sont pas toujours identifiables avec un outil comme Siegfried parce qu'aucune description de ce format n'existe dans PRONOM. La personne responsable de la préservation doit décider si des outils supplémentaires devraient être mis en œuvre pour identifier ce type de fichiers ou s'il vaut mieux attendre une éventuelle mise à jour de PRONOM qui permettrait de refaire le processus d'identification.

Si un fichier est mal formé, mais qu'il peut être ouvert et consulté, l'erreur signalée par l'outil ne nécessite peut-être pas un triage plus approfondi. Il est également important d'évaluer les résultats des actions de normalisation et de migration. Est-ce qu'un outil particulier de conversion produit un résultat qui répond mieux aux besoins de la Communauté d'utilisateurs cible en fonction de son contenu informatif ou de sa présentation? Sinon, des outils ou stratégies supplémentaires, tels que l'**émulation**, peuvent être nécessaires. Par exemple, la conversion de documents de la suite MS Office – notamment les présentations PowerPoint – en format PDF peut nécessiter l'accès aux polices originales du document à moins qu'elles aient été intégrées au fichier original. Sans accès à ces polices, la mise en page et l'aspect général de la version PDF peuvent être différents de l'original. Cet élément est-il important aux membres de la Communauté d'utilisateurs cible qui accèdent au fichier ou est-ce que le contenu informatif suffit? Ces évaluations peuvent être améliorées en ayant accès aux membres de la Communauté d'utilisateurs cible par l'intermédiaire de groupes consultatifs ou en interrogeant les membres de la communauté.

## Au niveau du logiciel et du système

Sur la base des exemples mentionnés précédemment, vous comprenez comment la réflexion sur les résultats à un niveau granulaire a un impact sur les décisions prises au niveau du système. L'utilisation d'un outil pour résoudre un ensemble de problèmes risque d'affecter d'autres fichiers du dépôt. Les actions de préservation peuvent être prises sur une base individuelle, un fichier à la fois, mais il est plus courant pour les responsables de la préservation de s'appuyer sur des outils de flux de tâches qui automatisent une série d'actions à plus grande échelle. Ces responsables ont aussi une deuxième responsabilité, celle d'évaluer la fonctionnalité et

l'impact des logiciels de flux de tâches, y compris s'ils peuvent effectuer les actions de préservation requise, en plus de valider les résultats.

Certains organismes peuvent se créer des scripts ou des outils internes personnalisés pour effectuer leurs actions de préservation, tandis que d'autres peuvent avoir recours à des **logiciels ouverts** ou commerciaux développés par des tiers. Toutefois, pour les actions de préservation individuelles, la plupart des outils de flux de tâches pour la préservation (y compris les logiciels commerciaux) utilisent plusieurs des outils libres mentionnés précédemment, notamment Siegfried et JHOVE. Archivematica (https://archivematica.org/fr/) est un exemple de ce type de logiciel; il s'agit d'une application libre pour le flux de tâches qui produit des données sous forme de paquets prêts pour la préservation et le stockage à long terme. Archivematica comprend des processus pour créer et valider les sommes de contrôle, pour effectuer les tâches d'identification des formats de fichiers, pour exécuter la validation et la normalisation des formats pour la préservation et l'accès ainsi que des processus pour se connecter aux systèmes de stockage, ce qui permet de déposer des fichiers dans des espaces de stockage à long terme. Il regroupe également les métadonnées de préservation en utilisant les normes METS et PREMIS XML. Lorsqu'un établissement définit ses priorités en matière de préservation et comprend bien les collections qu'il souhaite préserver, il prend des décisions plus éclairées sur les outils de préservation à mettre en place et sur la façon de les configurer.  La prise de ces décisions permet de définir la stratégie et la planification de la préservation.

## Au niveau de la stratégie

Des méthodes ont aussi été créées pour faire le lien entre les outils comme Archivematica (https://archivematica.org/fr/) et des systèmes et logiciels dédiés aux données de recherche. Par exemple, une intégration entre la plateforme logicielle Dataverse (un logiciel pour les dépôts de données de recherche) et Archivematica permet aux responsables de la préservation de sélectionner et de traiter les jeux de données de recherche indépendamment du logiciel du dépôt, ce qui signifie que le stockage et la gestion des données de recherche déposées dans une collection Dataverse peuvent se faire à l'intérieur du cadre de la stratégie de préservation plus large de leur établissement. Pour plus d'informations sur la plateforme logicielle Dataverse et sur Archivematica, consultez l'article *Integrating Dataverse and Archivematica for Research Data Preservation* de Meghan Goodchild et Grant Hurley.

En revanche, les hôtes des installations de Dataverse peuvent aussi offrir des fonctionnalités de préservation. Par exemple, l'application Borealis (https://borealisdata.ca/fr/) (qui est une instance de Dataverse hébergée au Canada) comprend une stratégie de préservation au niveau des bits qui implique une vérification régulière de l'intégrité et un stockage répliqué sur différents périphériques. Une autre tâche de la personne responsable de la préservation est d'évaluer quels types d'actions sont  nécessaires pour l'ensemble des collections administrées par l'établissement. Par exemple, un établissement peut être à l'aise de s'appuyer sur une stratégie de préservation de base au niveau des bits pour les données qu'il administre pour une brève période ou qu'il ne

considère pas comme essentielles à ses collections institutionnelles. D'autres pourraient élaborer une politique d'évaluation ou d'archivage qui précise des exigences pour les jeux de données qui seront préservés à long terme. Les deux approches peuvent aussi être utilisées conjointement pour différentes collections; une stratégie au niveau des bits pourrait suffire pour du matériel à faible risque ou à valeur limitée tandis qu'une approche plus avancée qui utilise Archivematica pourrait être nécessaire pour le matériel de plus grande valeur pour l'établissement. Les mêmes questions s'appliquent aux choix de stockage de préservation, comme discuté plus tôt dans la section sur « Les sommes de contrôle, la préservation au niveau des bits et le stockage de préservation. » À ce niveau, la planification de la préservation nécessite l'établissement de politiques, plans et autres documentations. Consultez l'ouvrage *Digital Preservation Policy and Strategy : Where Do I Start?* de Christine Madsen et Megan Hurst pour une bonne introduction sur le sujet.

# Conclusion

Les données de recherche stockées sous forme numérique  font face à une variété de menaces qui jouent sur leur accessibilité à long terme. Ces menaces peuvent inclure la détérioration des fichiers eux-mêmes ou la perte des connaissances qui sont nécessaires pour consulter ou bien comprendre les objets numériques. Heureusement, de nombreuses normes et pratiques ont été développées pour atténuer ces risques. De telles interventions peuvent être à la fois de nature technique ou sous forme de politique, mais elles nécessitent deux éléments. Le premier est une planification méticuleuse, car il est difficile, voire impossible, de reconstituer les connaissances techniques nécessaires pour bien comprendre l'objet numérique si celles-ci ont été oubliées. Le deuxième est d'avoir une bonne connaissance de la Communauté d'utilisateurs cible – le groupe pour lequel les données ont été préservées. Ce savoir permet aux responsables de la préservation d'appliquer les actions appropriées pour assurer que les données restent compréhensibles, significatives et authentiques pour les personnes  auxquelles elles sont destinées.

**Questions de réflexion**

1. Quelles sont les menaces à la pérennité des données de recherche au fil du temps? Ces menaces varient-elles selon le type de données?
2. Pouvez-vous envisager un scénario où un établissement choisirait de prendre certaines actions de préservation plutôt que d'autres? Par exemple, pourquoi un établissement choisirait-il de s'engager dans la génération et la vérification de sommes de contrôle, mais ne

pas procéder à la normalisation des formats de fichier?

3. Prenez comme exemple un jeu de données qui vous est familier. Réfléchissez ensuite aux personnes qui pourraient vouloir accéder à ces données. Quelles questions les données peuvent-elles provoquer et pourquoi? Est-ce à propos du type de logiciel nécessaire pour ouvrir les fichiers du jeu de données ou est-ce au sujet de l'origine des données et de la façon dont elles ont été recueillies?

   Prenons maintenant les mêmes personnes dix ans plus tard. Pensez-vous que les questions des membres de ce groupe seront toujours les mêmes ou leurs préoccupations auront changé? Si oui, de quelle façon?

**Éléments clés à retenir**

- Les menaces courantes pour les données comprennent : l'obsolescence des médias, la dégradation des médias, l'obsolescence des formats et la perte de la provenance.
- Les actions potentielles de préservation comprennent : les sommes de contrôle et la préservation au niveau des bits, l'extraction technique des métadonnées, la validation des formats de fichiers ainsi que la normalisation et la migration.
- En évaluant les actions de préservation, tenez compte (1) des risques que vous traitez et (2) du rapport coût-efficacité de l'action.
- L'efficacité des actions de préservation peut varier selon qu'elles sont appliquées aux fichiers ou collections, aux systèmes ou dépôts ou à l'échelle de l'établissement.

## Lectures et ressources supplémentaires

Addis, M. (2020). *Which checksum algorithm should I use?* Digital Preservation Coalition. http://doi.org/10.7207/twgn20-12 (http://doi.org/10.7207/twgn20-12)

Borealis. (2022). *Plan de préservation*. https://borealisdata.ca/planpreservation (https://borealisdata.ca/planpreservation)

Dorey, J., Hurley, G. et Knazook, B. (2022). *Guide d'évaluation pour la préservation des données de recherche*. Zenodo. https://zenodo.org/record/6283886 (https://zenodo.org/record/6283886)

Goodchild, M. et Hurley, G. (2019). Integrating Dataverse and Archivematica for research data preservation. Dans M. Ras, B. Sierman et A. Puggioni (dir.), *iPRES 2019: 16th international conference on digital preservation* (p. 234-244). https://osf.io/wqbvy (https://osf.io/wqbvy)

Lavoie, B. (2014). *The Open Archival Information System (OAIS) reference model: Introductory guide* (2e éd.). Digital Preservation Coalition Technology Watch Report.

Madsen, C. et Hurst, M. (2019). Digital preservation policy and strategy: Where do I start? Dans J. Myntti et J. Zoom (dir.), *Digital preservation in libraries: Preparing for a sustainable future* (p. 37-47). ALA Editions Core, American Library Association.

# Bibliographie

Bettivia, R. S. (2016). The power of imaginary users: Designated communities in the OAIS reference model. *Proceedings of the Association for Information Science and Technology, 53*(1), 1-9.

CCSDS. (2012). *Modèle de référence pour un Système ouvert d'archivage d'information (OAIS)*. Pratique recommandée CCSDS 650.0-M-2 (F). https://public.ccsds.org/Pubs/650x0m2%28F%29.pdf (https://public.ccsds.org/Pubs/650x0m2%28F%29.pdf)

DANS. (2022, 20 juin). *File formats*. https://dans.knaw.nl/en/file-formats/ (https://dans.knaw.nl/en/file-formats/)

Digital Preservation Coalition. (2015). *Manuel de préservation numérique* (2e éd.). https://www.dpconline.org/docs/digital-preservation/handbook/translations-3/2519-handbook-2021-fr/file (https://www.dpconline.org/docs/digital-preservation/handbook/translations-3/2519-handbook-2021-fr/file)

Ledoux, T., de La Houssaye, J., Reecht, S., Caron, B., Phillips, M., Bailey, J., Goethals, A. et Owens, T. (2019). *NDSA Levels of Preservation version 1, traduction française*. HAL. https://bnf.hal.science/hal-02162334 (https://bnf.hal.science/hal-02162334)

Marks, S. (2015). *Becoming a trusted digital repository*. Trends in Archives Practice Module 8. Society of American Archivists.

McGovern, N. (2016). *Digital preservation management model document*. https://dpworkshop.org/

workshops/management-tools/policy-framework/model-document (https://dpworkshop.org/workshops/management-tools/policy-framework/model-document)

Owens, T. (2018). *The theory and craft of digital preservation*. Johns Hopkins.

Schaefer, S., McGovern, N., Goethals, A., Zierau, E. et Truman, G. (2018). *Digital preservation storage criteria, version 3*. http://osf.io/sjc6u/ (http://osf.io/sjc6u/)

Wheatley, P. (2018, 11 octobre). *A valediction for validation?* Digital Preservation Coalition Blog. https://www.dpconline.org/blog/a-valediction-for-validation (https://www.dpconline.org/blog/a-valediction-for-validation)

## À propos des auteurs

### Grant Hurley

Grant Hurley est bibliothécaire des canadiana à la bibliothèque des livres rares Thomas Fisher où il est responsable de la conservation des imposantes collections canadiennes, à la fois imprimées et manuscrites. De 2016 à 2022, Grant était bibliothécaire en préservation numérique à Scholars Portal et il était responsable d'infrastructures et de services partagés tels que le service Permafrost, le dépôt numérique fiable de Scholars Portal et Borealis, le dépôt Dataverse canadien. Il est également chargé de cours à la Faculty of Information de l'Université de Toronto pour le cours *Digital Archives Workflows*. En 2021, il a reçu le prix Alexander Fraser de l'Archives Association of Ontario pour services exceptionnels rendus à la communauté archivistique.

### Steve Marks

Steve Marks est bibliothécaire en préservation numérique aux bibliothèques de l'Université de Toronto où il est responsable d'alimenter et de s'occuper de la bibliothèque d'objets numériques. Auparavant, Steve a occupé des postes à l'Université York, à la University Health Network et à Scholars Portal. À Scholars Portal, Steve a été responsable de la certification réussie de Scholars Portal en tant que dépôt numérique fiable et a été le premier administrateur du Dataverse de Scholars Portal, maintenant connu sous le nom de Borealis.

12.

# PLANIFICATION DE LA GESTION DES DONNÉES POUR LES PROCESSUS DE TRAVAIL EN SCIENCE OUVERTE

Felicity Tayler; Mélanie Brunet; Kathleen Gregory; Lina Harper; et Stefanie Haustein

**Objectifs d'apprentissage**

À la fin de ce chapitre, vous pourrez :

1. Décrire la science ouverte comme un mouvement comprenant le partage et la réutilisation des données en tant que pratiques exemplaires.
2. Énoncer vos propres motivations axées sur la recherche en ce qui a trait au partage et à la citation des données.
3. Rédiger un plan de gestion des données qui décrit une approche de science ouverte pour des méthodes mixtes en sciences sociales.
4. Établir un lien entre les plans de gestion des données et leur relation avec les organismes de financement nationaux dans le contexte canadien et international.
5. Comprendre le concept de propriété intellectuelle relativement aux options d'attribution de licences de données ouvertes.

**Évaluation préliminaire**

*Un élément interactif H5P a été exclu de cette version du texte. Vous pouvez le consulter en ligne ici : https://ecampusontario.pressbooks.pub/gdrcanada/?p=54#h5p-7 (https://ecampusontario.pressbooks.pub/gdrcanada/?p=54#h5p-7)*

# Introduction

Ce chapitre aborde le sujet d'actualité de la **science ouverte** du point de vue de la gestion des données de recherche (GDR) en appui aux **données ouvertes** en sciences sociales et dans des contextes disciplinaires connexes. Nous présenterons un exemple de **plan de gestion de données** (PGD) à méthodologie mixte (qualitative et quantitative) pour vous aider à planifier un flux de travail en science ouverte. D'autres sujets font écho à ceux abordés dans d'autres chapitres de ce manuel, car les processus de travail en science ouverte et la GDR aux fins de partage et de réutilisation des données sont étroitement liés. À la conclusion de ce chapitre, il sera question de la propriété intellectuelle en ce qui a trait à la définition de la propriété des données, du droit d'auteur, de l'octroi de licences et de permissions, autant d'éléments qui ont une incidence sur la pratique des données ouvertes et des processus de travail en science ouverte.

Le PGD présenté en tant qu'étude de cas est issu d'un exemple réel du projet de recherche *Meaningful Data Counts* (MDC) dont les chercheuses principales proviennent de l'Université d'Ottawa et de l'Université de Kiel, en Allemagne. Ce partenariat international a pour but d'améliorer la compréhension du rôle des jeux de données dans la communication scientifique. Le projet génère des données empiriques sur les pratiques de données ouvertes, notamment la réutilisation et la citation des **données de recherche**. Voilà des éléments qui sont essentiels à l'élaboration d'indicateurs significatifs d'impact des données et à la valorisation des données de recherche à titre de résultats scientifiques de premier ordre. Le projet MDC nous renseigne sur les motivations et les comportements en matière de partage de données. L'approche à méthodes mixtes de la recherche constitue une étude de cas utile pour illustrer, concrètement, ce à quoi ressemble un flux de travail de recherche en science ouverte dans un PGD. Ce dernier a été partagé en tant que modèle intégré dans l'Assistant PGD de l'Alliance de recherche numérique du Canada (l'Alliance). L'Assistant PGD est un outil

gratuit en ligne mis à la disposition des chercheuses et chercheurs qui aide à la création d'un PGD sur la base de questions élémentaires en gestion de données appuyées par des pratiques exemplaires et des exemples.

Souvent, une chercheuse ou un chercheur décidera de partager ses données ou de participer aux pratiques de la science ouverte en fonction des normes disciplinaires. Ce chapitre se concentre sur les processus de travail en science ouverte et le partage de données en sciences sociales et dans des domaines connexes. Ces principes et ces pratiques sont transférables à d'autres domaines qui travaillent avec des méthodes qualitatives et quantitatives. Toutefois, il est important de souligner que la science ouverte est définie différemment selon les contextes disciplinaires. Par exemple, le présent chapitre n'aborde pas de pratiques particulières au domaine biomédical, comme l'enregistrement à des essais cliniques, les revues systématiques ou d'autres types d'études (exigeant un enregistrement) et le recours à des lignes directrices de documentation.[1]

Avant de passer à l'étude de cas donnée en exemple et aux pratiques exemplaires pour les processus de travail en science ouverte utilisant une approche de méthodologie mixte (quantitative et qualitative), définissons quelques expressions. La dernière section de ce chapitre traite des éléments à prendre en considération en matière de propriété intellectuelle, pratiques éthiques essentielles au travail avec des données ouvertes.

## Qu'est-ce que la science ouverte?

Vous avez peut-être entendu l'expression *science ouverte* employée dans des contextes différents et parfois contradictoires puisqu'il y a abondance d'approches selon les professions, les politiques, les articles et les mandats. De nature générique, elle est comprise par des personnes différentes de manières différentes et les discussions sous-tendent des points de vue différents qui ont tous leurs propres hypothèses, buts et affirmations. À la lumière du projet de recherche MDC, qui nous sert d'étude de cas pour démontrer comment les pratiques exemplaires en **gestion des données de recherche** (GDR) peuvent soutenir un processus de travail en science ouverte, nous définissions la science ouverte du point de vue du scientifique, ou du professionnel, de la façon suivante: « le mouvement ayant pour but de rendre la recherche scientifique, les données et la diffusion accessibles à tous les niveaux d'une société curieuse[2] » [traduction] (FOSTER, s.d.-b). Cette définition est le fruit de FOSTER (https://www.fosteropenscience.eu/), un projet européen dédié à

1. Une étude Delphi a identifié 19 pratiques de science ouverte dans le domaine biomédical. Les autrices tiennent à remercier David Moher du Centre for Journalology de l'Institut de recherche de L'Hôpital d'Ottawa pour les conversations entretenues au sujet des pratiques en science ouverte dans plusieurs disciplines. Cobey, K. D., Haustein, S., Brehaut, J. Dirnagl, U., Franzen, D. L., Hemkens, L. G., Presseau, J., Riedel, N., Strech, D., Alperin J. P., Costas, R., Sena, E. S., van Leeuwen, T., Ardern, C. L., Bacellar I. O. L, Camack, N., Correa, M. B., Buccione, R., Cenci, M. S., ... Moher, D. (2022). Establishing a core set of open science practices in biomedicine: A modified Delphi study. medRxiv. https://doi.org/10.1101/2022.06.27.22276964 (https://doi.org/10.1101/2022.06.27.22276964)
2. "*the movement to make scientific research, data and dissemination accessible to all levels of an inquiring society.*"

favoriser la mise en œuvre concrète de la science ouverte. Parce qu'il existe de très nombreuses manières d'y arriver, FOSTER adopte une approche taxonomiste pour cartographier un large champ d'activités et de résultats en lien avec ces pratiques. Par exemple, la pratique de la science ouverte comprend le **libre accès** aux publications, rend les données librement accessibles et réutilisables, a recours à des outils ouverts, participe à la science citoyenne et compte des méthodes ouvertes d'évaluation de la recherche.

La gamme complète d'activités possibles et de résultats en science ouverte est souvent réduite aux publications en libre accès. Toutefois, la science ouverte cherche à rendre l'entièreté du processus de recherche transparent et accessible – non seulement la publication finale! Qui plus est, l'importance des normes disciplinaires qui façonnent concrètement et de manière différente la manière de « faire » de la science ouverte est souvent négligée. Il s'agit là d'un problème, car des disciplines différentes disposent de normes différentes et de moyens différents pour rendre les publications ouvertes et pour partager les données. Les chercheuses et chercheurs réutilisent les données ouvertes à des fins variées. Les jeux de données existants peuvent servir à établir une nouvelle étude de cas, à enseigner des méthodes computationnelles en classe, à étalonner des instruments, à créer des modèles ou alimenter des algorithmes. C'est pourquoi les pratiques exemplaires en GDR recommandent aux chercheuses et chercheurs de verser les données dans un dépôt, car cette infrastructure est plus fiable sur le plan du stockage à long terme et de la gestion d'**identifiants pérennes** (p. ex., DOI) grâce auxquels d'autres personnes peuvent trouver puis citer les jeux de données. Toutefois, les chercheuses et chercheurs peuvent également partager les données par le biais de sites Web personnels, d'une personne à l'autre ou par l'entremise d'énoncés de disponibilité des données dans des articles.

Ce chapitre s'inspire d'une approche interdisciplinaire de méthodes mixtes pour le partage des données qui peut être utilisée dans de nombreux domaines d'études différents. Cependant, l'exploration de la science ouverte ne s'arrête pas là; il existe plusieurs autres moyens de mettre en application les pratiques ouvertes dans divers domaines d'études.

## Que sont les données ouvertes?

FOSTER (s.d.-a) définit les données ouvertes de la façon suivante: « données en ligne, gratuites, accessibles, que l'on peut utiliser, réutiliser et distribuer pour autant que la source des données est indiquée[3] » [traduction]. L'accessibilité n'est qu'un des éléments de cette équation : les données doivent également être dans un format utilisable (Boulton *et al*., 2011, cité dans Fecher et Friesike, 2014). Les pratiques exemplaires en GDR permettent aux données ouvertes d'être utilisables par le biais des **principes FAIR**, abordées au chapitre 2, « Les principes FAIR et la gestion des données de recherche » (Wilkinson *et al.*, 2016). Les

3. "*online, free of cost, accessible data that can be used, reused and distributed provided that the data source is attributed*."

données doivent être facilement trouvables, accessibles, interopérables et réutilisables, et ce, en mettant l'accent sur l'**interopérabilité** technologique. Rendre les données conformes aux principes FAIR n'est qu'une partie de la solution des pratiques exemplaires en GDR. Le partage et la réutilisation des données exigent, eux aussi, un contexte offert par des renseignements supplémentaires comme la documentation des données et les **métadonnées**.

**Toutes les données ne peuvent pas être ouvertes**. Les données avec des enjeux de confidentialité, comme les renseignements personnels, doivent demeurer restreintes. Les pratiques exemplaires en GDR mettent au premier plan toute une gamme d'approches en science ouverte tout en trouvant un équilibre entre **les données « aussi ouvertes que possible, aussi fermées que nécessaire. »**

Le partage et la réutilisation des données (ouvertes) sont des concepts importants en appui à la science ouverte, la préférence étant accordée aux données ouvertes lorsque les normes éthiques le permettent. Les avantages perçus du partage et de la réutilisation des données reflètent les avantages potentiels de la science ouverte : rendre la recherche plus reproductible et transparente afin d'économiser temps et argent et de réunir des données jusqu'alors isolées afin de former des combinaisons inédites. Les recommandations de l'UNESCO au chapitre de la science ouverte mettent en évidence son potentiel transformateur et son importance lorsqu'il est question de résoudre les problèmes actuels les plus difficiles comme les changements climatiques, les enjeux de santé, la pauvreté et les inégalités croissantes.

Les prochaines sections définissent l'étude de cas MDC et un modèle de plan de gestion des données (PGD). Il y est question de la mise en application des principes de science ouverte et de pratiques relatives à la documentation. Cette documentation, qui comprend un PGD, permet la collaboration entre personnes qui doivent comprendre les données et leur donner un sens afin qu'elles puissent être réutilisées adéquatement.

## Étude de cas : le projet *Meaningful Data Counts*

Le projet de recherche MDC est une étude de cas utile en matière de pratiques exemplaires en gestion des données de recherche (GDR), car il étudie à la fois les pratiques de données ouvertes à travers les disciplines et met en pratique la science ouverte par le biais d'une approche de méthodes mixtes (quantitative et qualitative) en sciences sociales. La méthodologie incorpore de la bibliométrie, des réponses à un sondage et des entrevues.

Ce projet s'inscrit dans le cadre de l'initiative *Make Data Count (https://makedatacount.org/)* qui encourage l'adoption des pratiques de bases pour les mesures de données ouvertes: l'utilisation normalisée des données et des procédés de citation des données dans les dépôts et chez les maisons d'édition. Le MDC présente des données probantes empiriques sur le comportement en ce qui a trait à l'utilisation et à la citation des données afin d'améliorer la compréhension du rôle des jeux de données dans la communication scientifique. Les modèles de partage et de citation des données sont étudiés en fonction des disciplines universitaires et des

étapes de la carrière des chercheuses et chercheurs. Le MDC s'attarde également à ce qui motive les chercheuses et chercheurs à partager ou à citer des jeux de données – ou à ne pas le faire. S'il existe une grande variété de pratiques en matière de citation des données, la plupart des personnes ayant répondu au sondage affirment qu'elles citaient les données pour appliquer des pratiques de recherche « idéales », comme la reconnaissance de la dette intellectuelle, l'aide à la localisation et à l'accès aux données et l'appui à la validité de leurs propres affirmations (Gregory *et al*., 2023). Inversement, les obstacles au partage de données comprennent la peur qu'éprouvent les chercheuses et chercheurs de se faire damer le pion, la crainte de voir leurs erreurs exposées au grand jour, la perception selon laquelle les efforts nécessaires pour préparer et publier les jeux de données n'en valent pas le coup et la croyance que le partage des données ne s'applique pas à leur recherche (Tenopir *et al*., 2020).

Le projet MDC met en œuvre le processus de travail en science ouverte afin de soulever les défis que doivent relever les membres d'équipe qui participent à des pratiques de science ouverte, notamment le partage et la citation de données de recherche. Par le partage des plans de recherche, des processus, du code, des résultats préliminaires et des données, un processus de travail en science ouverte tente, autant que possible, de rendre le déroulement de la recherche transparent aux personnes extérieures à l'équipe de recherche initiale.

Un élément clé de la pratique en science ouverte du MDC est l'élaboration d'un plan de gestion des données (PGD) avec la bibliothécaire en GDR de l'Université d'Ottawa; ce plan est un modèle de PGD endossé par l'Alliance. Comme vous l'avez vu au chapitre 1, « Les rudiments, » un PDG est un document qui décrit comment les données d'un projet de recherche seront manipulées, de la collecte à l'analyse puis comment elles seront traitées à la fin du projet. Les PGD sont des documents dynamiques qui peuvent être mis à jour tout au long de la vie du projet; cette approche itérative s'harmonise bien avec l'objectif de partage éthique des données. Les pratiques exemplaires en GDR sont essentielles aux universitaires qui adoptent la science ouverte et, de plus en plus, elles sont exigées pour en atteindre les objectifs (Tenopir *et al*., 2020). La Politique des trois organismes sur la gestion des données de recherche (https://science.gc.ca/site/science/fr/financement-interorganismes-recherche/politiques-lignes-directrices/gestion-donnees-recherche/politique-trois-organismes-gestion-donnees-recherche), par exemple, soutient les principes directeurs FAIR (facile à trouver, accessible, interopérable et réutilisable) pour la gestion et l'intendance des données de recherche. Les trois organismes fédéraux de financement de la recherche (IRSC, CRSNG et CRSH) intègrent le partage des données à leur processus de demande de subvention (dans la section « Mobilisation des connaissances »). Il est attendu que, pour obtenir du financement, les futures demandes de subvention devront être accompagnées d'un PGD bien défini.

Le PGD de MDC (https://zenodo.org/records/6473351) décrit la manière dont le projet gère divers types de données recueillies et analysées par l'équipe de recherche. Le PGD est une des méthodes qu'emploie l'équipe pour documenter le processus de travail du projet afin de communiquer les protocoles en matière d'éthique, le transfert de fichiers et les procédures de stockage, les normes relatives aux métadonnées et le code

informatique avec les membres de l'équipe qui travaillent à distance. Anton Ninkov, un postdoctorant membre de l'équipe de recherche responsable de gérer les données, a mentionné que la documentation du processus de travail s'apparente « à réfléchir au projet comme étant quelque chose de plus large qu'une tâche individuelle. Il s'agit du fonctionnement du projet dans son ensemble – son travail n'en est qu'une partie[4] » [traduction] (communication personnelle, 15 février, 2022).

Les jeux de données de MDC comprennent l'analyse bibliographique des pratiques de citation de données dans un corpus composé de 8 643 593 jeux de données dans DataCite (Ninkov *et al*., 2022), de réponses au sondage provenant de plus de 2500 chercheuses et chercheurs qui ont réfléchi aux pratiques relatives au partage et à la citation des données dans diverses disciplines et entrevues semi-structurées qui donnent davantage d'information sur ce qui les motive à partager et à citer les données – ou à ne pas le faire. Le PGD abordé à la section suivante présente la gestion de tous les jeux de données générés à la suite d'analyses bibliométriques, de sondages et d'entrevues; le but est de partager les données avec une licence ouverte tout au long du cycle de vie du projet, et non seulement au moment de la publication.

## Pratiques exemplaires pour un plan de gestion des données en appui à un processus de travail en science ouverte

Un PGD constitue une excellente occasion de mettre l'accent sur les pratiques de la science ouverte, comme le partage et la réutilisation des données; il peut également soutenir d'autres éléments d'un processus de travail de science ouverte. Lorsque les processus sont reliés par votre PGD à d'autres éléments de votre projet de recherche, vous rendez votre recherche accessible à diverses étapes du projet. De plus, les données qui étayent les résultats de la recherche présentés dans une publication sont transparentes et réplicables tout au long du processus – non seulement en fin de parcours. Plusieurs chercheuses et chercheurs se concentrent sur l'aspect de planification d'un PGD; un document est rédigé au début du projet puis ignoré. Toutefois, la recherche est rarement linéaire et les plans doivent souvent changer. La création de versions subséquentes peut s'avérer incroyablement utile afin de planifier le projet et de répertorier l'évolution du processus de recherche.

- La science ouverte met l'accent sur le partage et la réutilisation des données tout au long des projets de recherche, non seulement à l'étape finale (publication);
- Les processus de travail en science ouverte peuvent servir dans d'innombrables méthodes de recherche : méthodes mixtes, quantitatives et qualitatives, dans toutes les disciplines;

4. "*about thinking about the project as a bigger thing than an individual task. It's about the movement of the whole project, which my work is just one component of.*"

- Les versions mises à jour de votre PGD rendent compte de l'évolution de vos méthodes de recherche et de vos processus de travail.

Les pratiques en science ouverte de MDC ont mis de l'avant l'élaboration d'un plan de gestion des données (https://zenodo.org/records/6473351) exhaustif. La première version du plan, créée au début du projet, décrit comment une équipe de chercheuses et chercheurs originaires de plusieurs pays gérera divers types de données recueillies à l'aide de méthodes mixtes (qualitatives et quantitatives). Les PGD sont des documents dynamiques et l'équipe de MDC a récemment mis à jour le sien, conformément aux pratiques exemplaires en science ouverte, dont celle de l'examen régulier de la documentation des données et la confirmation qu'elles reflètent adéquatement les méthodes de recherche et les processus de gestion des données dont se sert l'équipe de recherche. La version 2 (https://doi.org/10.5281/zenodo.6473351) du PGD est versée dans le même dépôt.

La révision du PGD a aidé à la gestion efficace du projet. Stefanie Haustein, chercheuse principale, a constaté qu'en fin de compte, « certaines sections décrites dans le modèle de l'Assistant PGD ne s'appliquaient pas à son projet de recherche[5] » [traduction] (communication personnelle, 15 février, 2022). Le modèle utilisé en 2022 demandait aux chercheuses et chercheurs d'aborder la question de la préservation à long terme; toutefois, Mme Haustein remarque que « cette question n'est plus pertinente, car nous supposons que la technologie comme les **interfaces de programmation** (API) et la pertinence des données auront beaucoup évolué d'ici 20 ans[6] » [traduction] (communication personnelle, 15 février 2022). L'examen du PGD a favorisé une révision du processus de travail de l'équipe de recherche, notamment le travail des membres qui se sont joints à l'équipe après la publication de la première version du plan. Des changements avaient été apportés en matière de collecte et de traitement des données et il fallait que la documentation en rende compte. Il est important de coucher sur papier ces processus de travail méthodologiques dans le cadre de pratiques exemplaires en science ouverte, car pour comprendre et reproduire les données partagées, les membres extérieurs à l'équipe doivent disposer d'un certain contexte quant à la manière dont les données ont été collectées, structurées et analysées.

Les deux versions du PGD ont été créées à l'aide de l'outil recommandé par l'Alliance, l'Assistant PGD (https://dmp-pgd.ca/), en collaboration avec la bibliothécaire en GDR de l'Université d'Ottawa. L'équipe a également contribué à l'élaboration d'un modèle dans l'Assistant PGD (https://zenodo.org/records/4701021) en ce qui a trait aux processus de travail en science ouverte. Le modèle guide les équipes de recherche dans le choix des pratiques exemplaires qu'elles devraient inclure dans les PGD exigés par les organismes

5. *"Some sections prescribed by the DMP Assistant template did not apply to our research project after all."*
6. *"Long term preservation isn't as relevant to us, as we assume that the technology such as the APIs [application programming interfaces] and the relevance of the data will have changed in 20 years from now."*

subventionnaires. Le PGD de MDC a été examiné par des pairs, publié puis diffusé en tant qu'exemple national de pratique exemplaire en rédaction d'un PGD pour un processus de travail en science ouverte, une approche à méthodes mixtes et un partenariat de recherche international. Toutes les ressources de formation créées par l'Alliance sont sous licence CC BY-NC 4.0 (https://creativecommons.org/licenses/by-nc/4.0/deed.fr). Vous pouvez les partager et les adapter gratuitement selon vos besoins.

Cette section définit certaines des pratiques exemplaires incluses dans le PGD de MDC afin de documenter les processus et de favoriser la collaboration entre les membres de l'équipe ou avec d'autres personnes qui doivent comprendre les données et les doter de sens afin qu'elles puissent être réutilisées adéquatement. Cette liste n'est pas exhaustive; par conséquent, nous vous encourageons à consulter les sections « *Guidance* » de l'exemple de PGD (https://doi.org/10.5281/zenodo.4092122) pour obtenir plus d'informations.

Responsabilité et ressources

- Attribuez suffisamment de ressources humaines aux responsabilités d'intendance des données dans votre budget avant d'entamer la collecte de données. Habituellement, la chercheuse ou le chercheur principal est responsable de maintenir les normes d'accessibilité des données pour l'équipe. Affectez des gens à la structuration des données, à leur documentation et à la réponse aux questions portant sur l'accès à l'information et l'octroi d'un accès aux données;
- Créez un document d'intégration pour faire en sorte que tous les membres de l'équipe adhèrent aux mêmes processus de travail. Structures de fichiers logiques, conventions de nommage informatives et indications claires de la version des fichiers – autant d'éléments qui permettent une meilleure utilisation des données pendant et après le projet de recherche. Le recours à une feuille de travail pour la convention de nommage des fichiers peut s'avérer très utile;
- Documentez votre processus et révisez votre PGD, le cas échéant. Consultez régulièrement les membres de l'équipe pour saisir les éventuels changements apportés à la collecte, au traitement et à la publication des données qui doivent être reflétés dans la documentation.

Documentation et métadonnées

- Documentez les processus de travail à l'aide d'un fichier **LISEZ-MOI** qui accompagne tous les jeux de données. Une bonne documentation des données comprend de l'information au sujet de l'étude, une description des données et de tout autre renseignement contextuel nécessaire pour que d'autres chercheuses ou chercheurs puissent se servir des données;
- Utilisez des **formats de fichier ouverts** ou conformes aux normes de l'industrie (p. ex., celles utilisées couramment par la communauté) dans la mesure du possible;
- Utilisez des **schémas de métadonnées** spécifiques aux jeux de données ouvertes ou toute autre norme de métadonnée particulière au domaine. La documentation des jeux de données devrait être en format

ouvert et lisible par machine afin de permettre un échange efficace d'information entre les systèmes et les personnes qui les utilisent. DataCite a créé un ensemble de champs de métadonnées essentiels et des instructions pour rendre les jeux de données faciles à identifier et à citer.

Éthique et conformité légale

- Les processus de travail en science ouverte priorisent d'être « aussi ouverts que possible, aussi fermés que nécessaire ». Réfléchissez aux types de données qui doivent être partagés afin de satisfaire aux exigences des établissements ou des organismes subventionnaires et aux données dont l'accès devrait être limité en fonction d'enjeux liés à la confidentialité, à la vie privée ou à la propriété intellectuelle tels que décrits dans votre protocole éthique;
- Demandez le consentement adéquat auprès des participantes et participants afin que leurs données puissent être partagées. Votre énoncé de consentement éclairé peut préciser certaines conditions clarifiant l'utilisation des données. Informez les personnes participant à vos études si vous avez l'intention de publier une version anonymisée et dépersonnalisée des données recueillies et faites en sorte qu'elles acceptent ces modalités;
- Utilisez des licences ouvertes (p. ex., CC BY) pour favoriser le partage et la réutilisation des données. Les licences déterminent de quelles manières d'autres personnes peuvent utiliser vos données. Pensez à inclure une copie de votre licence d'utilisation dans votre PGD (sujet abordé plus loin).

Mobilisation des connaissances

- Aidez les gens à réutiliser et à citer vos données. Saviez-vous qu'un jeu de données constitue un résultat de recherche que vous pouvez ajouter à votre curriculum vitae, au même titre qu'un article scientifique? Si vous publiez vos données dans un dépôt de données (p. ex., Zenodo, Borealis, Dryad), d'autres personnes peuvent les trouver et s'en servir. Les **identifiants numériques d'objets** (DOI) uniques constituent un excellent moyen d'identifier et de citer des jeux de données;
- Servez-vous des médias sociaux, des bulletins d'information électroniques, des affiches, des conférences, des webinaires, des forums de discussions, ou des forums spécifiques à votre discipline pour mettre en lumière vos données publiées, promouvoir la transparence et encourager la découverte ainsi que la réutilisation des données. Citez vos jeux de données comme vous le faites avec d'autres types de publications.

# Qu'est-ce qui constitue une donnée ouverte? Limites en matière de partage de données

L'étude de cas de MDC établit un lien entre le partage de données et les PGD, car ils fonctionnent de concert en appui aux pratiques de science ouverte à l'échelle d'un projet de recherche. Cette section aborde les modalités légales et contractuelles qui permettent ou limitent le partage et la réutilisation des données alors qu'elles circulent dans les infrastructures numériques. Après un aperçu des éléments à prendre en considération en matière de confidentialité dans le cadre du projet de MDC, cette section se concentrera sur la propriété intellectuelle au moment de déterminer la propriété des données et le partage des données de recherche[7]. Alors que la discussion sur la propriété intellectuelle et les licences des données s'inscrit dans un contexte canadien, le PGD du MDC énonce clairement comment l'accès aux données sensibles sera limité dans le cadre d'un projet de recherche international. Il définit également comment les données qui ont été anonymisées seront partagées à l'aide d'une licence ouverte, ce qui permettra la réutilisation du jeu de données.

Une licence est une permission qu'octroie le titulaire du droit d'auteur à un tiers d'utiliser ses œuvres (dans ce cas, des données sous une forme ou une autre) à certaines fins et sous certaines conditions. Le droit d'auteur demeure avec son titulaire (Office de la propriété intellectuelle du Canada, 2019). Après avoir déterminé si les données sont protégées par des droits d'auteurs (et, le cas échéant, qui les détient et si elles peuvent être partagées), vous pouvez utiliser diverses licences ouvertes pour indiquer leur degré d'ouverture. Leur rôle est double : informer les personnes qui utiliseront les données des droits conservés par les titulaires des droits d'auteur et indiquer comment utiliser les œuvres sans devoir demander aux titulaires la permission à chaque occurrence. La propriété des droits d'auteur ne change pas. La licence ouverte indique simplement que le titulaire des droits d'auteur libère son œuvre de certaines des limites habituelles pour que vous puissiez la partager, la remixer et la réutiliser en toute légalité, tant et aussi longtemps que vous respectez les conditions de la licence. Plusieurs dépôts permettent de sélectionner facilement une licence ouverte et ajoutent l'information dans les métadonnées.

Si le partage des données constitue la pierre angulaire de la science ouverte, il n'est peut-être pas toujours conseillé, sécuritaire ou même légal de le faire. Les pratiques exemplaires en science ouverte accordent la

7. Des éléments de cette section sur la propriété intellectuelle sont adaptés de M. Brunet, J. Hatherill et C. Ripp. (2021). *Libre accès aux connaissances Partie 2 : Partagez vos données de recherche*, Bibliothèque de l'Université d'Ottawa, CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/deed.fr), http://hdl.handle.net/10393/43308 (http://hdl.handle.net/10393/43308) et de M. Brunet et T. Rouleau. (2021). *Droit d'auteur et données de recherche à l'Université d'Ottawa : Questions fréquemment posées*, Bibliothèque de l'Université d'Ottawa, CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/deed.fr), https://www.uottawa.ca/library/sites/g/files/bhrskd381/files/2022-12/droit_dauteur_et_donnees_de_recherche_faq.pdf (https://www.uottawa.ca/library/sites/g/files/bhrskd381/files/2022-12/droit_dauteur_et_donnees_de_recherche_faq.pdf).

priorité au respect des limites de nature éthique et légale en matière d'accès aux données pour équilibrer les objectifs plus larges de partage, de publication et de réutilisation des données. Afin de respecter ces pratiques exemplaires, vous devrez prendre en considération quels sont les types de données qui doivent être partagées afin de satisfaire aux exigences provenant des établissements ou des organismes subventionnaires et quelles types de données doivent avoir un accès restreint en raison d'éléments relatifs à la confidentialité, à la vie privée ou à la propriété intellectuelle énoncés dans votre protocole d'éthique. En effet, avant de rendre des données publiques et ouvertes, il est essentiel que vous déterminiez si vous pouvez le faire de manière éthique et en toute légalité. La sécurité et la vie privée des participantes et participants, la souveraineté des données autochtones et la nature confidentielle ou propriétaire des données peuvent limiter votre capacité à les partager. Vous devez également vérifier le statut des droits d'auteur en ce qui a trait à la propriété des données.

Dans notre étude de cas, le PGD de MDC indique que toutes les données et publications finales seront publiées en accès libre. Pour y arriver, le partenariat international multiétablissements doit également se conformer aux politiques de GDR de ses établissements hôtes, lesquels prennent en considération la législation pertinente, les normes de l'industrie et les pratiques exemplaires. En particulier, les processus de travail des données tiendront compte des éléments juridiques et éthiques de l'Université d'Ottawa et de l'*Énoncé de politique des trois conseils : Éthique de la recherche avec des êtres humains (https://ethics.gc.ca/fra/policy-politique_tcps2-eptc2_2022.html)* – **EPTC 2** (2022); toutefois, ils peuvent également se référer à la politique sur l'intégrité et l'éthique en recherche de l'Université de Kiel si l'ETPC 2 n'offre pas suffisamment d'orientations. La cochercheuse est affiliée à des établissements européens. Par conséquent, les méthodes de recherche doivent se conformer au Règlement général sur la protection des données (RGPD) (https://www.cnil.fr/fr/reglement-europeen-protection-donnees) de l'UE, lequel est plus limitatif que ses équivalents canadiens.

L'équipe de recherche a stocké les données sensibles sur un serveur sécurisé au Canada. Seules les chercheuses principales ont accès à la totalité du projet. Les autres membres de l'équipe avaient un accès limité lorsqu'ils travaillaient à la collecte des données et à l'anonymisation des données sensibles. La collecte de données qualitatives et personnelles a respecté l'approbation formelle en matière d'éthique du comité d'éthique de la recherche de l'Université d'Ottawa qui demandait d'obtenir le consentement explicite et éclairé des personnes participantes en utilisant le *Recommended Informed Consent Language for Data Sharing (https://www.icpsr.umich.edu/web/pages/datamanagement/confidentiality/conf-language.html)* (ICPSR, s.d.). Les médias sociaux et autres données en ligne publiques ont été recueillies et gérées en fonction du document *Internet Research: Ethical Guidelines 3.0 (https://aoir.org/reports/ethics3.pdf)* de l'Association of Internet Researchers (franzke *et al.*, 2019). Toute donnée jugée sensible est stockée en sécurité avec mot de passe et chiffrement. Les données sont anonymisées dans les publications découlant du projet, sauf entente explicite de les publier autrement. Une fois les données anonymisées, elles peuvent être partagées sous forme de données ouvertes avec une licence Creative Commons Attribution (CC BY) 4.0 International. Si ce n'est pas possible, l'équipe peut choisir une licence plus restrictive, Creative Commons Attribution – Pas de Modification (CC BY-ND).

# Puis-je partager les données? Définir la propriété des données

Vous vous demandez peut-être pourquoi l'équipe de recherche doit attribuer une licence à ses données pour les rendre ouvertes. Les données sont-elles même protégées par des droits d'auteur? Parce que les droits d'auteur protègent l'expression originale des idées ou des faits fixés sur un support tangible, il est facile de conclure que les données sont des faits et donc non protégées. En effet, de manière générale, les données brutes ou factuelles sont des données non interprétées qui ne sont pas protégées par le droit d'auteur. Toutefois, une compilation de données peut être protégée en raison du jugement, de la compétence ou de l'effort nécessaire pour déterminer les données à inclure et leur organisation (faisant des données une « expression originale »). De plus, si les données sont de nature littéraire, musicale, dramatique ou artistique, elles peuvent être protégées par le droit d'auteur. Le tableau 1 ci-dessous résume les types de données susceptibles d'être protégées par le droit d'auteur.

**Tableau 1 : Types de données et protection par le droit d'auteur**

| **Non protégées par le droit d'auteur** | **Peuvent être protégées par le droit d'auteur** |
|---|---|
| Donnée individuelle ou brute (c'est-à-dire, un chiffre ou une mesure) | Représentations de données (p. ex., tableaux et graphiques) |
| | Jeux de données |
| | Compilations de données |
| | Bases de données |
| | Données achetées (sous conditions d'utilisation) |
| | Œuvres littéraires, musicales, dramatiques ou artistiques (p. ex., photos) |

Si les données sont protégées par le droit d'auteur, qui est le propriétaire? Si vous possédez des données générées ou fournies par un tiers, même si elles sont accessibles gratuitement, cela ne signifie pas que vous possédez un droit d'auteur sur elles. Vérifiez toujours s'il y a une licence ou lisez les modalités d'utilisation. Le tableau 2 résume la propriété des droits d'auteur selon les types de données.

**Tableau 2 : Types de données et propriété des droits d'auteur**

| | |
|---|---|
| **Données primaires** | Données recueillies pour vos propres fins à partir d'expériences ou de recherches que vous avez menées et que vous avez fixées sur un support tangible. |
| Si des droits d'auteur existent, vous en êtes probablement le titulaire. Toutefois, vous devriez vérifier les ententes ou les contrats en lien avec votre projet de recherche. | |

| | |
|---|---|
| **Données secondaires** | Données recueillies à d'autres fins à partir d'expériences ou de recherches menées par d'autres. |
| S'il existe un droit d'auteur, il est probablement détenu par d'autres. | |
| **Données tertiaires** | Synthèse de données issues d'expériences ou de recherche menées par d'autres. |
| Articles, rapports, etc. rédigés par d'autres et dont vous ne détenez pas le droit d'auteur. | |

Plusieurs facteurs extérieurs à votre équipe de recherche ou à votre projet peuvent déterminer si des données sont protégées par le droit d'auteur et qui détient celui-ci.

- Des politiques ou des ententes contractuelles entre chercheuses ou chercheurs et établissements affiliés (p. ex., contrats d'embauche, conventions collectives);
- Des conventions ou des pratiques disciplinaires en matière d'attribution de paternité;
- Des politiques de l'agence ou de l'organisme qui finance la recherche (en partie ou en totalité);
- Des conditions ou des modalités d'utilisation de licence de données achetées – le fait d'acquérir des données auprès d'un tiers ne signifie pas que les droits d'auteur vous ont été transférés ou que vous êtes autorisé à partager les données.

Toutes les parties prenantes impliquées dans un projet de recherche devraient préciser les questions relatives aux données et au droit d'auteur en début de processus. Les statuts divers, qui se chevauchent parfois, des personnes qui collectent les données ou des membres de l'équipe de recherche, même au sein d'un établissement ou d'un organisme, constituent des facteurs importants pour déterminer qui détient les droits d'auteur des données de recherche. Il est vital de préciser la propriété des droits d'auteur, car les données protégées ne peuvent pas être davantage « ouvertes » sans la permission du détenteur des droits.

Il existe trois types principaux de licences ouvertes pour les données :

- Licences Creative Commons
- Licences Open Data Commons
- Licences du logiciel

Deux désignations Creative Commons sont utilisées fréquemment pour les données. Elles servent d'options dans les dépôts de données.

- CC BY 4.0 (licence Creative Commons Attribution 4.0 International (https://creativecommons.org/licenses/by/4.0/deed.fr)) : cette licence exige que l'autrice ou l'auteur soit crédité;
- CC0 (domaine public (https://creativecommons.org/publicdomain/zero/1.0/deed.fr)) : elle sert à indiquer que le titulaire des droits d'auteur renonce à ses droits sur l'œuvre. Lorsque les données sont du

domaine public, il n'y a plus de restrictions quant à leur utilisation ni à leur attribution. Certains dépôts de données, comme Borealis, proposent cette licence par défaut.

Pour des données sous forme de base de données, les licences Creative Commons s'appliquent à la fois au contenu de la base de données et à la base de données elle-même. Creative Commons ne recommande pas le recours aux conditions « Pas d'utilisation commerciale » (NC) ou « Pas de modification » (ND) pour les données, car elles limitent grandement l'utilisation à des fins savantes et scientifiques[8]. Bien que nous ne recommandions pas de limiter la réutilisation des données à des fins non commerciales, vous pouvez appliquer une licence Creative Commons Attribution – Pas d'utilisation commerciale (https://creativecommons.org/licenses/by-nc/4.0/deed.fr). Toutefois, il est important de souligner que cette condition s'applique généralement à l'*utilisation* et non à la *personne qui utilise*. En principe, elle ne préviendrait pas une entité commerciale d'utiliser les données si elle ne les revend pas ou si elle ne les utilise pas comme élément de base pour un produit ou service vendu à des fins profitables.

La Open Knowledge Foundation offre trois licences ouvertes dédiées aux bases de données, bien qu'elles ne soient pas disponibles dans tous les dépôts de données.

- ODbL 1.0 (*Open Data Commons Open Database License (https://opendatacommons.org/licenses/odbl/summary/)*)
- ODC-BY 1.0 (*Open Data Commons Attribution License (https://opendatacommons.org/licenses/by/summary/)*)
- PDDL 1.0 (*Open Data Commons Public Domain Dedication and License (https://opendatacommons.org/licenses/pddl/summary/)*)

Veuillez noter que les licences Open Data Commons s'appliquent aux bases de données uniquement et non pas au contenu d'une base de données.

Les licences de logiciel sont parmi les premières licences ouvertes; elles sont également utilisées dans les dépôts de données. Elles peuvent s'appliquer au logiciel ou au code en plus d'être associées aux fichiers de documentation connexes.

- Licence MIT (https://opensource.org/licenses/MIT)
- Licence GNU General Public version 3 (https://opensource.org/licenses/GPL-3.0)
- Licence Apache version 2.0 (https://opensource.org/licenses/Apache-2.0)

---

8. Consultez *Creative Commons Frequently Asked Questions about data and CC licences*, https://wiki.creativecommons.org/wiki/Data#Frequently_asked_questions_about_data_and_CC_licenses (https://wiki.creativecommons.org/wiki/Data#Frequently_asked_questions_about_data_and_CC_licenses).

Le tableau 3 ci-dessous compare les trois types de licences en fonction de ce qu'elles permettent et de la nécessité de citer correctement la source, du point de vue d'une *personne qui utilise des données* sous licence (et non du point de vue de la personne qui les crée).

**Tableau 3 : Comparaison des licences Creative Commons, Open Data Commons et de logiciels**

| **Licence*** | **Distribution** | **Modification** | **Octroi de sous-licence€** | **Attribution** |
|---|---|---|---|---|
| © Tous droits réservés | Permission requise | Permission requise | Permission requise | Exigée |
| CC BY | Permise | Permise | Permis | Exigée |
| CC0 | Permise | Permise | Interdit | Non exigée |
| ODbL | Permise | Permise | Interdit | Exigée |
| ODC-BY | Permise | Permise | Interdit | Exigée |
| PDDL | Permise | Permise | Permis | Non exigée |
| MIT | Permise | Permise | Permis | Exigée |
| GNU GPL | Permise | Permise | Permis | Exigée |
| Apache | Permise | Permise | Permis | Exigée |

Tableau comparatif sous licence CC BY-SA 4.0 (https://creativecommons.org/licenses/by-sa/4.0/deed.fr), fondé sur l'article *Comparison of Free and Open-Source Software licenses (https://en.wikipedia.org/wiki/Comparison_of_free_and_open-source_software_licences)*, Wikipedia, CC BY-SA 3.0 (https://creativecommons.org/licenses/by-sa/3.0/deed.fr).

* Les huit licences permettent une utilisation commerciale.

€ L'octroi d'une sous-licence indique que les dérivés peuvent être partagés sous une licence différente.

# Conclusion

Ce chapitre porte sur la planification de la gestion des données en tant que pratique exemplaire en GDR qui peut appuyer les données ouvertes et le partage des données en tant que partie intégrante d'un processus de travail en science ouverte dans les sciences sociales et d'autres contextes disciplinaires connexes. Les chercheuses et chercheurs choisissent de rendre leurs données ouvertement accessibles pour diverses raisons, notamment pour que leurs travaux soient davantage cités; toutefois, le mouvement de la science ouverte a pour objectif de rendre la recherche plus facilement reproductible et transparente, d'économiser temps et

argent et de réunir de manière novatrice des données jusque-là isolées. Par le biais du PGD dans l'étude de cas, *Meaningful Data Counts*, vous avez découvert la valeur d'un PGD dans le cadre global de planification de projets, sous l'angle des objectifs de la science ouverte. Le PGD permet une gestion conséquente et éthique de tous les jeux de données produits par plusieurs membres de l'équipe de recherche par le biais d'analyses bibliométriques, de sondages et d'entrevues. En outre, il fait en sorte que les données soient partagées tout au long du cycle de vie du projet, et non seulement au moment de publier les résultats de recherche. Les éléments clés du partage de données définis dans le PGD comprennent le dépôt de jeux de données dans un dépôt reconnu à l'aide d'une licence ouverte. L'octroi de licences ouvertes par MDC permet à d'autres chercheuses et chercheurs de réutiliser leurs travaux; grâce au dépôt de données, les chercheuses et chercheurs peuvent trouver les jeux de données et les citer adéquatement. La dernière section de ce chapitre aborde les éléments à prendre en considération en ce qui a trait à la confidentialité : avant de rendre les données ouvertes, vous devez vérifier si elles sont protégées par le droit d'auteur; le cas échéant, vous devez découvrir qui le détient. Lorsque vous avez confirmé qu'il est possible de partager les données ouvertement, le choix d'une licence ouverte qui permet les modifications favorise la réutilisation des données à des fins savantes et scientifiques. Toutes les données ne peuvent pas devenir des données ouvertes; cependant, si vous souhaitez adopter les principes du mouvement de science ouverte par le biais du partage de données et de versement dans des dépôts de données, un PGD vous aidera à normaliser et à communiquer les étapes à suivre aux membres de l'équipe et à la communauté disciplinaire élargie.

## Questions de réflexion

*Un élément interactif H5P a été exclu de cette version du texte. Vous pouvez le consulter en ligne ici : https://ecampusontario.pressbooks.pub/gdrcanada/?p=54#h5p-8 (https://ecampusontario.pressbooks.pub/gdrcanada/?p=54#h5p-8)*

## Éléments clés à retenir

- La science ouverte est un mouvement visant à rendre la recherche scientifique, les données et les publications accessibles par le biais du libre accès. Elle soutient l'ouverture des données et leur réutilisation avec des outils ouverts, la participation à la science citoyenne et l'accès à des méthodes ouvertes pour évaluer la recherche.
- Les motivations des chercheuses et chercheurs pour partager les données et les citer reposent souvent sur des normes disciplinaires; toutefois, les chercheuses et chercheurs qui publient et citent des données participent à un processus de valorisation des données en tant que résultat de recherche de premier ordre au statut équivalent à celui d'autres résultats de recherche.
- La création d'un plan de gestion des données (PGD) avec un processus de travail en science ouverte est un excellent moyen de satisfaire aux exigences des organismes de financement en ce qui a trait à la gestion efficace des données de recherche d'un projet avec pour objectif de permettre un partage de données éthique.
- Lorsque vous liez les processus de travail documentés de votre PGD à d'autres éléments de votre projet de recherche, vous faites en sorte que votre recherche sera partagée à grande échelle à différentes étapes du projet et que les données sous-jacentes aux résultats de la recherche rapportées dans une publication sont transparentes et réplicables d'un bout à l'autre du projet (et non seulement à sa conclusion).
- Les PGD sont des documents vivants. Il peut s'avérer utile de les revoir et de les mettre à jour tout au long du projet. La création de versions subséquentes peut aider à répertorier l'évolution de votre processus de recherche.
- En plus de considérations éthiques, avant de rendre les données ouvertes, il faut préciser l'existence et la propriété des droits d'auteur; le cas échéant, obtenez la permission de verser les données dans un dépôt ouvert.
- Après avoir confirmé qu'il est possible de partager les données ouvertement, choisissez une licence ouverte qui permet d'apporter des modifications autant que possible : une condition « Pas de modification » limite grandement son utilisation à des fins savantes et scientifiques ainsi que les avantages de rendre les données ouvertes.

# Bibliographie

Brunet, M., Hatherill J. et Ripp, C. (2021). *Libre accès aux connaissances Partie 2 : Partagez vos données de recherche.* Bibliothèque de l'Université d'Ottawa. http://hdl.handle.net/10393/43308 (http://hdl.handle.net/10393/43308)

Brunet, M. et Rouleau, T. (2021). *Droit d'auteur et données de recherche à l'Université d'Ottawa : Questions fréquemment posées*, Bibliothèque de l'Université d'Ottawa. https://www.uottawa.ca/library/sites/g/files/bhrskd381/files/2022-12/droit_dauteur_et_donnees_de_recherche_faq.pdf (https://www.uottawa.ca/library/sites/g/files/bhrskd381/files/2022-12/droit_dauteur_et_donnees_de_recherche_faq.pdf)

Cobey, K. D., Haustein, S., Brehaut, J. Dirnagl, U., Franzen, D. L., Hemkens, L. G., Presseau, J., Riedel, N., Strech, D., Alperin J. P., Costas, R., Sena, E. S., van Leeuwen, T., Ardern, C. L., Bacellar I. O. L, Camack, N., Correa, M. B., Buccione, R., Cenci, M. S., ... Moher, D. (2022). *Establishing a core set of open science practices in biomedicine: A modified Delphi study*. medRxiv. https://doi.org/10.1101/2022.06.27.22276964 (https://doi.org/10.1101/2022.06.27.22276964)

Fecher, B. et Friesike, S. (2014). Open science: One term, five schools of thought. Dans S. Bartling et S. Friesike (dir..), *Opening Science: The evolving guide on how the Internet is changing research, collaboration and scholarly publishing* (p. 17–47). Springer. https://doi.org/10.1007/978-3-319-00026-8_2 (https://doi.org/10.1007/978-3-319-00026-8_2)

FOSTER. (s.d.-a). *Open Data*. https://www.fosteropenscience.eu/taxonomy/term/6 (https://www.fosteropenscience.eu/taxonomy/term/6)

FOSTER. (s.d.-b). *Open Science*. https://www.fosteropenscience.eu/taxonomy/term/7 (https://www.fosteropenscience.eu/taxonomy/term/7)

franzke, a. s., Bechmann, A., Zimmer, M., Ess, C. et Association of Internet Researchers (2020). *Internet Research: Ethical Guidelines 3.0*. https://aoir.org/reports/ethics3.pdf (https://aoir.org/reports/ethics3.pdf)

Gregory, K., Ninkov, A. B., Ripp, C., Roblin, E. Peters, I. et Haustein, S. (2023). *Tracing data: A survey investigating disciplinary differences in data citation*. Zenodo. https://doi.org/10.5281/zenodo.7555266 (https://doi.org/10.5281/zenodo.7555266)

ICPSR. (s.d.) *Recommended Informed Consent Language for Data Sharing*. https://www.icpsr.umich.edu/web/pages/datamanagement/confidentiality/conf-language.html (https://www.icpsr.umich.edu/web/pages/datamanagement/confidentiality/conf-language.html)

Ninkov, A., Gregory, K., Ripp. C., Morissette, E., Harper, L., Peters, I., Tayler, F. et Haustein, S. (2022).

*Research data management plan for the meaningful data counts project (v.2)*. Zenodo. https://doi.org/10.5281/zenodo.6473351 (https://doi.org/10.5281/zenodo.6473351)

Office de la propriété intellectuelle du Canada (OPIC). (2019). *Le guide du droit d'auteur.* Gouvernement du Canada. https://ised-isde.canada.ca/site/office-propriete-intellectuelle-canada/fr/guide-droit-dauteur (https://ised-isde.canada.ca/site/office-propriete-intellectuelle-canada/fr/guide-droit-dauteur)

Tenopir C., Rice, N.M., Allard, S., Baird, L., Borycz, J., Christian, L., Grant, B., Olendorf, R. et Sandusky, R.J. (2020). Data sharing, management, use, and reuse: Practices and perceptions of scientists worldwide. *PLOS ONE, 15*(3): e0229003. https://doi.org/10.1371/journal.pone.0229003 (https://doi.org/10.1371/journal.pone.0229003)

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., ... Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data, 3*(1), 160018. https://doi.org/10.1038/sdata.2016.18 (https://doi.org/10.1038/sdata.2016.18)

## À propos des auteurs

### Felicity Tayler

Felicity Tayler, MLIS, PhD, est bibliothécaire chargée de la gestion des données de recherche à l'Université d'Ottawa et associée de recherche au Labo de données en sciences humaines. Elle est cocandidate du partenariat SpokenWeb (https://spokenweb.ca/), financé par le CRSH, qui met l'accent sur une approche coordonnée et collaborative de l'étude historique littéraire et du développement numérique avec diverses collections d'enregistrements oraux provenant de l'ensemble du Canada et d'ailleurs. En tant que membre du Groupe d'experts national sur la formation de l'Alliance de recherche numérique du Canada, Tayler a été l'autrice principale de la REL bilingue *Data Primer: Making Digital Humanities Research Data Public (https://ecampusontario.pressbooks.pub/dataprimer/)* / *Manuel d'introduction aux données : rendre publiques les données de recherche en sciences humaines numériques (https://ecampusontario.pressbooks.pub/introdonnees/)*. Également artiste visuel et commissaire d'exposition, Tayler a réalisé des expositions et publié des ouvrages scientifiques explorant les relations de coédition dans les communautés littéraires et artistiques. ORCID : 0000-0001-8865-2836 (https://orcid.org/0000-0001-8865-2836)

### Mélanie Brunet

Mélanie Brunet est bibliothécaire à l'Université d'Ottawa; elle a commencé au service du droit d'auteur et

travaille maintenant dans l'éducation ouverte. Elle fait de la sensibilisation sur l'abordabilité des manuels scolaires et les ressources éducatives libres à l'Université d'Ottawa et elle est membre du groupe de travail sur l'éducation ouverte de l'ABRC depuis 2019, dirigeant son groupe de travail francophone sur l'éducation ouverte. Elle a coédité le *Guide REL par discipline : Université d'Ottawa (https://ecampusontario.pressbooks.pub/uottawareldisciplineversion2/)*, un outil pour aider le corps professoral et la communauté étudiante à se familiariser avec les REL dans leurs disciplines. Mélanie est titulaire d'une maîtrise en information et d'un doctorat en histoire de l'Université de Toronto.

## Kathleen Gregory

Kathleen Gregory est chercheuse postdoctorale à l'Université de Vienne et au Scholarly Communications Lab au Canada. Elle est également chercheuse affiliée au Center for Science and Technology Studies (CWTS) de l'Université de Leiden. Les recherches de Mme Gregory portent sur les pratiques en matière de données dans la communication savante et scientifique, en particulier les pratiques de gestion et de curation des données et l'examen de ce que ces pratiques permettent.

## Lina Harper

Lina Harper est bibliothécaire responsable de la curation des données et travaille avec le Dépôt fédéré de données de recherche et la communauté Borealis à l'Alliance de recherche numérique du Canada (l'Alliance). Elle est titulaire d'un baccalauréat en études féminines et en communications (Université Concordia) et a rédigé un mémoire de maîtrise en sciences de l'information (Université d'Ottawa). Également coprésidente du groupe de travail sur les activités de curation de l'Alliance, Lina s'intéresse aux sciences de l'information, à la conception de l'information, à l'engagement communautaire, aux humanités numériques et à l'équité. lina.harper@alliancecan.ca (mailto:lina.harper@alliancecan.ca) | ORCID : 0000-0002-8735-7621 (https://orcid.org/0000-0002-8735-7621)

## Stefanie Haustein

Stefanie Haustein est professeure associée à l'École des sciences de l'information de l'Université d'Ottawa et codirectrice du Scholarly Communications Lab (https://www.scholcommlab.ca/) (ScholCommLab), un groupe interdisciplinaire de chercheuses et chercheurs basés à Ottawa et à Vancouver qui analysent l'érudition à l'ère numérique. Elle est également chercheuse affiliée à l'Institut de recherche sur la science, la société et la politique publique (https://www.uottawa.ca/recherche-innovation/issp), au Centre for Journalology (https://ohri.ca/journalology/), et à l'Institut de recherche LIFE (https://www.uottawa.ca/research-innovation/life) de l'Université d'Ottawa, ainsi qu'au Centre interuniversitaire de recherche sur la science et la technologie (https://www.cirst.uqam.ca/) de l'Université du Québec à Montréal. Les recherches de Mme Haustein

portent sur la communication savante, l'évaluation de la recherche et la science ouverte, y compris le libre accès, le partage et la réutilisation des données de recherche.

PARTIE IV

# TYPES DE DONNÉES DE RECHERCHE

13.

# LES DONNÉES SENSIBLES: DES CONSIDÉRATIONS PRATIQUES ET THÉORIQUES

Dr. Alisa Beth Rod et Kristi Thompson

**Objectifs d'apprentissage**

À la fin de ce chapitre, vous pourrez :

1. Définir les termes suivants: dépersonnalisation, renseignements identificatoires, données sensibles, évaluation statistique des risques de divulgation.
2. Reconnaître que l'établissement de niveaux de risque pour des données sensibles (p. ex., faible, moyen, élevé, très élevé) dépend du contexte de recherche.
3. Comprendre les politiques et les règlements canadiens en matière d'éthique relatifs aux données de recherche.

**Évaluation préliminaire**

*Un élément interactif H5P a été exclu de cette version du texte. Vous pouvez le consulter en ligne ici : https://ecampusontario.pressbooks.pub/gdrcanada/?p=58#h5p-4 (https://ecampusontario.pressbooks.pub/gdrcanada/?p=58#h5p-4)*

# Introduction

Que sont les **données sensibles**? La *Boîte à outils pour les données sensibles – destinée aux chercheurs (https://zenodo.org/record/4088986#.Y9wq-BOZPap)* (Groupe d'experts sur les données sensibles du réseau Portage, 2020a) définit les données sensibles comme des « informations qui doivent être protégées contre l'accès non autorisé ou la divulgation » et donne plusieurs exemples. Toutefois, cette définition des données sensibles soulève la question suivante : pourquoi ces informations doivent-elles être protégées? L'examen des exemples peut nous aider à y voir plus clair, car ils incluent des éléments tels que des renseignements personnels sur la santé et d'autres types de renseignements d'ordre confidentiel, certains renseignements géographiques (p. ex., des localisations d'espèces en péril) ou des données protégées en vertu de politiques institutionnelles. La caractéristique commune à tous ces exemples est la notion du risque, celui associé aux personnes dont la confidentialité est compromise, aux espèces en péril qui seraient perturbées ou chassées, aux politiques qui seraient enfreintes. Autrement dit, les données sensibles s'appliquent aux données qui ne peuvent être partagées sans risquer de trahir la confiance ou de nuire à une personne, une entité ou une communauté.

Dans ce chapitre, nous discuterons du travail avec les données sensibles dans le cadre des politiques fédérales et provinciales canadiennes. (Les données autochtones ont des implications en matière d'éthique et de propriété et elles sont abordées dans un autre chapitre.) Nous terminerons en donnant un aperçu des options pour la préservation sécuritaire, le partage et l'archivage approprié des données sensibles.

# Les données de la recherche avec des êtres humains

Au Canada, tant au fédéral, au provincial que dans les institutions, une variété de cadres juridiques, stratégiques et réglementaires régissent les données sensibles qui impliquent des êtres humains. Dans la plupart des cas, ces exigences réglementaires ont été conçues pour assurer la confidentialité et un degré élevé de protection de la vie privée. Ainsi, le cadre réglementaire associé aux données sensibles relève de la catégorie des données impliquant des êtres humains.

## Les politiques sur la vie privée au Canada

Ce n'est pas toujours facile de savoir laquelle des lois sur la protection de la vie privée s'applique à chacune des situations. Les principaux règlements sur la protection de la vie privée en lien avec les **données de recherche** se retrouvent généralement au niveau des juridictions provinciales et territoriales puisque les universités ne relèvent pas du champ d'application des deux principales lois fédérales sur la protection de la vie privée (Commissariat à la protection de la vie privée au Canada, s.d.). Toutefois, certaines informations sensibles, comme les dossiers médicaux, peuvent être recueillies par des chercheuses et chercheurs universitaires en partenariat avec des organismes privés et publics; celles-ci relèvent alors de la juridiction fédérale de la Loi sur la protection des renseignements personnels, qui s'applique aux organismes gouvernementaux, ou de la Loi sur la protection des renseignements personnels et les documents électroniques (LPRPDE), qui s'applique aux entités commerciales du secteur privé. Le gouvernement canadien a développé un outil pratique (https://web.archive.org/web/20220102063509/https:/www.priv.gc.ca/fr/signaler-un-probleme/leg_info_201405/) pour aider à déterminer laquelle des législations s'applique aux différents scénarios impliquant des informations sensibles.

Au niveau national, l'énoncé de politique des **trois organismes subventionnaires** sur la conduite éthique de la recherche avec des êtres humains (*Énoncé de politique des trois conseils : Éthique de la recherche avec des êtres humains (https://ethics.gc.ca/fra/policy-politique_tcps2-eptc2_2022.html)* – **EPTC 2**) établit les paramètres liés à la vie privée, la justice, le respect et la préoccupation pour le bien-être des participantes et des participants. L'EPTC 2 supervise aussi la gouvernance des comités d'éthique de la recherche (CÉR), qui sont responsables d'évaluer les projets de recherche qui dépendent d'êtres humains. Contrairement aux États-Unis, où une loi fédérale régit les renseignements médicaux (HIPPA (https://www.hhs.gov/hipaa/index.html)), la gestion des dossiers médicaux ou des données cliniques au Canada est régie au niveau provincial et territorial (https://www.priv.gc.ca/fr/a-propos-du-commissariat/ce-que-nous-faisons/collaboration-avec-les-provinces-et-les-territoires/lois-et-organismes-de-surveillance-provinciaux-et-territoriaux-en-matiere-de-protection-de-la-vie-privee/).

Toutes les provinces et tous les territoires disposent d'au moins une loi sur la protection de la vie privée qui s'applique à la recherche.

Traditionnellement, les systèmes judiciaires au Canada et ailleurs en Occident garantissent le droit des individus (et par extension, des sociétés) à la vie privée, à la propriété de leurs renseignements et à la protection contre les préjudices directs en contexte de recherche. Toutefois, les données ont le potentiel de nuire à certains groupes ou certaines communautés – par exemple, pour stigmatiser des groupes racisés ou des minorités sexuelles et de genre (voir Ross *et al*., 2018.) De tels préjudices sont trop peu abordés dans les politiques et législations canadiennes actuelles. Les principes de propriété, contrôle, accès et possession (https://fnigc.ca/fr/les-principes-de-pcap-des-premieres-nations/) (**PCAP®**) constituent un modèle alternatif. Il s'agit d'un protocole de recherche établi pour protéger les intérêts des Premières Nations piloté par le Centre de gouvernance de l'information des Premières Nations (https://fnigc.ca/fr/). Ces principes « s'appliquent aux Premières Nations spécifiquement, et non à tous les Autochtones » (Centre de gouvernance de l'information des Premières Nations, s.d.) et leur utilisation n'est pas destinée à d'autres contextes. Même si ces principes cherchent d'abord à privilégier les intérêts d'une communauté, ils peuvent aussi s'appliquer de façon plus générale à des recherches auprès de communautés marginalisées. Pour en savoir plus sur les modèles autochtones de recherche éthique, consultez le chapitre « Souveraineté des données autochtones. »

De nombreuses provinces mettent à jour leurs lois sur la protection de la vie privée, ce qui aura un impact sur la gestion des données de recherche impliquant des êtres humains. Par exemple, la loi 25 (https://www.quebec.ca/gouvernement/ministeres-et-organismes/institutions-democratique-acces-information-laicite/acces-documents-protection-renseignements-personnels/pl64-modernisation-de-la-protection-des-renseignements-personnels) (aussi appelée Loi modernisant des dispositions législatives en matière de protection des renseignements personnels) a été adoptée au Québec pour solidifier les exigences en matière de consentement, de surveillance et de conformité. La loi 25 s'inspire du Règlement général sur la protection des données (RGPD) (https://www.cnil.fr/fr/reglement-europeen-protection-donnees) de l'Union européenne, largement considéré comme étant la plus complète des législations sur la protection de la vie privée au monde. Les impacts potentiels de la loi 25 comprennent l'obligation du consentement pour chacune des utilisations secondaires des données de recherche, l'introduction du droit à l'effacement ou du « **droit à l'oubli** » (Wolford, 2018) et l'obligation de faire une évaluation formelle des impacts sur la vie privée avant tout transfert de données d'une personne à l'extérieur du Québec (Commissariat à la protection de la vie privée du Canada, 2020). Nous aborderons le consentement plus en détail dans la section intitulée « Le langage du consentement et l'EPTC 2 ». Cette modification légale s'explique par le fait que les chercheuses et chercheurs pouvaient demander un consentement général pour l'utilisation des renseignements des personnes participantes (p. ex., l'échantillon d'un patient participant à un essai clinique pouvait être utilisé pour d'autres études sans avoir à fournir de détails sur ces études particulières). Maintenant, la loi 25 ne permet plus de consentement général pour les études. Elle exige que les responsables des recherches obtiennent le consentement à chaque fois que l'échantillon est utilisé à de nouvelles fins. Bien que la loi 25 soit propre au

Québec, elle représente un modèle en matière de réforme de loi sur la protection de la vie privée et pourrait donc avoir d'importantes conséquences pour les données de recherche impliquant des êtres humains.

## Risques et préjudices

La  divulgation de renseignements personnels est un risque qui peut survenir lorsqu'il est possible d'isoler des individus dans un jeu de données et d'associer leurs renseignements à des sources externes, ce qui, avec un effort raisonnable, permet d'identifier les personnes. L'ampleur des préjudices causés à des participantes et participants de recherche dépend de la population et du sujet des données. Généralement, les risques plus élevés sont faciles à identifier et à définir (p. ex., si des renseignements sur la santé d'un individu étaient rendus publics). Les enfants sont considérés comme une population vulnérable puisqu'ils ne peuvent donner eux-mêmes leur consentement. Les recherches impliquant des enfants représentent donc un niveau élevé de risque puisque la divulgation de leurs renseignements peut causer de graves préjudices. Des sujets jugés tabous par la société peuvent aussi comporter des risques de préjudices importants pour les gens qui participent à des recherches. Ce qui constitue un sujet tabou peut varier d'une culture ou d'une situation à l'autre. Les éléments suivants peuvent néanmoins être considérés comme extrêmement sensibles, ce qui augmente le risque de préjudices potentiel qu'une participante ou un participant de recherche pourrait subir si ses renseignements sont divulgués :

- usage de drogues ou d'alcool (y compris la cigarette);
- pratiques sexuelles / MTS;
- questions familiales privées;
- violence conjugale / familiale;
- perte ou décès dans la famille;
- statut de victime;
- comportements criminels / délinquants;
- questions et problèmes de santé physique / mentale.

Les populations vulnérables, telles que les communautés suivantes, sont plus susceptibles d'être affectées par une divulgation de renseignements, et ce indépendamment du sujet de recherche :

- les communautés autochtones;
- les communautés racisées;
- les groupes à faibles revenus;
- les enfants/adolescents;
- les opprimés politiques.

Les recherches impliquant des êtres humains issus de populations vulnérables et/ou abordant des sujets sensibles peuvent nécessiter l'ajout de mesures de protection supplémentaires en matière de stockage et de sécurité des données. Certaines données peuvent être partagées sans se préoccuper du risque de divulgation, comme dans le cas de personnes qui participent à une recherche et consentent au partage de leurs renseignements identificatoires (p. ex., des partages d'histoires orales lors d'entrevues) ou dans le cas de renseignements recueillis de sources publiques sans attentes particulières liées à la vie privée (p. ex., une liste des membres d'un conseil d'administration). Autrement, les données doivent être évaluées pour déterminer leur risque de divulgation et elles ne pourront être partagées que si elles tombent en deçà d'un seuil de risque acceptable. Des mesures courantes de curation et de statistiques permettent d'évaluer quantitativement les risques et ainsi de réduire le potentiel de divulgation de données confidentielles. Les premières étapes comprennent l'analyse du caractère unique des données de chaque personne à l'intérieur d'un jeu de données plus large.

## Les identificateurs

Plusieurs personnes peuvent consentir à ce que leurs renseignements soient utilisés à des fins de recherche sans vouloir que leur identité soit divulguée. Les données de recherche peuvent comporter des **identifiants directs**, tels que les coordonnées des personnes participantes, leur numéro d'étudiant ou autre type de **renseignements identificatoires**. Les données de recherche sans identifiants directs sont quand même susceptibles de porter atteinte à la confidentialité en raison des **identifiants indirects** ou ***quasi-identifiants*** – des détails personnels qui, utilisés en combinaison avec d'autres, peuvent mener à l'identification d'une personne. Ces données peuvent inclure des sondages ou des entrevues avec des personnes participantes qui ont consenti à fournir des renseignements à des fins de recherche. Les données des personnes participantes peuvent également comprendre des renseignements issus de dossiers médicaux, de déclarations d'impôts, des médias sociaux ou de toute autre source de renseignements de nature personnelle.

### Les identifiants directs

*Les identifiants directs* mettent immédiatement les participantes et participants d'une recherche à risque d'être réidentifiés. On parle d'éléments tels que le nom, le numéro de téléphone, mais aussi d'autres détails moins évidents. Par exemple, la loi américaine *Health Insurance Portability and Accountability Act* (HIPAA) (https://www.cdc.gov/phlp/publications/topic/hipaa.html) considère toute zone géographique de moins de 20 000 personnes comme un identifiant direct. Des dates précises liées à des événements personnels, telles que les dates de naissance, sont également considérées comme identificatoires.

Le HIPAA a établi une liste de 18 identifiants personnels (https://www.hipaajournal.com/considered-phi-hipaa/) tandis qu'un ensemble de lignes directrices du *British Medical Journal (https://www.bmj.com/content/340/*

*bmj.c181)* contient une liste de 14 identifiants directs et 14 indirects basés sur des lignes directrices internationales. À partir de ces sources et d'autres, nous avons compilé une liste d'identifiants directs pour la recherche canadienne. Ils devraient toujours être éliminés des données avant de rendre celles-ci publiques, à moins que les participantes et les participants de recherche n'aient autorisé le partage de leur identité (à quelques exceptions près).

1. nom entier ou partiel ou initiales
2. dates précises d'événements personnels tels que la naissance, la graduation, l'admission à l'hôpital (seul le mois ou l'année peut être acceptable)
3. adresse complète ou partielle (de grandes zones géographiques, telles que des villes, appartiennent à la catégorie d'identifiants indirects et doivent être révisées)
4. code postal complet ou partiel (les trois premiers chiffres peuvent être acceptables)
5. numéros de téléphone ou de télécopieur
6. adresse courriel
7. identifiants ou noms d'utilisateurs Web ou de médias sociaux tels que le pseudonyme Twitter
8. numéros de protocole Internet ou IP; renseignements précis relatifs au navigateur Web ou au système d'exploitation (ces informations peuvent être recueillies par certains types d'outils de sondage ou de formulaires Web)
9. identifiants de véhicule tels que la plaque d'immatriculation
10. identifiants liés à des dispositifs médicaux ou autres
11. tout autre numéro d'identification unique lié directement ou indirectement à un individu tel que le numéro d'assurance sociale, numéro d'étudiant ou numéro d'identification d'un animal de compagnie
12. photos d'individus ou de leur domicile ou emplacement; des enregistrements vidéos les montrant; des images médicales
13. enregistrements audio de personnes (Han *et al*., 2020)
14. données biométriques
15. tout attribut personnel unique ou reconnaissable (p. ex., maire de Kapuskasing ou gagnant du prix Nobel)

De plus, tous les fichiers numériques partagés, tels que des photos ou documents, doivent être vérifiés au cas où des renseignements comme un nom ou un emplacement y seraient intégrés (voir Henne *et al*., 2014.)

## Les identifiants indirects

Les risques de violation de la confidentialité posés par les identifiants directs sont évidents – si vous détenez l'adresse ou le numéro d'assurance sociale d'une personne, sa vie privée peut être compromise. Mais quels sont les identifiants indirects et pourquoi sont-ils problématiques? Les identifiants indirects (aussi désignés comme

quasi-identifiants) se rapportent à des attributs qui ne peuvent pas en soi identifier une personne, mais qui, combinés, peuvent révéler l'identité de quelqu'un. Une variable a le potentiel d'être un identifiant (direct ou indirect) seulement si elle peut être liée à des renseignements d'autres sources pour identifier une personne.

Il est impossible de dresser une liste complète des quasi-identifiants, mais on devrait toujours tenir compte des éléments suivants:

- âge (peut être un identifiant direct dans le cas de personnes très âgées);
- identité de genre;
- revenu;
- emploi ou secteur d'activité;
- variables géographiques;
- variables ethniques ou d'immigration;
- appartenances à des organismes ou utilisation de services particuliers.

Ces variables doivent être considérées en relation avec tout renseignement contextuel d'un jeu de données – par exemple, la documentation d'un sondage ou d'une recherche publiée peut indiquer de façon claire que les personnes qui ont participé à la recherche habitent un endroit particulier ou pratiquent un métier particulier.

Les autres variables d'un jeu de données constituent des renseignements non identificatoires (peu susceptibles d'être reconnus comme provenant d'individus spécifiques ou qui n'apparaissent pas dans des bases de données externes). Celles-ci peuvent comprendre des opinions, des notes sur l'**échelle de Likert**, des mesures temporaires (telles que le rythme cardiaque au repos) et autres. Ces renseignements ne font pas partie de l'évaluation de la confidentialité, mais doivent quand même être pris en compte dans l'évaluation globale des risques. Une des questions en lien avec les variables non identificatoires se rapporte au niveau de sensibilité des données; un jeu de données qui contient des renseignements confidentiels sur la santé ou un sondage qui pose des questions de nature délicate sur des comportements antérieurs doit être traité avec plus de soins qu'un jeu de données qui évalue des produits.

Un ensemble d'enregistrements qui comporte les mêmes valeurs de quasi-identifiants s'appelle une classe d'équivalence. Une **classe d'équivalence** de 1 se rapporte à une personne dont les attributs sont uniques dans le jeu de données. Cette personne risque donc d'être identifiée et elle est désignée comme étant **unique à l'échantillon**. Si une étude contient un échantillon complet d'une population quelconque (p. ex., tous les employés d'un endroit particulier), cette personne est aussi **unique à la population** en fonction de ces attributs (son identité peut donc être évidente pour les gens qui connaissent la population). Parallèlement, les membres d'une classe d'équivalence plus importante – avec 10, 20 ou 50 membres – sont indissociables les uns des autres et ne peuvent donc pas être identifiés en fonction de leurs quasi-identifiants; ils ne sont donc pas considérés comme étant à risque d'être identifiés.

Vous connaissez maintenant les quasi-identifiants et comprenez de quelle façon ils peuvent être utilisés pour identifier les personnes qui participent à des sondages. Alors, que faites-vous avec? Vous pouvez tout simplement les retirer de vos données comme vous le feriez pour les identifiants directs. Par contre, ce procédé pourrait avoir de sérieux impacts sur la capacité des autres membres de la communauté de recherche à utiliser ce jeu de données. Vous devriez plutôt évaluer les quasi-identifiants pour en déterminer le niveau de risque.

Comme première étape, une curatrice ou un curateur des données peut isoler les variables et les étudier dans le contexte d'autres informations sur les données. Les variables quasi-identifiantes qui contiennent des groupes avec un petit nombre de personnes répondantes (p. ex., une variable sur la religion avec trois réponses de « Bouddhisme ») peuvent représenter un risque élevé. Des valeurs peu courantes (p. ex., plus de six enfants) peuvent aussi comporter un risque élevé. Ces valeurs peuvent être évaluées en effectuant un calcul de fréquence sur les données. Toutefois, la grosseur des groupes identifiables tant dans l'enquête que dans la population en général doit être prise en compte. Il n'y a peut-être qu'une seule personne de Winnipeg dans votre sondage téléphonique à composition aléatoire, mais si l'enquête ne cherche pas à cibler davantage, la personne ne risque pas d'être identifiée.

Une approche pratique et sensée à la **dépersonnalisation** des données est de décrire une personne en n'utilisant que les valeurs des variables démographiques dans un jeu de données :

> « Je pense à une personne mariée de sexe féminin, vivant en Ontario, détentrice d'un diplôme universitaire qui est âgée de 40 à 55 ans. »

Cette personne ne semble pas être à risque à moins que l'information contextuelle fournisse des indices supplémentaires – par exemple, s'il s'agit d'une enquête sur les gens qui exercent la profession d'arbitre de hockey.

Des valeurs à combinaison inhabituelle ou atypique pour les variables peuvent être problématiques. L'âge, le niveau d'éducation ou l'état civil ne sont pas forcément identifiants, mais que se passe-t-il si, par exemple, une personne dans le jeu de données est dans le groupe d'âge de moins de 17 ans et qu'elle indique être divorcée ou diplômée universitaire? Cette personne peut alors être identifiée et représente un exemple de valeur extrême cachée qui n'apparaît pas en examinant la fréquence de toutes les variables d'un jeu de données. Plus il y a d'identifiants indirects dans les données, plus il y a de chances de combinaisons atypiques cachées et plus elles sont difficiles à déceler. Le besoin d'une méthode formelle d'évaluation des quasi-identifiants et de quantification de leur niveau de risque s'impose. Ce processus s'appelle l'évaluation statistique des risques de divulgation.

## L'évaluation statistique des risques de divulgation

Il existe différentes techniques pour évaluer et limiter les risques de réidentification, mais la plus connue

d'entre elles est la ***k*-anonymisation**. Il s'agit d'une approche permettant de démontrer mathématiquement qu'un jeu de données a été anonymisé. Elle a d'abord été avancée en 1998 par des informaticiens (Samarati et Sweeney) et constitue depuis lors la base de tous les efforts formels d'anonymisation des données. L'approche part du principe que ce ne devrait pas être possible d'isoler moins de « k » cas individuels dans votre jeu de données et ce, pour toutes les combinaisons possibles de variables identificatoires – « k » correspond au numéro établi par la chercheuse ou le chercheur; dans la pratique, il correspond généralement à cinq.

Imaginons un sondage sur le personnel d'une usine d'outils et de matrices avec trois variables démographiques : le groupe d'âge, le genre et le groupe ethnique. Prenons, par exemple, une personne dans le jeu de données qui n'est pas une minorité visible et qui est de sexe masculin entre 25 et 30 ans. Pour que les données soient *k*-anonymes avec k=5, au moins quatre autres personnes dans le jeu de données doivent avoir le même ensemble d'attributs. Cela doit également être vrai pour tous les autres individus dans le jeu de données; chaque personne doit avoir au moins quatre **jumeaux de données**.

Dans la figure 1, les cas 1, 6 et 13 représentent une classe d'équivalence où k=3. Chaque cas de cette classe d'équivalence a deux jumeaux de données. Même si un pirate informatique savait qu'une personne ciblée se trouvait dans le jeu de données et qu'il pouvait faire correspondre ses attributs aux données, il ne pourrait identifier cette personne parmi les trois cas. Le cas 14 quant à lui, n'a aucun jumeau de données – il est unique à l'échantillon.

| ID | Genre | GrpAge | GrpEthn |
|---|---|---|---|
| 1 | M | 25-34 | 1 |
| 2 | F | 16-24 | 1 |
| 3 | M | 25-34 | 2 |
| 4 | M | 16-24 | 1 |
| 5 | F | 35-44 | 1 |
| 6 | M | 25-34 | 1 |
| 7 | F | 16-24 | 1 |
| 8 | F | 35-44 | 1 |
| 9 | F | 35-44 | 2 |
| 10 | M | 25-34 | 2 |
| 11 | M | 16-24 | 1 |
| 12 | F | 25-34 | 1 |
| 13 | M | 25-34 | 1 |
| 14 | F | 16-24 | 2 |
| 15 | F | 35-44 | 1 |

Figure 1. Classes d'équivalence. Vert: classe où k=3. Orange: classe où k=1

Pour réussir la *k*-anonymisation où la valeur de « k » d'un jeu de données correspond à cinq, vous pouvez utiliser des techniques de réduction des données, dont la réduction globale des données et la **suppression locale**. La **réduction globale des données** implique de modifier certaines variables de tout un jeu de données, tel que de regrouper certaines réponses en catégories (p. ex., regrouper l'âge en tranches de 10 ans). La suppression locale se rapporte à l'élimination de réponses ou de cas individuels (p. ex., supprimer la réponse sur l'état civil de la personne participante de moins de 17 ans plutôt que de regrouper les variables sur l'état civil et l'âge qui sont autrement inoffensives).

La *k*-anonymisation est facile à vérifier en utilisant des logiciels standards de statistiques, même si la plupart de ces progiciels n'ont pas de fonctionnalités intégrées pour faire ces vérifications. La ressource *Directives sur la dépersonnalisation des données (https://zenodo.org/record/4047176#.Y-EyXBOZPao)* de l'Alliance de recherche numérique du Canada (L'Alliance) (https://alliancecan.ca/fr/services/gestion-des-donnees-de-recherche/reseau-dexperts) fournit du code pour faire les vérifications dans R et Stata.

La *k*-anonymisation a pour but de garantir l'anonymisation des données, c'est-à-dire que chaque enregistrement des données anonymisées ne pourra être distingué de « k » moins un autre enregistrement d'un même jeu de données. Toutefois, les personnes qui participent à une recherche ne sont généralement pas

informées qu'il sera impossible de savoir quelle ligne d'un fichier de données correspond à leurs renseignements personnels. On leur dit plutôt que leurs réponses resteront confidentielles. Même si les enregistrements d'une personne ne sont pas uniques aux données, la possibilité de trouver quand même des renseignements personnels les concernant n'est pas exclue.

Quelques années après la publication sur la *k*-anonymisation comme solution à l'enjeu de la vie privée, des chercheurs ont remarqué une potentielle lacune importante : l'**attaque par homogénéité**. Lorsque tous les membres d'une classe d'équivalence (un jeu de jumeaux de données) partagent les mêmes valeurs d'attributs sensibles, des pirates informatiques pourraient déduire les attributs des personnes qui ont répondu à un sondage sans les identifier. Reprenons comme exemple l'échantillon de l'enquête sur le personnel d'une usine. La figure 2 illustre les variables démographiques accompagnées d'une question sensible, si les travailleuses et travailleurs sont pour ou contre la mise en place d'un syndicat. Les cas 1, 6 et 13 forment toujours une classe d'équivalence de k=3. Alors même si vous connaissez les personnes qui correspondent à ces attributs, vous ne pouvez savoir quelle personne correspond à quel cas. Par contre, ces trois personnes ont donné la même réponse à la question sur le syndicat. Vous connaissez donc la réponse de toutes ces personnes. Il y a donc une atteinte à la confidentialité.

| ID | Genre | GrpAge | GrpEthn | Syndicat |
|---|---|---|---|---|
| 1 | M | 25-34 | 1 | O |
| 2 | F | 16-24 | 1 | N |
| 3 | M | 25-34 | 2 | N |
| 4 | M | 16-24 | 1 | O |
| 5 | F | 35-44 | 1 | O |
| 6 | M | 25-34 | 1 | O |
| 7 | F | 16-24 | 1 | N |
| 8 | F | 35-44 | 1 | O |
| 9 | F | 35-44 | 2 | O |
| 10 | M | 25-34 | 2 | N |
| 11 | M | 16-24 | 1 | O |
| 12 | F | 25-34 | 1 | O |
| 13 | M | 25-34 | 1 | O |
| 14 | F | 16-24 | 2 | N |
| 15 | F | 35-44 | 1 | O |

Figure 2. Divulgation d'attributs. Vert : classe où k=3. Orange : classe où k=1

Des techniques qui affinent la *k*-anonymisation, telles que la ***k*-anonymisation p-sensible** et la ***l*-diversité** (Domingo-Ferrer et Torra, 2008) ont été élaborées pour aborder la question de la divulgation des attributs. Toutefois, leur mise en œuvre est difficile et tend à dégrader la valeur de recherche du jeu de données. Examinons une des variantes les plus simples.

La *l*-diversité est appliquée à un jeu de données quand chaque groupe d'enregistrements qui partage une même série d'attributs démographiques comporte au moins « *l* » valeurs différentes pour chacune des variables confidentielles. Dans notre exemple, chacun des groupes de jumeaux de données devrait comporter les deux réponses, oui et non, à la question sur le syndicat; deux représente donc la valeur maximale possible de « *l* » pour cette question. La valeur établie de « *l* » doit s'appliquer pour toute réponse confidentielle dans le jeu de données. Imaginez maintenant un sondage typique avec une douzaine de questions – la *l*-diversité doit être considérée pour chaque réponse dans chaque classe d'équivalence. Autrement dit, l'application des techniques comme la *l*-diversité n'est pratique que pour les jeux de données avec peu de variables.

La plus grande menace d'une attaque par homogénéité survient lorsqu'un jeu de données représente l'échantillon d'une population entière. Imaginez que le jeu de données du sondage à l'usine n'implique que 25% de la population. Cela signifie qu'il y a possiblement d'autres personnes qui ne font pas partie du jeu de données et qui possèdent les mêmes quasi-identifiants que les cas 1, 6, et 13; leur position sur l'établissement d'un syndicat est inconnue. En ignorant quelles personnes font partie du jeu de données, impossible de connaître leur opinion, et ce, même si elles appartiennent à la même classe d'équivalence. Cette supposition ne peut être faite que si le jeu de données représente un petit échantillon d'une population plus large et que si le principe de la *k*-anonymisation a été appliqué. Inversement, si le jeu de données représente l'échantillon d'une population entière ou d'une proportion importante de celle-ci, il doit être traité avec beaucoup de précautions – il est presque impossible de garantir la dépersonnalisation d'un tel jeu de données.

## Les identifiants cachés

En évaluant le risque, il faut tenir compte de l'ampleur du jeu de données (nombre de personnes participantes et nombre de variables). Dans le cas de jeux de données plus importants, les pirates informatiques peuvent utiliser des approches d'apprentissage automatique. Les cotes et classements personnels ne sont pas considérés comme des éléments identifiants. Malgré cela, Zhang *et al*. (2012) ont fait la description d'un cas où un système d'apprentissage artificiel a été entraîné à traiter une importante collection de profils qui incluaient des classements de films; le système a pu déduire avec une certaine fiabilité quels comptes étaient liés à de multiples personnes utilisatrices. Il est facile d'imaginer d'autres attaques avec des approches semblables – par exemple, faire la comparaison de critiques de livres publiées sur des sites comme Goodreads à des réponses de sondage incluant des classements de livres utilisés dans les thérapies axées sur les traumatismes.Thompson et Sullivan (2020) ont démontré une autre approche où des variables inattendues étaient susceptibles de

réidentifier des personnes ayant répondu à un sondage par le biais d'une attaque utilisant des renseignements géographiques. Elles ont démontré qu'une variable indiquant la distance de la grande ville la plus proche pouvait être combinée au renseignement sur le domicile d'une personne ayant répondu au sondage et habitant une réserve autochtone; cette combinaison permettait de localiser certaines personnes ayant répondu au sondage. Le procédé est difficile à faire à la main, mais très facile à l'ordinateur.

Ces cas sont la preuve qu'il n'y aura jamais une seule mécanique simple pour encadrer la dépersonnalisation des jeux de données. Vous devrez toujours évaluer à quel point des sources externes de renseignements ou de données peuvent chevaucher les données de votre population de recherche et partager certains renseignements identiques. Il y a risque de réidentification quand des informations externes auxquelles des pirates informatiques ont accès peuvent être combinées à des informations d'un jeu de données archivé; chaque jeu de données doit être évalué de façon individuelle.

## La dépersonnalisation des données qualitatives

Nous utilisons généralement des méthodes statistiques d'anonymisation de données sur des données structurées comme celles d'un tableur ou d'une feuille de calcul. Toutefois, les **données qualitatives** sont souvent stockées et analysées dans des formats non structurés (p. ex., des entrevues, des groupes de discussion ou transcriptions d'histoires orales en format texte, audio ou vidéo, des observations ethnographiques notées sur le terrain, etc.). L'anonymisation de données qualitatives non structurées demeure possible, de nombreux logiciels ou outils numériques existent pour faciliter ou automatiser certains des processus (pour un excellent aperçu, visionnez cette discussion entre spécialistes sur la dépersonnalisation des données qualitatives (en anglais uniquement) : https://youtu.be/MbKw3LR2rVo (https://youtu.be/MbKw3LR2rVo)).

Il arrive parfois qu'un individu qui participe à une recherche s'identifie lui-même par inadvertance en répondant à des questions d'entrevue ou en discutant d'une expérience vécue. Prenons par exemple une étude où un bibliothécaire mène des entrevues avec des bibliothécaires d'autres universités; si une personne répond qu'elle travaille pour tel établissement (McGill) et occupe tel poste (spécialiste en gestion des données de recherche), ces renseignements utilisés en combinaison avec d'autres peuvent l'identifier. Le défi des données qualitatives repose sur le fait que les informations identificatoires ne se retrouvent pas dans des catégories prédéterminées (p. ex., l'âge, la religion, le genre) de sorte qu'il n'est pas possible de prédire la quantité de renseignements identificatoires contenus dans un jeu de données avant la collecte et l'analyse de celui-ci.

La chercheuse ou le chercheur peut supprimer les informations identificatoires – une approche similaire à celles utilisées avec les données structurées – cependant, les informations contextuelles sont souvent essentielles aux recherches qualitatives. Par conséquent, la chercheuse ou le chercheur pourra attribuer des codes pour remplacer les catégories d'informations identificatoires. Le Finnish Social Science Data Archive (FSD) recommande l'utilisation de crochets dans les transcriptions pour indiquer les éléments qui ont été

dépersonnalisés afin d'éviter la ponctuation couramment utilisée (Finnish Social Science Data Archive, 2020). Par exemple, une chercheuse peut remplacer un nom d'individu par [Participant1], ou un endroit précis comme Pohénégamook (un petit village au Québec) par [village]. Si le contexte géographique est important, le code peut alors être modifié pour représenter un endroit en général plutôt qu'un village précis tel que [région du Bas-Saint-Laurent].

Lorsque vous dépersonnalisez des informations qualitatives ou rassemblez des informations plus détaillées en catégories, il est important de documenter les décisions et les catégories dans un **guide de codification** qui accompagnera le jeu de données. Par exemple, la chercheuse peut décider d'éliminer le nom des villages dont la population est inférieure à 1000 habitants. La documentation doit détailler les motifs et les définitions afin de permettre la réutilisation potentielle du jeu de données.

Les transcriptions d'entrevues devraient être anonymisées même si la chercheuse ou le chercheur n'envisage pas de publier ses données. Ainsi, le risque de préjudices est réduit en cas de fuite. L'anonymisation devrait être irréversible; en anonymisant, la chercheuse ou le chercheur doit tenir compte à la fois du préjudice potentiel pour les personnes participantes si des informations identificatoires étaient rendues publiques et de sa capacité à analyser les données sans perdre de nuances. Si l'objectif d'une recherche est d'analyser un sujet sensible, les données ne devraient peut-être pas être dépersonnalisées, elles auront donc besoin de mesures supplémentaires de protection.

## Le langage de consentement et l'ETPC 2

Lors de la curation de données d'êtres humains, vous devez être au courant des mesures de protection offertes aux personnes participantes et sous quelles conditions les comités d'éthique de la recherche (CÉR) ont autorisé la recherche. Au Canada, c'est la politique de l'EPTC 2 qui établit les lignes directrices éthiques pour les recherches avec des êtres humains. Dans la plupart des établissements, le CÉR examinera le langage de consentement plus que toute autre composante de la demande afin d'assurer la confidentialité et la protection de la vie privée des personnes participantes, en plus de s'assurer que celles-ci aient été bien informées de la portée et de la nature de leur participation dans la recherche. En vertu des lignes directrices de l'EPTC 2, les formulaires de consentement devraient comprendre les informations suivantes :

- la participation est volontaire;
- les personnes participantes peuvent se retirer de la recherche et ce, même si elle est en cours;
- un énoncé en langage clair (p. ex., sans jargon) qui décrit l'étude de façon sommaire et qui énumère les risques et bénéfices potentiels pour les personnes participantes – cette information est particulièrement importante pour les études qui impliquent des populations vulnérables, des sujets jugés tabous, de la coercition (p. ex., des incitations) et/ou de la duperie quand les personnes participantes ne connaissent

pas le but véritable de la recherche;
- dans l'éventualité où les données seraient accessibles à d'autres chercheuses ou chercheurs ou au public, sous quelles conditions, est-ce qu'elles seront stockées dans des dépôts particuliers*, sous quels formats et avec quels renseignements (p. ex., si les données peuvent comporter des identifiants directs ou indirects).

*Les CÉR peuvent exiger que les chercheuses et chercheurs identifient le dépôt où seront stockées ou publiées les données sur des personnes participantes. Par exemple, un CÉR peut exiger que les données soient uniquement stockées ou publiées dans des dépôts dont les serveurs sont au Canada ou dont l'accès est contrôlé (p. ex., la possibilité de limiter l'accès à certaines personnes particulières).

Les formulaires de consentement devraient préciser les détails entourant le dépôt ou la publication éventuelle de données sur des êtres humains; quelques chercheuses ou chercheurs peuvent avoir l'intention ou l'obligation (p. ex., par les organismes subventionnaires ou les politiques d'un périodique) de rendre leurs données accessibles à la suite de la publication d'une recherche connexe. Autrement, si une chercheuse ou un chercheur est tenu ou choisit de partager ses données, le consentement des personnes participantes devra possiblement à nouveau être demandé par le biais d'un formulaire de consentement amendé, ce qui peut s'avérer difficile, voire impossible, si tous les identifiants directs ont été anonymisés de façon permanente.

Certaines ressources fournissent des modèles ou des exemples de langage pour les formulaires de consentement et les demandes de CÉR en lien avec le stockage et le partage de données d'êtres humains. L'Alliance de recherche numérique du Canada (l'Alliance) a publié une *Boîte à outils pour les données sensibles – destinée aux chercheurs (https://zenodo.org/record/4107186#.Y-O7fxOZPao)* (Groupe d'experts sur les données sensibles du réseau Portage, 2020b) dans laquelle vous pouvez puiser pour rédiger vos formulaires de consentement et bien expliquer les éléments suivants aux personnes participantes : la différence entre l'anonymat et la confidentialité; les obstacles au retrait de l'étude; les paramètres de l'utilisation future des données y compris les processus de supervision (p. ex., l'établissement d'ententes de réutilisation des données ou l'obligation pour les prochains projets de recherche d'obtenir l'autorisation d'un CÉR avant de pouvoir accéder aux données); si les données peuvent être utilisées à d'autres fins que celles du sujet de recherche original et si les données sont ouvertes au public soit entièrement ou en partie. Vous trouverez ci-dessous un texte passe-partout (adapté à partir de plusieurs sources) que vous pouvez modifier et utiliser dans des formulaires de consentement advenant que vos données soient susceptibles d'être partagées. La *Boîte à outils pour les données sensibles – destinée aux chercheurs* (Groupe d'experts sur les données sensibles du réseau Portage, 2020b) contient des exemples supplémentaires pour d'autres types de cas :

> Les organismes subventionnaires et les maisons d'édition demandent souvent aux chercheuses et chercheurs de rendre leurs données de recherche accessibles une fois leur étude terminée. L'accès aux données permet à la communauté de recherche de reproduire les conclusions scientifiques et encourage l'exploration des jeux de données existants. Afin d'assurer la confidentialité et l'anonymat, toute donnée partagée sera dépouillée des renseignements pouvant identifier une personne participante.

Pour plus de ressources et des exemples de langage pour les formulaires de consentement, veuillez vous référer aux guides très complets fournis (en anglais uniquement) par le Inter-university Consortium for Political and Social Research (ICPSR) (https://www.icpsr.umich.edu/web/pages/datamanagement/confidentiality/conf-language.html) et le Finnish Social Science Data Archive (https://www.fsd.tuni.fi/en/services/depositing-data/guidelines-for-research-projects-planning-to-archive-their-data/).

Le Qualitative Data Repository (QDR) basé à l'Université Syracuse à New York propose un guide pour le consentement éclairé (en anglais uniquement) en lien avec les études qualitatives, notamment pour des entrevues et des histoires orales où les identifiants directs sont conservés dans la publication du jeu de données (Qualitative Data Repository, s.d.-b). Le QDR propose aussi des modèles (https://qdr.syr.edu/guidance/templates) (en anglais uniquement) pour la publication de documents d'archives et pour l'obtention du consentement pour la diffusion de données dépersonnalisées ou identificatoires (Qualitative Data Repository, s.d.-a). L'exemple suivant provient du QDR pour le dépôt d'informations potentiellement identificatoires :

> *Les données générées à partir de l'information que vous avez fournie dans le cadre de notre échange pourront être partagées avec la communauté de recherche (généralement sous forme numérique via Internet) afin de faire progresser les connaissances scientifiques. En raison de la nature des informations, une dépersonnalisation complète des données peut s'avérer impossible. Conséquemment, d'autres mesures seront appliquées avant le partage. Je prévois déposer les données dans le DÉPÔT X, ou dans un dépôt similaire dans le domaine des sciences sociales. Vos données NE SERONT ACCESSIBLES QUE SOUS LES CONDITIONS SUIVANTES. Malgré ces mesures, il n'est pas possible de prédire comment les personnes qui accèdent aux données les utiliseront*[1] [traduction].

Le Data Curation Network fournit un guide complet pour la curation des données d'êtres humains (en anglais uniquement), y compris pour la révision du langage de consentement. Le guide d'introduction sur la curation des données avec des êtres humains offre du soutien aux responsables de la curation et des dépôts de données sur les bonnes questions à poser en lien avec, notamment, le processus de consentement, le langage de consentement et les lacunes potentielles entre le jeu de données et le langage de consentement.

## Autres catégories de données sensibles

Les données d'êtres humains sont souvent considérées sur un pied d'égalité avec les données sensibles. Toutefois, d'autres catégories de données sensibles qui n'impliquent pas des êtres humains sont tout aussi

1. "Data generated from the information you provide in our interaction may be shared with the research community (most likely in digital form via the Internet) to advance scholarly knowledge. Due to the nature of the information, full de-identification of those data might not be possible. As a result, other measures will be taken before sharing. I plan to deposit the data at REPOSITORY X, or at a similar social science domain repository. Your data will BE MADE AVAILABLE UNDER THE FOLLOWING ACCESS CONDITIONS. Despite my taking these measures it is not possible to predict how those who access the data will use them."

importantes. Dans le cas de chercheuses et chercheurs qui travaillent en collaboration avec des partenaires d'industrie pour développer des technologies et des inventions, les données peuvent être considérées comme relevant du « secret commercial » et doivent alors être protégées conformément aux obligations contractuelles (Gouvernement du Canada, 2019). La poursuite du profit comme principal objectif d'une recherche est, en théorie, contraire aux vocations universitaires. Par contre, ce type de partenariat permet un plus large accès à des ressources et des infrastructures que celles qui sont disponibles par l'entremise de sources de financement universitaires ou publiques. Par exemple, les vaccins de la COVID-19 ont été développés beaucoup plus rapidement en raison des partenariats entre la communauté de recherche universitaire et les compagnies pharmaceutiques privées.

Voici d'autres catégories de données sensibles :

- la propriété intellectuelle;
- les données à double usage;
- les données assujetties à un contrôle des importations/exportations;
- les données sous licence tierce;
- la localisation d'espèces en péril.

Des préoccupations en matière de propriété intellectuelle peuvent survenir lorsque des données sont associées à des demandes de brevet en instance, à des recherches brevetées ou à d'autres informations protégées par le droit d'auteur. Les gens qui détiennent les droits sur les données peuvent décider d'accorder l'accès et permettre leur réutilisation. Si la propriété intellectuelle est liée à une source potentielle de revenus, le partage et l'accès aux données sont peu courants. Voici quelques considérations importantes en matière de propriété intellectuelle : qui est propriétaire des données, quelles sont les conditions d'utilisation (ou la licence) des données ainsi que toute autre condition liée à l'utilisation ou la réutilisation des données. Le chapitre 12 « Planification de la gestion des données pour les processus de travail en science ouverte » aborde plus en détail les considérations en matière de propriété intellectuelle.

Les données à double usage se rapportent aux données développées à des fins civiles qui peuvent également être utilisées pour des applications militaires. Par exemple, la technologie de reconnaissance faciale peut avoir été développée pour des applications de téléphones intelligents, mais les jeux de données sous-jacents peuvent être utilisés pour former des modèles d'apprentissage automatique afin de suivre des dissidents politiques ou de déployer des drones armés. Un autre exemple de données sensibles à double usage comprend des informations techniques sur des infrastructures essentielles. Le Canada a mis en place une réglementation et des procédures d'évaluation pour déterminer si la recherche est à double usage et quelles sont les mesures de protection à appliquer selon le niveau de risque.

Les données assujetties au contrôle des importations/exportations (marchandises contrôlées (https://www.tps

gc-pwgsc.gc.ca/pmc-cgp/quellessont-whatare-fra.html)) sont liées aux données à double usage dans la mesure où elles ont des incidences sur les applications et renseignements militaires pouvant traverser les frontières canadiennes (Gouvernement du Canada, s.d.). Dans la définition des marchandises contrôlées sont incluses les armes en provenance des États-Unis. Ces règlements sont en place pour empêcher la participation des chercheuses et chercheurs, qu'elle soit intentionnelle ou non, dans le trafic d'armes ou de technologies de l'armement.

Les tiers se rapportent aux entités autres que la personne responsable de la recherche et son établissement. L'utilisation des données par un tiers nécessite une licence de l'entité propriétaire des données. Par exemple, les démographes peuvent acheter des jeux de données de Statistiques Canada à condition que les données soient utilisées ou partagées que par des chercheuses ou chercheurs du même établissement. Des ententes sur l'utilisation des données précisent plusieurs choses : qui peut accéder aux données, à quelle(s) fin(s), quand, où les données seront stockées, si une partie des données peut être déposée et si les données doivent être détruites ou préservées à la fin de l'étude. Dans la plupart des cas, ces ententes interdisent aux chercheuses et chercheurs le dépôt ou la publication des jeux de données sous-jacents utilisés pour leur recherche.

Les informations sur la localisation des espèces menacées sont considérées comme une catégorie de données sensibles en raison de leur possible utilisation malveillante qui pourrait nuire à ces espèces. Prenons l'exemple du projet d'un chercheur qui place des étiquettes numériques de géolocalisation sur des rhinocéros pour étudier leurs déplacements. Des braconniers qui accèdent à ces données pourraient les utiliser pour localiser et chasser les rhinocéros qui constituent une espèce en péril.

L'identification des personnes participantes n'est pas nécessairement une préoccupation aussi importante lorsque les chercheuses et chercheurs travaillent avec ces autres catégories de données sensibles mais les responsables de la recherche auront à se soucier davantage des mesures de protection et de cybersécurité, des responsabilités légales et de la conformité. La **gestion des données de recherche** (GDR) pour ce type de données implique d'autres éléments comme des accès chiffrés ou protégés par mot de passe (p. ex., l'authentification multifactorielle, la transmission sécurisée de données via un réseau privé virtuel (VPN)), des procédés sécurisés de stockage et de sauvegarde, l'interdiction d'utiliser des appareils personnels pour interagir avec les données et des vérifications robustes de sécurité pour l'identification potentielle de fuites.

## La préservation et le partage de données sensibles

Certains dépôts de données numériques acceptent les données sensibles, notamment le Inter-university Consortium for Political and Social Research (ICPSR) (https://www.icpsr.umich.edu/), le Qualitative Data Repository (QDR) (https://data.qdr.syr.edu/), et le Finnish Social Science Data Archive (https://www.fsd.tuni.fi/en/). Toutefois, aucun dépôt canadien ne les accepte.

L'Alliance travaille actuellement sur un projet pilote pluriannuel visant à établir un partenariat avec des universités canadiennes dans le but de soutenir la mise en œuvre d'une infrastructure qui permettrait de contrôler l'accès à des données sensibles. La technologie doit se conformer aux politiques et lois institutionnelles, provinciales et fédérales et doit dépendre d'une infrastructure située au Canada. Le projet a débouché sur un outil intégrant le cryptage à divulgation nulle de connaissance afin que les jeux de données sensibles puissent être transférés d'un environnement de dépôt sécurisé vers les chercheuses ou chercheurs ou vice versa. La divulgation nulle de connaissance signifie que les gens qui administrent un système ne possèdent pas la clé pour déchiffrer les fichiers dans leur système. Les clés de déchiffrement des données sont stockées sur une plateforme indépendante. La chercheuse ou le chercheur qui souhaite accéder à un jeu de données sensibles doit télécharger les données chiffrées du dépôt de données pour ensuite recevoir le mot de passe de la plateforme de gestion de la clé.

Plusieurs dépôts de données d'établissements universitaires canadiens ont accès à une instance de Dataverse, et plusieurs utilisent Borealis, une instance de Dataverse gérée par Scholar's Portal. Les conditions d'utilisation de Borealis ne permettent pas le dépôt de données sensibles. Toutefois, le consortium responsable du développement et de la maintenance de Borealis a choisi de s'en remettre aux CÉR pour déterminer si un jeu de données est de nature sensible. La sensibilité n'est pas un concept binaire – une donnée peut être plus ou moins sensible – le processus pour déterminer le niveau de sensibilité des données peut exiger des calculs complexes. Les dépôts de données peuvent accepter les jeux de données anonymisées d'êtres humains sans nécessairement les définir comme étant de nature sensible.

Pour préserver et partager des données sensibles, une chercheuse ou un chercheur peut parfois garder ses données localement tout en publiant une documentation et des **métadonnées** dans la collection Dataverse de son établissement pour que d'autres chercheuses et chercheurs puissent y découvrir les renseignements et les procédures pour accéder aux données. Les bibliothèques peuvent soutenir ce type d'initiative en créant des espaces sûrs et isolés du réseau pour assurer une préservation et une sauvegarde sécurisée, particulièrement dans le cas de données stockées à long terme. La bibliothèque doit alors travailler de concert avec la personne qui dépose des données pour assurer la mise en place de protocoles d'accès appropriés. Dans l'encadré ci-dessous, vous trouverez un exemple de langage pour le formulaire de dépôt :

**Formulaire de dépôt : Conditions pour le dépôt, le stockage, le partage et la réutilisation**

La personne qui dépose accorde à la bibliothèque le droit de stockage et de gestion sécuritaire

des données y compris pour la transformation, le déplacement vers d'autres plateformes et la création de copies de sauvegarde pour assurer la préservation.

- indéfiniment ou jusqu'à leur retrait
- jusqu'à la date suivante, après laquelle les données seront supprimées

Est-ce qu'un enregistrement de ce jeu de données peut être partagé dans <archive locale> afin qu'il puisse être découvert? Si oui, veuillez fournir toute restriction en lien avec le partage de la documentation.

Veuillez préciser comment et sous quelles conditions les données peuvent être partagées avec la communauté de recherche extérieure à l'équipe originale. **À noter que votre formulaire de consentement original doit, selon le cas, permettre cette réutilisation**.

- Les données ne peuvent être partagées qu'avec la permission explicite de la ou des personne(s) suivante(s) (p. ex., le déposant, les membres de l'équipe de recherche originale, le comité de révision des données, etc.).
    - Veuillez indiquer les personnes et fournir leurs coordonnées.
- Les données peuvent être partagées conformément à certaines conditions (p. ex., l'autorisation d'un comité d'éthique en recherche, l'établissement d'un plan de gestion sécuritaire des données qui fait état des mesures prévues pour assurer la sécurité des données lors de leur réutilisation, la signature d'un document avec des conditions).

Veuillez préciser les restrictions éthiques en matière de réutilisation. Selon le cas, joignez à votre dépôt de données une copie du formulaire original de consentement.

# Conclusion

En évaluant les risques et les préjudices potentiels, les chercheuses et chercheurs doivent tenir compte d'une foule d'éléments : les politiques, lois et règlements au niveau institutionnel, provincial, fédéral; les exigences des organismes subventionnaires, les normes disciplinaires ainsi que les obligations contractuelles. Les préjudices peuvent atteindre plusieurs parties concernées, y compris les personnes participantes, les établissements, la chercheuse ou le chercheur, la communauté, le pays et toutes autres entités associées.

Voilà pourquoi plusieurs établissements mettent en place de façon formelle un classement pour les données sensibles avec une échelle de niveaux de risques (p. ex., très élevé, élevé, modéré et faible). Les établissements

doivent tenir compte de différents facteurs locaux et de la gouvernance dans l'évaluation des risques, ce qui entraîne parfois quelques difficultés. Par exemple, le classement des données de recherche dans plusieurs établissements se fait selon les mêmes balises que les **données administratives** ou d'entreprise, complexifiant l'application des différents niveaux de risques dans certains contextes – notamment à l'Université de la Colombie-Britannique (2021) où toutes les informations électroniques sont classées uniformément avec seulement une référence générale aux données de recherche. D'autres universités ont établi des lignes directrices qui comprennent des exemples précis en lien avec la recherche comme l'Université de Calgary (2015) qui inclut les recherches « avec des êtres humains identifiables » comme un exemple de situations comportant des risques très élevés. L'Université Harvard (2020) a mis en place un système consacré à l'identification des niveaux de risques et de préjudices dans les données de recherche. Les données comportant des risques mortels pour le sujet appartiennent à la catégorie de risques la plus élevée. Cette catégorie est définie ainsi : « toute donnée sensible qui est susceptible d'entraîner de graves préjudices pour le sujet ou toute donnée comportant des exigences contractuelles en matière de mesures de sécurité exceptionnelles[2] » [traduction].

Les bibliothèques fournissent les outils, l'information et la formation nécessaires aux gens qui font de la recherche pour qu'ils puissent préserver et partager leurs données de façon éthique et responsable. Mais c'est aux personnes responsables de la recherche qu'il incombe de faire preuve de diligence raisonnable en lien avec les risques.

### Questions de réflexion

1. Quelle est la principale politique canadienne en matière d'éthique à propos de la recherche qui utilise des données d'êtres humains?
2. Énumérez trois identifiants directs et trois quasi-identifiants dans les données d'êtres humains.
3. Une étudiante de cycle supérieur mène une étude sur le terrain au sujet d'une espèce de tortue menacée le long du fleuve Saint-Laurent. Sur une feuille de calcul stockée localement sur son ordinateur, elle fait le suivi des tortues et enregistre les informations suivantes à chaque observation : les latitudes et longitudes, la distance des sites industriels à proximité et

2. *"sensitive data that could place the subject at severe risk of harm or data with contractual requirements for exceptional security measures."*

le nombre de tortues observées. Dans quelle mesure la chercheuse travaille-t-elle avec des données sensibles?

Voir le solutionnaire pour les réponses.

**Éléments clés à retenir**

- La dépersonnalisation est le processus d'élimination dans un jeu de données de tout renseignement susceptible de porter atteinte à la vie privée des personnes qui participent à une recherche.
- Les données sensibles sont des données qui ne peuvent être partagées sans risquer de trahir la confiance ou de causer des préjudices à une personne, une entité ou une communauté.
- Les renseignements identificatoires sont des renseignements d'un jeu de données qui, soit seuls ou en combinaison avec d'autres, peuvent entraîner la divulgation de l'identité d'une personne.
- L'évaluation statistique des risques de divulgation est le processus d'évaluation mathématique des quasi-identifiants d'un jeu de données pour démontrer l'anonymisation des données.
- En évaluant le niveau de risque d'un jeu de données, vous devez tenir compte des éléments suivants : les détails à l'intérieur du jeu de données qui ont le potentiel de réidentifier un individu, soit individuellement, soit en combinaison avec d'autres; les renseignements extérieurs au jeu de données qui pourraient être jumelés aux données dans le jeu de données ou qui fournissent des informations supplémentaires sur la population de l'étude; le niveau de préjudices potentiels aux individus ou aux communautés en cas de diffusion des données.
- Les principales réglementations sur les données de recherche se trouvent au niveau provincial et territorial puisque les universités ne relèvent pas de la juridiction fédérale des lois sur la protection de la vie privée. La Loi sur la protection des renseignements personnels s'applique aux organismes gouvernementaux tandis que la Loi sur la protection des renseignements personnels et les documents électroniques (LPRPDE) s'applique aux entités commerciales du secteur privé. Les chercheuses et chercheurs qui travaillent avec ces organismes ou qui utilisent des données recueillies par eux (p. ex., les dossiers médicaux)

doivent être au courant de ces réglementations. Les provinces et territoires canadiens comportent tous au moins une loi en lien avec la protection de la vie privée qui peut s'appliquer à la recherche; il est donc important de s'informer des lois de votre juridiction. Au niveau fédéral, l'*Énoncé de politique des trois conseils sur l'éthique de la recherche avec des êtres humains* (EPTC2) constitue le plus important cadre régissant la conduite de la recherche.

## Lectures et ressources supplémentaires

Alder, S. (2023, 16 mai). *What is Considered PHI Under HIPAA*? The HIPAA Journal. https://www.hipaajournal.com/what-is-considered-phi-under-hipaa/ (https://www.hipaajournal.com/what-is-considered-phi-under-hipaa/)

Groupe en éthique de la recherche. (2022). *Énoncé de politique des trois conseils : Éthique de la recherche avec des êtres humains – EPTC 2 (2022)*. Gouvernement du Canada. https://ethics.gc.ca/fra/policy-politique_tcps2-eptc2_2022.html (https://ethics.gc.ca/fra/policy-politique_tcps2-eptc2_2022.html)

Krafmiller, E. et Prasad, R. (2021, 16 juin). *Dataverse and OpenDP* [Présentation]. Dataverse Community Meeting 2021. https://youtu.be/q3irpQ4rOyU?t=250 (https://youtu.be/q3irpQ4rOyU?t=250)

Réseau Portage, Groupe de travail sur la COVID-19. (2020). *Directives sur la dépersonnalisation des données (Version 2)*. Zenodo. https://zenodo.org/record/4452825#.Y-Wj7xOZPao (https://zenodo.org/record/4452825#.Y-Wj7xOZPao)

Sweeney, L. (2000). *Simple demographics often identify people uniquely* (Data Privacy Working Paper 3). Carnegie Mellon University. http://ggs685.pbworks.com/w/file/fetch/94376315/Latanya.pdf (http://ggs685.pbworks.com/w/file/fetch/94376315/Latanya.pdf)

Thorogood, A. (2018). Canada: will privacy rules continue to favour open science? *Human Genetics, 137*(8), 595–602. https://doi.org/10.1007/s00439-018-1905-0 (https://doi.org/10.1007/s00439-018-1905-0)

## Bibliographie

Centre de gouvernance de l'information des Premières Nations. (s.d.). *FAQ sur les principes de PCAP®*.

https://fnigc.ca/fr/les-principes-de-pcap-des-premieres-nations/ (https://fnigc.ca/fr/les-principes-de-pcap-des-premieres-nations/)

Commissariat à la protection de la vie privée au Canada. (s.d.). *Aperçu des lois sur la protection des renseignements personnels au Canada*. https://www.priv.gc.ca/fr/sujets-lies-a-la-protection-de-la-vie-privee/lois-sur-la-protection-des-renseignements-personnels-au-canada/02_05_d_15/ (https://www.priv.gc.ca/fr/sujets-lies-a-la-protection-de-la-vie-privee/lois-sur-la-protection-des-renseignements-personnels-au-canada/02_05_d_15/)

Commissariat à la protection de la vie privée au Canada. (2020). *Questions et réponses – projet de loi no 64*. https://www.priv.gc.ca/fr/nouvelles-du-commissariat/nouvelles-et-annonces/2020/qa_20200924/ (https://www.priv.gc.ca/fr/nouvelles-du-commissariat/nouvelles-et-annonces/2020/qa_20200924/)

Domingo-Ferrer, J. et Torra, V. (2008). A critique of k-anonymity and some of its enhancements. Dans S. Jakoubi, S. Tjoa, et E. R. Weippl (dir.), *ARES 2008: Third International Conference on Availability, Reliability and Security Proceedings, March 4-7, 2008*. (pp. 990-993). IEEE Computer Society. https://doi.org/10.1109/ARES.2008.97 (https://doi.org/10.1109/ARES.2008.97)

Finnish Social Science Data Archive. (2020). *Anonymisation and personal data*. https://www.fsd.tuni.fi/en/services/data-management-guidelines/anonymisation-and-identifiers/ (https://www.fsd.tuni.fi/en/services/data-management-guidelines/anonymisation-and-identifiers/)

Gouvernement du Canada. (s.d.). *Quelles sont les marchandises contrôlées*. https://www.tpsgc-pwgsc.gc.ca/pmc-cgp/quellessont-whatare-fra.html (https://www.tpsgc-pwgsc.gc.ca/pmc-cgp/quellessont-whatare-fra.html)

Gouvernement du Canada. (2019). *Lignes directrices sur la sécurité nationale pour les partenariats de recherche*. https://science.gc.ca/site/science/fr/protegez-votre-recherche/lignes-directrices-outils-pour-mise-oeuvre-securite-recherche/lignes-directrices-securite-nationale-pour-partenariats-recherche

Groupe d'experts sur les données sensibles. (2020a). *Boîte à outils pour les données sensibles — destiné aux chercheurs Partie 1: Glossaire terminologique sur l'utilisation des données sensibles à des fins de recherche*. Zenodo. https://doi.org/10.5281/zenodo.4088986 (https://doi.org/10.5281/zenodo.4088986)

Groupe d'experts sur les données sensibles. (2020b). *Boîte à outils pour les données sensibles — destiné aux chercheurs Partie 3 : Langage en matière de gestion de données de recherche pour le consentement éclairé*. Zenodo. https://doi.org/10.5281/zenodo.4107186 (https://doi.org/10.5281/zenodo.4107186)

Han, Y., Li, S., Cao, Y., Ma, Q. et Yoshikawa, M. (2020). Voice-indistinguishability: Protecting voiceprint in privacy-preserving speech data release. *2020 IEEE International Conference on Multimedia and Expo*

*(ICME)*, 1-6, https://doi.org/10.1109/ICME46284.2020.9102875 (https://doi.org/10.1109/ICME46284.2020.9102875)

Henne, B., Koch, M. et Smith, M. (2014). On the awareness, control and privacy of shared photo metadata. Dans N. Christin et R. Safavi-Naini (dir.), *Financial Cryptography and Data Security. FC 2014. Lecture Notes in Computer Science* (pp. 77-88). Springer. https://doi.org/10.1007/978-3-662-45472-5_6 (https://doi.org/10.1007/978-3-662-45472-5_6)

Qualitative Data Repository. (s.d.-a). *Templates for researchers*. https://qdr.syr.edu/guidance/templates#informed%20consent (https://qdr.syr.edu/guidance/templates#informed%20consent)

Qualitative Data Repository. (s.d.-b). *Informed consent*. https://qdr.syr.edu/guidance/human-participants/informed-consent

Ross, M. W., Iguchi, M. Y. et Panicker, S. (2018). Ethical aspects of data sharing and research participant protections. *American Psychologist, 73*(2), 138-145. http://dx.doi.org/10.1037/amp0000240 (http://dx.doi.org/10.1037/amp0000240)

Samarati, P. et Sweeney, L. (1998). *Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression* (Technical Report SRI-CSL-98-04). Computer Science Laboratory, SRI International. https://epic.org/wp-content/uploads/privacy/reidentification/Samarati_Sweeney_paper.pdf (https://epic.org/wp-content/uploads/privacy/reidentification/Samarati_Sweeney_paper.pdf)

Thompson, K. et Sullivan, C. (2020). Mathematics, risk, and messy survey data. *IASSIST Quarterly, 44*(4), 1-13. https://doi.org/10.29173/iq979 (https://doi.org/10.29173/iq979)

Université Harvard. (2020, 22 avril). *Data Security Levels – Research Data Examples*. https://security.harvard.edu/data-security-levels-research-data-examples (https://security.harvard.edu/data-security-levels-research-data-examples)

Université de la Colombie-Britannique. (2021). *Security classification of UBC electronic information*. https://cio.ubc.ca/information-security-standards/U1 (https://cio.ubc.ca/information-security-standards/U1)

Université de Calgary. (2015, 1 janvier). *Information Security Classification Standard*. https://www.ucalgary.ca/legal-services/sites/default/files/teams/1/Standards-Legal-Information-Security-Classification-Standard.pdf (https://www.ucalgary.ca/legal-services/sites/default/files/teams/1/Standards-Legal-Information-Security-Classification-Standard.pdf)

Wolford, B. (2018, 5 novembre). *Everything you need to know about the 'Right to be forgotten'*. GDPR.EU. https://gdpr.eu/right-to-be-forgotten/ (https://gdpr.eu/right-to-be-forgotten/)

Zhang, A., Fawaz, N., Ioannidis, S. et Montanari, A. (2012). Guess who rated this movie: identifying users through subspace clustering. *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, 944-953. https://dl.acm.org/doi/10.5555/3020652.3020750 (https://dl.acm.org/doi/10.5555/3020652.3020750)

## À propos des auteurs

### Dr. Alisa Beth Rod

Alisa Beth Rod est spécialiste de la gestion des données de recherche à la bibliothèque de l'Université McGill. Alisa détient une maîtrise et un doctorat en sciences politiques de l'Université de Californie, Santa Barbara, et un baccalauréat en bioéthique de l'American Jewish University. Avant de rejoindre McGill, Alisa a été méthodologiste d'enquête chez Ithaka S+R, puis directrice associée de l'Empirical Reasoning Center au Collège Barnard de l'Université Columbia. Elle possède une expérience approfondie en collecte et utilisation de données d'êtres humains dans le contexte de la recherche par sondage, des méthodes qualitatives et des systèmes d'information géographique (SIG).

### Kristi Thompson

Kristi Thompson est bibliothécaire en gestion des données de recherche à l'Université Western. Elle a précédemment occupé les postes de bibliothécaire des données à l'Université de Windsor et de spécialiste des données à l'Université Princeton. Elle détient un baccalauréat en informatique de l'Université Queen's et une maîtrise en science de l'information de l'Université Western. Kristi soutient des projets de recherche, administre des logiciels d'archivage de données, travaille avec les comités d'éthique de la recherche de l'Université Western et participe au niveau national au développement de l'infrastructure des données de recherche. Elle a coédité le livre *Databrarianship : the Academic Data Librarian in Theory and Practice* et a publié sur des sujets allant des algorithmes d'anonymisation des données à la psychologie intergénérationnelle.
kthom67@uwo.ca (mailto:kthom67@uwo.ca) | ORCID 0000-0002-4152-0075 (https://orcid.org/0000-0002-4152-0075)

14.

# LA GESTION DES DONNÉES DE RECHERCHE QUALITATIVES

Dr. Joel T. Minion

**Objectifs d'apprentissage**

À la fin de ce chapitre, vous pourrez :

1. Identifier les caractéristiques des données qualitatives par rapport à d'autres formes de données de recherche.
2. Comprendre les procédés interactifs avec lesquels les chercheuses et chercheurs génèrent et gèrent leurs données qualitatives.
3. Décrire comment la gestion des données de recherche pourrait mieux inclure les données qualitatives et les besoins des chercheuses et chercheurs qui les utilisent.
4. Faire la promotion d'une plus grande intégration de tout type de données de recherche dans les principes, les politiques, les stratégies et les pratiques de gestion des données de recherche.

## Introduction

Une saine gestion des données est essentielle pour l'excellence en recherche. La plupart des établissements d'enseignement supérieur appuient les initiatives dans ce domaine, mais peu s'attardent sur les **données qualitatives** et leurs chercheuses et chercheurs. Si vous participez à des formations typiques, on devra vous pardonner de penser que la gestion des données de recherche s'applique surtout aux données comprenant des chiffres ou des images géospatiales. À quelques exceptions près (p. ex., les principes de PCAP® des Premières Nations (https://fnigc.ca/fr/les-principes-de-pcap-des-premieres-nations/) – propriété, contrôle, accès et

possession), la reconnaissance des données de recherche qualitatives semble souvent secondaire. C'est sans doute en raison de la nature des données qualitatives. Elles sont souvent textuelles ou orales, plutôt que numériques, et recueillies uniquement de l'être humain. Cette qualité rend les données qualitatives particulièrement identifiables. Les données qualitatives obligent également les chercheuses et chercheurs à tenir compte des relations et contextes sociaux et les données sont communément générées par des études qui traitent de questions sensibles ou de communautés marginalisées. De tels défis peuvent expliquer pourquoi les données de recherche qualitatives ne réussissent que rarement à s'insérer dans les structures actuelles de GDR.

Dans ce chapitre, nous examinerons les raisons qui expliquent le peu d'études dans ce domaine de GDR et nous proposerons des moyens de pallier ces lacunes. Le contenu reflète mes 25 dernières années d'expérience en tant que bibliothécaire, chercheur qualitatif en santé, gestionnaire de données et formateur au Canada et en Europe. Comme la majorité de mes collègues en recherche qualitative, j'ai du mal à m'insérer dans les principes, les politiques, les stratégies et les pratiques existantes en GDR. Les spécialistes sont peu nombreux et les ressources, limitées; le chapitre n'est donc pas un guide pratique en la matière.

Les données qualitatives existent sous plusieurs formes et il existe une multitude de moyens de les générer, les analyser, les archiver, les partager, et parfois, les réutiliser. Nous aborderons de quelle façon la gestion des données de recherche qualitatives s'insère dans le processus de recherche. Pour la chercheuse ou le chercheur, cela implique d'explorer la façon de penser à ses données et comment mieux les organiser. Pour les bibliothécaires, archivistes ou autres types de spécialistes des données, la discussion devrait améliorer vos compétences en gestion de l'information avec une compréhension plus approfondie des origines des données qualitatives.

Le chapitre est divisé en trois sections : (1) la nature des données de recherche qualitatives; (2) la façon dont ces données se reflètent dans le processus de recherche qualitative; et (3) les défis de GDR au niveau de la collecte des données qualitatives. Pour terminer, nous discuterons des moyens à employer pour améliorer la gestion des données de recherche qualitatives.

## La nature des données qualitatives

Les données qualitatives ne sont pas créées et analysées de la même façon que les données quantitatives ou que celles des **humanités numériques**. Toutefois, il ne faut pas croire que les différents types de données de recherche soient incompatibles et qu'elles ne peuvent être utilisées ensemble. Plusieurs chercheuses et chercheurs emploient une variété de méthodes dans leur recherche, par exemple une combinaison d'entrevues et d'évaluations psychométriques pour répondre à des questions du genre « Comment un trouble dépressif clinique est-il vécu par des individus qui soignent leur partenaire de vie en phase initiale de démence? » Les approches de ce genre démontrent bien l'interdépendance des différents types de données de recherche.

## Qu'est-ce qui fait qu'une donnée de recherche est qualitative?

Les données qualitatives ne partagent pas entre elles une seule et unique philosophie ou un seul ensemble de principes méthodologiques. Elles sont générées par des recherches qui examinent des aspects sociaux de la condition humaine en utilisant des méthodes descriptives plutôt que des mesures. Les chercheuses et chercheurs peuvent observer des individus ou communiquer avec eux par une multitude de façons afin de mieux comprendre comment ils interagissent entre eux et comment ils conçoivent leur environnement soit à la maison, au travail, dans la communauté, en recevant des soins de santé, etc. La recherche qualitative trouve ses origines dans les **sciences sociales**, particulièrement en anthropologie, sociologie et psychologie, bien que la perspective qualitative puisse aussi être adoptée dans d'autres disciplines. Les chercheuses et chercheurs en sciences infirmières, par exemple, utilisent couramment des données qualitatives pour étudier certaines expériences vécues par des patientes et patients.

Les données qualitatives peuvent être recueillies au cours d'un seul contact ou par le biais d'interactions sur une plus longue période. Ce qui est retenu est toujours filtré à travers le prisme de l'expérience des chercheuses et chercheurs et de leur interprétation des interactions avec les personnes participantes. Les responsables de la recherche eux-mêmes font donc partie des données. Les données qualitatives sont importantes parce qu'elles fournissent des informations qui ne peuvent être mesurées ou comptées autrement, par exemple, comment la population afghane réfugiée interprète-t-elle les services gouvernementaux en arrivant au Canada ou quelle est l'expérience de compétition d'athlètes paralympiques ou qu'est-ce qui attire certains individus vers des mouvements d'extrême droite.

## À quoi ressemblent les données qualitatives ?

Les méthodes les plus courantes pour générer des données qualitatives se font par le biais d'entrevues, de groupes de discussion et d'observations (p. ex., vous pouvez vous entretenir avec des personnes réfugiées sur leurs expériences par le biais d'entrevues individuelles, mener des groupes de discussion avec des athlètes ou observer ce qui se passe lors d'évènements d'extrême droite). Ces méthodes sont courantes parce que les techniques sont assez simples à apprendre et à mettre en pratique. D'autres méthodes comprennent des histoires orales, des journaux personnels de personnes participantes, des photographies/vidéos, des analyses de documents, des artéfacts (p. ex., de la nourriture, des vêtements) et des réponses ouvertes aux questions de sondages.

Ces méthodes peuvent être utilisées en combinaison avec d'autres, entraînant ainsi une interdépendance des jeux de données. Une chercheuse qui s'intéresse à la nature des collaborations entre climatologues peut effectuer des observations lors d'une conférence où elle peut aussi mener des entrevues avec des personnes

présentes et recueillir les documents des présentations. Les chercheuses et chercheurs en recherche qualitative tiennent souvent un journal **réflexif** pour réfléchir à leur positionnement au sein du projet, pour noter de nouvelles idées et pour identifier de nouvelles pistes de questionnement. Les chercheuses et chercheurs peuvent aussi se tourner vers les médias sociaux, en parcourant par exemple les discussions en ligne où les interactions se font librement, sans intervention de leur part. Même si cette méthode est de plus en plus appliquée pour relever une perspective qualitative, le présent chapitre traite plutôt des données recueillies par la chercheuse ou le chercheur.

### Exercice: Travailler avec des formes moins courantes de données qualitatives

Vous travaillez dans une université où vous êtes bibliothécaire de données. Un étudiant à la maîtrise vous demande conseil sur la façon de gérer les données qui seront recueillies lors d'une étude sur des femmes suivant des traitements contre le cancer du col de l'utérus. Les méthodes de collecte impliqueront conjointement des entrevues et des *photovoix*, une approche que vous connaissez peu. Consultez le document ci-dessous et relevez des questions que vous pourrez poser au sujet des photographies qui seront recueillies et sur la façon dont elles pourraient être gérées.

Wang, C. et Burris, M. A. (1997). Photovoice: Concept, methodology, and use for participatory needs assessment. *Health Education & Behavior, 24*(3), 369-387. https://doi.org/10.1177/109019819702400309 (https://doi.org/10.1177/109019819702400309)

Bien que les données qualitatives prennent plusieurs formes, l'écrit est la forme la plus courante. Les entrevues et groupes de discussion sont généralement enregistrés avec des appareils audios portatifs; les enregistrements sont ensuite retranscrits pour l'analyse.

### La complexité de la transcription

La transcription est souvent longue et difficile à faire sans les équipements appropriés (p. ex., de bons casques d'écoute ou des logiciels spécialisés). Plusieurs chercheuses et chercheurs en

recherche qualitative confient ces tâches à des sous-traitants, ce qui soulève des préoccupations au sujet des coûts et de la possibilité d'avoir besoin de transférer les données à l'extérieur du Canada.

Le processus oblige également les chercheuses et chercheurs à faire des choix quant à l'ampleur des détails à transcrire. Chaque « umm », « euh » ou faux départ doit-il être noté (qualifiée de « transcription intégrale »)? Ou l'objectif est-il simplement de produire une version lisible de ce qui a été dit (une transcription dite « épurée »)? Ces décisions sont cruciales puisque certaines formes d'analyse qualitative peuvent nécessiter des niveaux particuliers de transcription.

Enfin, l'exactitude de toute transcription doit être vérifiée avant d'être analysée, ce qui implique d'écouter les enregistrements en lisant en parallèle les transcriptions correspondantes pour déceler toute erreur ou tout oubli.

Les enregistrements vidéo sont moins courants parce que certaines personnes participantes les considèrent comme plus intrusifs; leur utilisation peut nécessiter plus de démarches pour assurer le consentement des gens. Les vidéos peuvent aussi être plus exigeantes à analyser. Selon les circonstances et les versions, la capture de données d'observation est généralement faite à l'aide de notes écrites à la main ou à l'ordinateur et/ou de notes audios enregistrées sur le terrain (p. ex., les notes écrites peuvent être prises sur les lieux en même temps que des notes audios qui seront retranscrites plus tard).

Le traitement des formes moins courantes de données qualitatives varie énormément. Les copies papier, telles que des procès-verbaux ou des documents de présentation de conférence, ont souvent besoin d'être numérisées avant l'entreposage ou l'analyse. Les journaux personnels des personnes participantes peuvent nécessiter une transcription avant qu'ils puissent être analysés. Les photos numériques peuvent être stockées sous différents formats, selon les besoins ou les préférences de la chercheuse ou du chercheur. Pour ce qui est des artéfacts, la chercheuse ou le chercheur peut décider de les photographier ou de prendre des notes directement à partir des objets eux-mêmes.

## Comment les données qualitatives se traduisent-elles en nouvelles connaissances?

Une gestion efficace des données qualitatives exige une compréhension du processus d'analyse. Le biologiste qui mesure les populations de poissons dans les lacs du nord utilisera probablement un logiciel pour faire une analyse statistique des données (p. ex., SPSS, Stata, R). Mais comment la sociologue peut-elle extraire le sens contenu dans une transcription d'entrevue? Il faut souvent travailler de façon inductive, remonter aux données pour identifier des concepts et des tendances plus larges. L'objectif est de regarder au-delà de ce qui a

été dit, entendu, observé, photographié, etc. pour relever les idées transversales dans un jeu de données complet. L'analyse des données sous forme textuelle peut comprendre le codage ou l'utilisation de logiciels d'analyse de données qualitatives (p. ex., NVivo, Quirkos). Ces logiciels peuvent traiter de grandes quantités de données, mais ils ne peuvent pas les analyser. C'est le travail de la chercheuse ou du chercheur. De plus, ce n'est pas l'ensemble des chercheuses et chercheurs qui veulent coder ou utiliser des logiciels; quelques personnes préfèrent utiliser des transcriptions papier, des surligneurs, des stylos et des fiches.

À certains égards, les données peuvent représenter ce qu'il y a de plus simple dans le processus de recherche qualitative. Après tout, les transcriptions de discussions de groupe peuvent se ressembler d'une étude à l'autre; des pages de textes sur lesquelles a été consigné ce qui a été dit et par qui. Toutefois, le contenu des discussions est un reflet de la personne qui a mené chacune des études et des raisons qui l'ont poussée à s'engager dans la recherche. Les psychologues et anthropologues sont susceptibles d'aborder le même sujet autrement et de poser des questions différentes. Les données ne prennent leur sens qu'une fois analysées. Ce processus est complexe car il n'existe pas d'ontologie, d'épistémologie, de théorie ou de mode d'analyse unique pour toutes les formes de recherche qualitative. Les chercheuses et chercheurs travaillent à partir de leur propre point de vue; les mêmes données peuvent donc être interprétées de façons différentes selon qui mène l'analyse et à quelles fins.

## Comprendre la recherche qualitative

Pour gérer les données qualitatives de façon efficace, vous devez comprendre comment la recherche qualitative est menée. Nous allons considérer les pratiques de recherche qualitative en fonction des données. Le but est de faire le lien entre les données qualitatives et trois éléments clés du processus de recherche : la structure des équipes de recherche, les répercussions de ces structures dans la production des données et le rôle des personnes participantes dans la recherche qualitative.

Contrairement à son équivalent quantitatif qui utilise souvent des processus bien établis (p. ex., des essais cliniques randomisés en recherche médicale, des instruments de sondage validés en psychologie), la recherche qualitative est plus souple et évolutive. Par exemple, les questions d'entrevues peuvent évoluer de façon importante au fil des multiples échanges avec différentes personnes participantes et mener à de nouvelles pistes de questionnement. Il est même possible, en cours de recherche, d'avoir à ajouter ou à éliminer certains types de données (ce serait le cas, par exemple, si les photos s'avéraient moins utiles que prévu). De telles décisions sont rarement prises à la légère, mais le fait que ces décisions puissent être prises reste une caractéristique propre aux recherches qualitatives.

# Les équipes de recherche qualitative

La composition des équipes de recherche qui utilisent des approches qualitatives pour recueillir leurs données peut varier considérablement. Elle peut prendre l'une ou l'autre des formes suivantes :

- Une personne responsable de la recherche avec une étudiante ou un étudiant de cycle supérieur (p. ex., un individu ayant reçu une petite subvention pour mener 20 entrevues sur la façon dont les parents monoparentaux gèrent les défis de garde d'enfants);
- Un groupe de recherche au sein d'une université ou d'un département (p. ex., une chercheuse de haut niveau et deux boursiers en recherches postdoctorales qui mènent des groupes de discussion et utilisent des documents de planification urbaine pour étudier les changements proposés aux configurations de circulation urbaine);
- Une équipe plus large de multiples universités qui regroupe différentes disciplines (p. ex., six chercheuses et chercheurs à mi-carrière en techniques énergétiques, en affaires et en psychologie organisationnelle qui utilisent des observations et des entrevues pour explorer les réseaux de communication des équipes chargées de l'installation de parcs éoliens en mer);
- Un groupe composé de chercheuses et chercheurs internationaux et multidisciplinaires dont la collaboration s'étend sur plusieurs pays et terrains d'enquête (p. ex., une trentaine de chercheuses et chercheurs, des membres de la communauté étudiante des cycles supérieurs, du personnel clinique et des malades partenaires qui étudient les impacts de la COVID-19 longue sur la santé au Canada, aux États-Unis et au Royaume-Uni, en utilisant des entrevues, des analyses d'archives médicales et une enquête longitudinale).

Les chercheuses et chercheurs peuvent s'impliquer dans l'un ou plusieurs de ces types d'équipes au fil de leur carrière, mais la plupart finissent par développer une préférence ou des compétences spécialisées pour une ou deux approches précises.

Lorsque la composition d'une équipe de recherche implique plus d'une chercheuse ou d'un chercheur, il y a presque toujours une hiérarchisation des relations et une variété de niveaux de compétences menant à l'attribution de rôles et responsabilités particulières. Comme pour les recherches quantitatives, une personne est désignée chercheuse principale ou un chercheur principal pour diriger le projet et elle peut être appuyée par une ou plusieurs cochercheuses ou cochercheurs. La chercheuse principale ou le chercheur principal détient l'autorité sur l'étude et est redevable aux établissements, aux organisations subventionnaires et aux comités d'éthique à toutes les étapes du projet. Pour les équipes plus importantes, la personne responsable du projet peut avoir que très peu d'implications directes sur les activités de recherche au quotidien, notamment en matière de production et de gestion des données. Les chercheuses et chercheurs en début de carrière (p. ex.,

les membres de la communauté étudiante des cycles supérieurs et postdoctoraux) sont fréquemment chargés de la collecte, du traitement, de l'organisation et de la protection des données.

## Le rapport entre les équipes de recherche et les données

La structure d'une équipe de recherche a des conséquences sur la façon dont la recherche qualitative est menée et sur le type de données générées. Dans ce rapport entre les équipes et les données, deux éléments valent la peine d'être soulignés.

Le premier élément concerne la nature **itérative** des recherches qualitatives et son impact sur les données. Les données qualitatives sont souvent analysées dès qu'elles sont recueillies, ce qui implique que les données du présent peuvent influencer celles de l'avenir. Par exemple, une chercheuse pourra utiliser ce qu'elle a appris lors d'une entrevue pour déterminer les questions de l'entrevue suivante. Les modifications peuvent être minimes (p. ex., la reformulation d'une question pour la rendre plus claire) ou importantes (p. ex., l'ajout ou le retrait d'une série complète de questions). Les études plus importantes dépendent souvent de séances de **récapitulation entre collègues** tout au long du processus de collecte des données afin de générer de nouvelles idées, de discuter des défis d'ordre pratique ou d'améliorer les compétences de la chercheuse ou du chercheur. Ainsi, les données qualitatives sont recueillies progressivement et de façon réflexive.

Le deuxième élément de ce rapport entre équipe et données implique l'attribution des rôles et responsabilités à l'intérieur des équipes et la façon dont cette attribution peut influencer les données. Puisque les données qualitatives sont intimement liées aux circonstances de leur collecte, celle ou celui qui recueille les données aura un impact sur le choix de ce qui est recueilli. À défaut de capturer des détails sur le processus de collecte, les données qualitatives peuvent perdre leur capacité à soutenir des analyses rigoureuses. Par exemple, la transcription des échanges d'un groupe de discussion nécessite l'ajout de détails particuliers sur les personnes participantes (p. ex., l'âge, le travail, le niveau d'éducation) et de notes sur le ton de la discussion et sur la nature des interactions entre les personnes participantes (p. ex., un roulement des yeux ne peut être perçu dans un enregistrement audio).

Si chaque membre d'une équipe de recherche capture et traite les données selon ses propres balises, certaines variations peuvent survenir ou certains détails critiques, être omis. Pour éviter ces dérapages, certains individus peuvent être désignés pour agir à titre de **personne responsable de l'intendance des données** (idéalement, plus d'une personne devrait être responsable de l'intendance des données au cas où l'une d'elles quitte le projet). Ces gens seront chargés d'assurer une uniformité dans la gestion des données (p. ex., la désignation des fichiers, la structure des dossiers).

### Exercice: Capturer le contexte

Félicitations! Vous venez d'être nommé stagiaire postdoctoral pour une étude qui observe les échanges verbaux et comportementaux dans les salles d'audience au Canada. L'objectif est d'étudier les différences dans les interactions des juges, des services juridiques, des parties demanderesses et défenderesses issus des minorités visibles. L'équipe du projet comprend un chercheur principal, une cochercheuse d'une deuxième université, deux récents titulaires de doctorat situés ailleurs, un coordonnateur d'étude et une étudiante à la maîtrise.

L'étude impliquera jusqu'à 600 heures d'observation de la part de quatre membres de l'équipe dans cinq villes différentes. Puisqu'aucun d'entre vous ne pourra ni discuter avec les personnes que vous observez ni les enregistrer, la collecte de données se limitera à la prise de notes sur le terrain que chaque personne devra ensuite retranscrire et partager.

L'équipe a discuté du type d'échanges à privilégier pour sa collecte. Mais force est de constater que les personnes chargées de recueillir les données doivent aussi capturer des détails liés au contexte, et ce, de façon uniforme. L'exercice vous demande donc d'identifier (1) quels types d'informations doivent être enregistrées au-delà des échanges eux-mêmes, (2) comment ces informations peuvent être recueillies de façon systématique et (3) comment faire le lien entre les données relatives au contexte et celles de la salle d'audience.

Une **piste de vérification** est une autre pratique utile en recherche qualitative. Cette documentation fait le suivi des activités et de la prise de décisions pendant toute la durée du projet, précisant ce qui s'est passé, à quel moment et pourquoi. Certaines informations seront capturées à même les données, mais pour les projets plus importants, un document distinct accessible par plusieurs membres de l'équipe peut s'avérer nécessaire. L'information enregistrée fait le lien en temps réel entre ce qui se passe au niveau de l'équipe et de la collecte des données. Par exemple, une piste de vérification permet à une équipe de ne pas avoir à se rappeler quand et pourquoi elle a décidé, à mi-parcours d'un projet, d'introduire un nouveau site ou une nouvelle méthode de collecte de données. Malheureusement, peu de normes existent sur la façon de créer et de gérer de tels documents en recherche qualitative. Comme nous le verrons en fin de chapitre, ce genre de défi peut être abordé en réunissant des chercheuses et chercheurs et des spécialistes des données.

## La dimension sociale de la recherche qualitative

Un survol du processus de recherche qualitative ne peut être complet sans discuter du rôle des personnes qui participent à une étude. La recherche qualitative est de nature relationnelle, c'est-à-dire que les chercheuses et chercheurs doivent souvent interagir directement avec les gens qui prennent part à l'étude. La relation peut être passagère, comme dans le cas d'entrevues uniques menées par téléphone. Cela dit, même ces échanges peuvent nécessiter des efforts pour qu'un bon rapport soit établi au cours de l'étape de recrutement. L'établissement de bons rapports est essentiel pour les études qui impliquent des contacts réguliers et prolongés. Il est d'autant plus critique lorsque les chercheuses ou chercheurs doivent interagir avec des personnes de communautés marginalisées ou stigmatisées qui hésitent à participer à des recherches de peur que des informations sensibles soient divulguées ou que leurs données soient utilisées contre elles ou leur communauté. Bien que des mécanismes éthiques de surveillance soient en place pour régir ce type de relation, la complexité des rapports demeure réelle dans la pratique. Quand un rapport étroit devient-il trop étroit? À quel point faut-il croire les informations transmises par les personnes participantes? Un témoin privilégié est-il représentatif de sa communauté ou est-il un cas particulier? Pour bien aborder ces préoccupations, les chercheuses et chercheurs doivent poser un regard critique sur leur rôle dans le processus de recherche et sur leur impact sur les données.

# La production des données

La production des données qualitatives comporte son lot de défis. Les chercheuses et chercheurs doivent se conformer à diverses exigences (p. ex., éthiques, institutionnelles, professionnelles) qui gouvernent leurs choix et leurs actions, sans compter les nombreux défis inhérents à l'étude de l'être humain dans un contexte naturaliste. Nous discuterons de la gouvernance en matière de production des données qualitatives et nous aborderons trois questions particulières qui influent sur la façon dont les données sont recueillies : le recrutement des personnes participantes, le lieu du terrain d'enquête et l'évolution du rapport entre les personnes qui participent à l'étude et leurs données.

## La gouvernance liée à la production des données

Pour toute recherche qualitative, les données ne sont jamais générées sans une quelconque forme d'autorisation ou d'exemption. Outre l'obtention du consentement éclairé des personnes participantes, l'**approbation éthique** est l'autorisation la plus importante. Pour l'obtenir, les chercheuses et chercheurs doivent soumettre les détails de l'étude à un comité d'éthique indépendant, précisant notamment le type de données à recueillir, les moyens prévus pour l'organisation et le stockage ainsi que les mesures pour assurer que la décision des personnes de contribuer aux données soit prise de façon éclairée. Lorsque les études

impliquent des êtres humains, les chercheuses et chercheurs au Canada (y compris la communauté étudiante de premier cycle et des cycles supérieurs) doivent généralement suivre le cours EPTC 2 : FER-2022 (https://tcps2core.ca/welcome?lang=fr) avant de recevoir une approbation éthique. La formation présente les obligations des chercheuses et chercheurs lors de la collecte et de la manipulation des données, en plus des droits des gens qui participent à l'étude.

Lorsqu'une étude a reçu son approbation éthique, les chercheuses et chercheurs doivent respecter leurs engagements. Toute modification (p. ex., aux méthodes de recrutement ou à l'ampleur des données recueillies) doit être soumise et approuvée avant d'être mise en œuvre. L'approbation éthique accompagne aussi d'autres exigences en matière de données, dont celles en place dans les universités (p. ex., la durée de conservation des données). Les données générées à l'extérieur de ces balises sont inutilisables.

Dans certains cas, la collecte des données ne nécessite pas d'approbation éthique, par exemple lorsqu'une étude qualitative est menée pour évaluer un service ou pour mettre en place une initiative d'amélioration de la qualité. Cette approche peut être utilisée dans des recherches en santé qui n'implique pas des gens du public et qui comporte peu de risques pour les personnes participantes (p. ex., le personnel clinique) comme dans le cas d'une étude sur l'expérience des physiothérapeutes dans le traitement de leur clientèle en fauteuil roulant. Un outil d'évaluation peut être utilisé pour déterminer si une évaluation éthique complète est nécessaire (p. ex., ARECCI (https://arecci.albertainnovates.ca/)). Bien que la production des données soit moins rigide dans le cas d'évaluations de service (p. ex., le consentement éclairé n'est pas toujours exigé), elle est rarement moins rigoureuse. Les données ressemblent à celles de toute autre étude qualitative et sont généralement analysées et rapportées de la même façon.

## Les défis courants

La production des données n'est pas toujours sans embûches. Deux défis particuliers reviennent souvent – le recrutement et l'emplacement du terrain d'enquête – tandis qu'un troisième est en évolution : le rapport des personnes participantes avec leurs données de recherche.

### Le recrutement

Les données qualitatives ne peuvent exister sans participantes ou participants. Les méthodes mises en place pour le recrutement et le suivi des individus dans l'étude se traduisent dans leurs données. Les personnes participantes doivent d'abord être identifiées, un processus parfois long et exigeant. Les échantillons d'une étude doivent parfois tenir compte de facteurs tels que l'âge, le genre ou l'éducation. La collecte des détails liés au processus de recrutement peut aider à donner un sens aux données. Le choix des détails recueillis peut varier d'une étude à l'autre, mais les éléments suivants seront probablement inclus :

- la personne qui a été contacté, combien de fois et comment elle ou il a répondu;
- les renseignements personnels/professionnels de l'individu (p. ex., le rôle clinique, ses pronoms);
- les dates, heures et détails de la collecte des données (p. ex., le nom de la chercheuse ou du chercheur, le lieu du terrain d'enquête);
- l'état des données (p. ex., transcrites, anonymisées, prêtes pour l'analyse);
- les restrictions liées à l'utilisation des données (p. ex., si un participant ne veut pas être directement cité);
- la possibilité d'un contact subséquent.

Tous les registres de recrutement doivent demeurer confidentiels et conservés séparément des données afin de prévenir la réidentification. Les détails de recrutement ne sont généralement pas considérés comme des données, mais ils peuvent constituer une forme importante de métadonnées. De telles informations peuvent mettre en lumière, par exemple, le moment où un témoin privilégié s'est joint à l'étude ou si l'entrevue a eu lieu avec une professeure ou son assistant. Ce type de détail n'est pas toujours recueilli dans les données.

## Le lieu du terrain d'enquête

Le deuxième défi en production des données concerne le lieu du terrain d'enquête. Les chercheuses et chercheurs se rendent souvent là où sont les personnes participantes, ce qui comporte son lot d'obstacles. Imaginez que vous êtes chercheuse ou chercheur dans un lieu insolite (p. ex., une communauté isolée de l'Arctique, un campement urbain sous la tente, le département des urgences d'un hôpital à deux heures du matin) et posez-vous les questions suivantes:

- Comment vais-je recueillir les données? (p. ex., avec une enregistreuse audio, du papier et un stylo, des photos)
- Quels sont mes recours si mon moyen de collecte ne fonctionne pas? (p. ex., les piles s'épuisent, le stylo n'écrit plus dans le froid)
- Comment vais-je numériser, sécuriser et/ou faire une sauvegarde de mes données? (p. ex., l'envoi d'enregistrements pour la transcription nécessite une connexion Internet sécurisée)
- Comment vais-je partager mes données avec les autres membres de l'équipe?
- Les données auront-elles besoin d'être traduites? Comment puis-je assurer la confidentialité et l'intégrité des données tout au long du processus?

Certes, les chercheuses et chercheurs en recherche quantitative font face à des défis semblables, mais leurs homologues qualitatifs ont la difficulté supplémentaire d'avoir à recueillir des données personnelles identifiables et potentiellement sensibles, ajoutant ainsi à la complexité du travail sur le terrain.

**Exercice: Recueillir des données loin de chez soi**

Le Dr James Cummings est un sociologue britannique qui a mené une étude ethnographique sur les expériences des homosexuels masculins à Hainan en Chine. Dans un article de journal, il a discuté des défis d'avoir à travailler avec des participants de recherche qui devaient cacher certains aspects de leur vie. Lisez l'article et examinez les expériences du Dr Cummings dans la production et la gestion de ses données de recherche. À quels obstacles a-t-il été confronté? Comment ces obstacles seraient-ils semblables ou différents à ceux d'un chercheur qui mène une étude comparable dans une communauté au Canada? Comment les pratiques de GDR peuvent-elles être améliorées pour mieux appuyer ce type de recherche sur le terrain?

Cummings, J. (2018, 11 mars). The double lives of gay men in China's Hainan province. *The Conversation*. https://theconversation.com/the-double-lives-of-gay-men-in-chinas-hainan-province-153945 (https://theconversation.com/the-double-lives-of-gay-men-in-chinas-hainan-province-153945)

## Les personnes participantes et leurs données

Les données ne sont plus perçues comme un élément sur lequel les personnes participantes ont peu de contrôle. Quelques participantes et participants (p. ex., du personnel clinique ou des fonctionnaires qui prennent part aux recherches dans l'exercice de leurs fonctions) demandent de réviser ou de modifier leurs données avant de consentir à leur utilisation. Ces demandes restent peu fréquentes, ce qui explique pourquoi le suivi des modifications aux données (p. ex., des révisions apportées à une transcription d'entrevue) s'avère compliqué et pourquoi il y a si peu de meilleures pratiques.

Ce lien entre la personne participante et ses données évolue de façon importante. L'idée que l'individu qui participe à une recherche a le droit d'être informé des conclusions résultant de ses données se répand de plus en plus dans les débats éthiques. Des questions sur l'accès aux données qualitatives pour une **analyse secondaire** sont également soulevées. Dans quelle mesure les personnes participantes devraient-elles avoir leur mot à dire sur la manière dont leurs données sont utilisées aujourd'hui ou à l'avenir? Et quelles sont les implications éthiques, juridiques et sociales de ces choix pour les chercheuses et chercheurs?

Une des approches proposées est le consentement dynamique; il permet aux participantes et participants de recherche de maintenir un lien avec leurs données sur une plus longue période et de revenir sur leur décision

de consentir (selon leur souhait). Quand des transcriptions d'entrevues sont déposées dans des dépôts de données (p. ex., Borealis (https://borealisdata.ca/fr/) au Canada, Qualitative Data Repository (https://qdr.syr.edu/) à l'Université Syracuse et UK Data Service (https://ukdataservice.ac.uk/) au Royaume Uni), l'accès est souvent réduit étant donné la nature identifiable du matériel. Le consentement dynamique permet à d'autres chercheuses ou chercheurs de communiquer avec des gens qui ont participé à d'anciennes études pour leur demander la permission de réutiliser leurs données à de nouvelles fins. Les patientes et patients ainsi que les familles qui ont participé à une quelconque étude (p. ex., les maladies rares) cherchent souvent à maximiser l'utilisation de leurs données, dans l'espoir que leurs efforts puissent contribuer à une percée médicale. Bien que le consentement dynamique soit principalement utilisé pour les données quantitatives, le concept sous-jacent reflète des changements plus larges dans la relation entre les personnes qui participent à la recherche et toutes les formes de données.

# Les données de recherche qualitative dans le contexte de la GDR

Cette dernière section nous rappelle la discussion du début : la nécessité de tenir compte des spécificités des données qualitatives dans les principes, les politiques, les stratégies et les pratiques de GDR.

## Le traitement des données d'entrevues

Les données qualitatives n'arrivent pas d'emblée prêtes à être analysées. Un traitement considérable est presque toujours nécessaire et chacune de ces étapes peut créer des versions supplémentaires d'une même donnée principale; les pratiques de GDR peuvent donc être aussi itératives que la recherche qu'elles appuient.

Dans l'exemple suivant, le tableau fait le suivi des modifications d'une seule entrevue, entre le moment où la discussion est enregistrée et celui où les données sont prêtes à être analysées (dans ce cas-ci, elles sont codifiées en utilisant le logiciel NVivo (https://www.qsrinternational.com/nvivo-qualitative-data-analysis-software/home)). Chaque rangée représente la création d'un nouveau fichier.

**Tableau 1. Les itérations d'une entrevue qualitative.**

| Donnée du fichier | Nom du fichier | Observations | Complications |
|---|---|---|---|
| Enregistrement audio (données originales) | CG_CLIN_INT_P14 | Données très identifiables qui ne sont que très rarement partagées avec des gens de l'extérieur de l'équipe de recherche | Les données peuvent être divisées en deux (ou +) fichiers si les entrevues sont très longues ou interrompues; deux enregistrements semblables peuvent exister si un |

| Donnée du fichier | Nom du fichier | Observations | Complications |
|---|---|---|---|
| | | | enregistreur d'appoint est utilisé |
| Transcription — originale (version reçue de la transcription) | CG_CLIN_INT_P14_o | Contient probablement plusieurs erreurs de transcription | Le format peut avoir besoin d'être uniformisé, si différentes personnes effectuent la transcription |
| Transcription — vérifiée (version suivant la comparaison avec l'enregistrement original) | CG_CLIN_INT_P14_v | Le suivi des modifications peut être utile mais peut créer des sous-versions (p. ex., version- suivi, version-acceptée) | Les variations sont plus probables si la même personne ne vérifie pas toutes les entrevues |
| Transcription — modifiée (version suivant les modifications demandées par une personne participante) | CG_CLIN_INT_P14_m | Peut nécessiter des notes explicatives; généralement fait à partir de la version vérifiée des données | Peut obliger une décision sur l'ajout ou non des données si les modifications demandées sont importantes |
| Transcription — anonymisée [Pour plus d'info sur l'anonymisation, consultez le chapitre 13, « Les données sensibles. » ] | CG_CLIN_INT_P14_a | On doit décider si l'anonymisation des entrevues est individuelle ou collective | Les **clés d'anonymisation** peuvent facilement mener à la réidentification d'une personne participante et doivent rester séparées des données |
| Transcription — NVivo (version importée dans le logiciel et révisée davantage) | CG_CLIN_INT_P14_NV | Les révisions dans NVivo ne sont pas capturées dans les versions antérieures | La version demeure dans l'écosystème NVivo à moins d'être téléchargée |

Ce tableau illustre les différentes versions d'une seule transcription. La plupart ne sont que des versions transitoires, bien que cet exemple soit assez sommaire. Plusieurs différents facteurs peuvent ajouter à la complexité du traitement de données d'entrevues. Il s'agit notamment des facteurs suivants :

- les personnes participantes qui sont rencontrées plus d'une fois;
- les entrevues qui doivent être traduites pendant ou après la transcription;
- les transcriptions qui doivent être accompagnées d'autres fichiers de données (p. ex., des notes ou des photos prises sur le terrain, associées à la même personne participante).

**Exercice: Menez une entrevue avec une chercheuse ou un chercheur en recherche qualitative**

L'exercice vous invite à mener une entrevue avec une chercheuse ou un chercheur en recherche qualitative pour en savoir plus sur la façon dont ses données sont gérées. Commencez par identifier un individu qui travaille régulièrement avec des méthodes qualitatives et qui possède de solides connaissances sur le traitement des données qualitatives (ce qui exclut peut-être la communauté étudiante de cycle supérieur). Demandez de voir la structure de ses dossiers (pour des raisons éthiques, vous ne pourrez consulter des données spécifiques). Demandez-lui de vous expliquer quels types de fichiers sont conservés. Examinez l'organisation et la désignation de ses dossiers et fichiers. Posez-lui des questions sur la nature de ce qui a été conservé, l'emplacement de ses données et les raisons de ses choix. En tenant compte de ce que vous avez appris, pouvez-vous proposer des moyens pour améliorer les approches actuelles de la chercheuse ou du chercheur dans la gestion de ses données?

Le traitement des données n'est pas toujours aussi complexe que dans l'exemple de travail ci-haut. Pendant des décennies, les recherches qualitatives ont intégré des approches à la fois simples et efficaces. Cela dit, il y a toujours place à l'amélioration, particulièrement avec l'émergence de nouvelles exigences en GDR. Le savoir ouvert (*open scholarship*) exige des chercheuses et chercheurs que leur gestion des données qualitatives se fasse, dans la mesure du possible, selon des modalités compatibles avec l'excellence en recherche, mais aussi avec la perspective du partage et de la réutilisation des données. Cette transition peut avoir des conséquences sur deux pratiques encore peu courantes en recherche qualitative : les **métadonnées** et l'archivage des données.

Joindre des métadonnées aux données de recherches qualitatives peut être problématique puisque ce type de données nécessite des détails contextuels. Cependant, le contexte peut mener à l'identification des personnes participantes. Comment les données peuvent-elles être adéquatement décrites sans compromettre la confidentialité? Des métadonnées qui indiquent, par exemple, que des données proviennent d'une étude sur les perspectives de cliniciennes et cliniciens qui fournissent des soins de compressothérapie dans une clinique communautaire sont probablement trop simples. Indiquer que les personnes participantes font partie du corps infirmier ayant suivi la même formation spécialisée, que la clinique joue un rôle de premier plan dans le développement d'une approche innovante pour les vêtements de compression et que les patientes et les patients souffrent tous de diabètes de type 2 accroît l'utilité des données. Toutefois, de tels détails augmentent les risques de divulgation et de réidentification. Le problème est moins grand dans le cas de métadonnées

utilisées par des chercheuses ou chercheurs en solo ou à l'intérieur des équipes de recherche lors de la collecte et l'analyse de données primaires. Mais qu'en est-il des métadonnées utilisées pour faciliter l'analyse secondaire par la communauté de recherche externe? Les normes spécifiques aux métadonnées dans le contexte des recherches qualitatives sont difficiles à trouver. En 2023, la question est moins pressante, puisque les données qualitatives sont encore rarement conservées dans des dépôts de données et encore moins accessibles sans restriction.

Plusieurs chercheuses et chercheurs en recherche qualitative hésitent à archiver leurs données et à les ouvrir pour la réutilisation. En plus, les organismes subventionnaires ne les y obligent pas. Le partage des données qualitatives soulève aussi des questions liées au recrutement; la majorité des chercheuses et chercheurs s'engagent auprès de leurs participantes et participants à rendre leurs données accessibles uniquement aux membres de l'équipe de recherche. Ces pratiques sont susceptibles d'évoluer à mesure que les principes d'ouverture des données s'intègrent dans des disciplines plus qualitatives et que les attentes des organismes subventionnaires changent. Nous constatons déjà ce phénomène dans le mouvement de **souveraineté des données autochtones**, ce qui suscite des questions fondamentales sur les métadonnées et leur propriété (pour plus d'informations, consultez le chapitre 3, « Souveraineté des données autochtones » ). De nombreuses préoccupations semblables sont soulevées par et au sujet d'autres groupes identifiables de la société. Qui faut-il consulter, par exemple, quand des décisions de GDR doivent être prises sur la collecte de données auprès de communautés religieuses ou de minorités ethniques? Qui décide de la façon dont ces données seront décrites, archivées, et possiblement réutilisées? Lisez le chapitre 12 « Planification de la gestion des données pour les processus de travail en science ouverte » pour en savoir plus sur les données ouvertes.

Pour terminer, le plus important des défis illustrés avec l'exemple de travail précédent est de déterminer laquelle des versions des données constitue la version définitive. Les enregistrements originaux sont les plus fidèles et descriptifs, mais ils sont plus susceptibles de mener à la réidentification des personnes participantes. Les transcriptions vérifiées et anonymisées semblent donc le choix le plus sûr, mais comment s'assurer du retrait complet de tout **détail identifiant**? Les versions intermédiaires sont-elles conservées et si oui, pendant combien de temps? Si un établissement hôte oblige que les données soient conservées pendant une période de cinq ans suivant la fin de l'étude, est-ce obligatoire pour l'ensemble des versions ou certaines d'entre elles peuvent être supprimées? Ce type de questionnement peut s'appliquer à tout type de données générées en recherche qualitative, rendant ainsi la gestion des données qualitatives extrêmement complexe.

## Un modèle de coproduction de GDR et de données qualitatives

L'exemple de travail vu plus tôt incite à se demander : est-ce que la gestion efficace des données qualitatives est réaliste dans le contexte des principes, des politiques, des stratégies et des pratiques de GDR. L'apparition de

certaines initiatives suggère que oui, notamment les principes de **PCAP®** des Premières Nations, une structure significative d'envergure dont la mise en pratique est en cours et dont les détails sont précisés dans le chapitre 3, « Souveraineté des données autochtones. » Comment atteindre l'objectif d'une gestion efficace des données qualitatives? Règle générale, les chercheuses et chercheurs en recherche qualitative n'ont pas les compétences nécessaires en gestion de l'information pour l'établissement de meilleures pratiques en GDR. Inversement, les bibliothécaires, archivistes et gestionnaires de données sont souvent moins versés sur les complexités des données qualitatives et les processus de recherche qui y sont associés.

En 2020, alors que je travaillais sur une étude portant sur la coproduction dans le domaine de la santé, j'ai assisté à une session de formation en GDR qui, une fois de plus, ne tenait pas compte de mon type de recherche ou de mes préoccupations en matière de gestion des données. Toutefois, je me suis rendu compte à quel point les champs de compétences des chercheuses et chercheurs et des spécialistes des données/de l'information étaient complémentaires. En faisant équipe, un meilleur système de gestion des données qualitatives pourrait être créé.

Pour être efficace, la coproduction doit être hautement collaborative et faire appel à ce qui se fait de mieux dans chacun des domaines. Terminons donc notre discussion sur les avenues potentielles d'une telle coopération :

- Les chercheuses et chercheurs qualitatifs seraient responsables des éléments suivants:
  - s'assurer que les partenaires en GDR comprennent bien les données qualitatives et les processus de recherche;
  - garantir l'uniformité des pratiques de gestion des données à travers les équipes de recherches et maximiser la valeur analytique des données;
  - trouver les fonds nécessaires pour couvrir les coûts du projet associé à la GDR (p. ex., l'embauche d'une personne qualifiée en archivistique ou en recherche);
  - faire la promotion du partage des données dans la culture de recherche;
  - faire appel à leur propre réseau et statut professionnel pour informer les organismes subventionnaires et les établissements de la nature des défis et des coûts inhérents à la gestion de données qualitatives;
- Les bibliothécaires, archivistes et autres spécialistes des données seraient responsables des éléments suivants :
  - appliquer les principes de bibliothéconomie, de sciences des données / de l'information et des meilleures pratiques à la gestion des données qualitatives;
  - soutenir les chercheuses et chercheurs dans la création de jeux de données finaux (avec les métadonnées qui y sont associées) qui répondent ou dépassent les exigences d'excellence en recherche;
  - faire appel à leurs relations professionnelles pour rester au courant et pour faire circuler des

développements en matière de pratiques de GDR qualitatives;

- Les deux groupes seraient conjointement responsables des éléments suivants:
  - établir et promouvoir des normes efficaces en gestion des données qualitatives;
  - développer et dispenser des formations en GDR;
  - faire la promotion d'une plus grande inclusivité de toutes les formes de données de recherche dans les futurs principes et futures politiques, stratégies et pratiques de GDR.

# Conclusion

L'époque actuelle est à la fois excitante et frustrante pour les personnes impliquées dans la gestion de données de recherche qualitatives. Les opportunités d'avancement de nouveaux principes, politiques, stratégies et pratiques ne manquent pas. En même temps, la plupart des chercheuses et chercheurs en recherche qualitative peinent à se retrouver dans les structures actuelles de GDR. Les établissements, les organismes de financement et les spécialistes de GDR doivent s'atteler à trouver des moyens de répondre aux besoins de la communauté de recherche. Même si les données qualitatives ne sont pas particulièrement uniques (après tout, elles sont régulièrement utilisées en collaboration avec d'autres types de données de recherche), elles sont quand même distinctes à plusieurs égards. De telles complexités font ressortir non seulement les limites des approches plus globales de GDR, mais aussi la nécessité d'élargir la gestion des données pour mieux intégrer toutes les disciplines, les domaines de recherche et les méthodes d'enquête.

**Questions de réflexion**

1. Identifiez au moins trois caractéristiques importantes des données qualitatives.
2. En plus des entrevues, des groupes de discussion et des observations, nommez deux autres formes de données qualitatives.
3. Quel est le but de la piste de vérification?
4. Identifiez deux défis liés aux données qu'une chercheuse ou un chercheur en recherche qualitative peut rencontrer sur le terrain dans un lieu éloigné.
5. Les données d'entrevues existent généralement en multiples versions entre la collecte et l'analyse. Identifiez deux de ces versions.
6. En une phrase, décrivez l'objectif d'un modèle de coproduction pour la GDR qualitative.

Voir le solutionnaire pour les réponses.

*Un élément interactif H5P a été exclu de cette version du texte. Vous pouvez le consulter en ligne ici : https://ecampusontario.pressbooks.pub/gdrcanada/?p=60#h5p-6 (https://ecampusontario.pressbooks.pub/gdrcanada/?p=60#h5p-6)*

**Éléments clés à retenir**

- Les données générées par le biais de recherche qualitative sont complexes parce qu'elles traitent de l'être humain, sont itératives, ancrées dans le contexte et très difficiles à dépersonnaliser.
- De telles données sont difficiles à placer dans le contexte actuel des principes, des politiques, des stratégies et des pratiques de la GDR.
- Une bonne gestion des données de recherche qualitatives doit comprendre et tenir compte des processus de recherche en jeu, y compris l'évolution des attentes en matière d'archivage et de réutilisation des données, ainsi que le transfert des responsabilités des participantes et participants de recherche.
- Ensemble, les chercheuses et chercheurs et les spécialistes des données/de l'information sont bien placés pour coproduire de nouvelles approches en GDR qui répondent mieux aux besoins des données des chercheuses et chercheurs en recherche qualitative .

# Remerciements

Le Dr Minion remercie sincèrement la Dre Naomi Adelson et la Dre Tamara McCarron pour leurs précieux commentaires sur les versions antérieures de ce chapitre.

# Lectures et ressources supplémentaires

Adelson, N. et Mickelson, S. (2022). The Miiyupimatisiiun research data archives project: Putting OCAP® principles into practice. *Digital Library Perspectives, 38*(4), 508-520.

Budin-Ljøsne, I., Teare, H. J. A., Kaye, J., Beck, S., Bentzen, H. B., Caenazzo, L., Collett, C., D'Abramo, F., Felzmann, H., Finlay, T., Javaid, M.K., Jones, E., Katić, V., Simpson, A. et Mascalzoni, D. (2017). Dynamic consent: A potential solution to some of the challenges of modern biomedical research. *BMC Medical Ethics, 18*(1), 1-10. https://doi.org/10.1186/s12910-016-0162-9 (https://doi.org/10.1186/s12910-016-0162-9)

Chauvette, A., Schick-Makaroff, K. et Molzahn, A. E. (2019). Open data in qualitative research. *International Journal of Qualitative Methods, 18*, 1-6. https://doi.org/10.1177/160940691882386 (https://doi.org/10.1177/1609406918823863)

Corti, L. (2019). Archiving qualitative data. Dans P. Atkinson, S. Delamont, A. Cernat, J.W. Sakshaug et R.A. Williams (dir.), *SAGE Research Methods Foundations*. SAGE Publications. https://dx.doi.org/10.4135/9781526421036813114 (https://dx.doi.org/10.4135/9781526421036813114)

Cummings, J. (2018, 11 mars). The double lives of gay men in China's Hainan province. *The Conversation*. https://theconversation.com/the-double-lives-of-gay-men-in-chinas-hainan-province-153945 (https://theconversation.com/the-double-lives-of-gay-men-in-chinas-hainan-province-153945)

Diaz, P. (2021). Introduction: Archiving qualitative data in practice: Ethical feedback. *Bulletin of Sociological Methodology, 150*(1), 7-27. https://doi.org/10.1177/0759106321995678 (https://doi.org/10.1177/0759106321995678)

DuBois, J. M., Strait, M. et Walsh, H. (2018). Is it time to share qualitative research data? *Qualitative Psychology, 5*(3), 380-393. https://doi.org/10.1037/qup0000076 (https://doi.org/10.1037/qup0000076)

Mannheimer, S., Pienta, A., Kirilova, D., Elman C. et Wutich, A. (2019). Qualitative data sharing: Data repositories and academic libraries as key partners in addressing challenges. *American Behavioral Scientist, 63*(5): 643-664. https://doi.org/10.1177/0002764218784991 (https://doi.org/10.1177/0002764218784991)

Moon, K. et Blackman, D. (2014). A guide to understanding social science research for natural scientists. *Conservation Biology, 28*(5), 1167-1177. https://doi.org/10.1111/cobi.12326 (https://doi.org/10.1111/cobi.12326)

Pels, P., Boog, I., Florusbosch, J. H., Kripe, Z., Minter, T., Potsma, M., Sleeboom-Faulkner, M., Simpson, B., Dilger, H., Schönhuth, M., Poser, A., Castillo, R. C. A., Lederman, R. et Richards-Rissetoo, H. (2018). Data

management in anthropology: The next phase in ethics governance? *Social Anthropology, 26*(3), 391-396. https://doi.org/10.1111/1469-8676.12526 (https://doi.org/10.1111/1469-8676.12526)

Saunders, B., Kitzinger, J. et Kitzinger, C. (2015). Anonymising interview data: Challenges and compromise in practice. *Qualitative Research, 15*(5), 616-632. https://doi.org/10.1177/1468794114550439 (https://doi.org/10.1177/1468794114550439)

Steinhardt, I., Fischer, C., Heimstädt, M., Hirsbrunner, S. D., Ikiz-Akinci, D., Kressin, L., Kretzer, S., Möllekamp, A., Porzelt, M., Rahal, R., Schimmler, S., Wilke, R. et Wünsche, H. (2021). *Opening up and sharing data from qualitative research: A primer*. https://www.ssoar.info/ssoar/bitstream/handle/document/74039/ssoar-2021-steinhardt_et_al-Opening_up_and_Sharing_Data.pdf (https://www.ssoar.info/ssoar/bitstream/handle/document/74039/ssoar-2021-steinhardt_et_al-Opening_up_and_Sharing_Data.pdf)

Suter, W. N. (2012). Qualitative data, analysis, and design. *Introduction to Educational Research, 2*, 342-386.

Van den Eynden, V. et Chatsiou, K. (2011). *Data management for qualitative data: Using NVivo 9*. https://dam.ukdataservice.ac.uk/media/622387/ukda-datamanagement-nvivo.pdf (https://dam.ukdataservice.ac.uk/media/622387/ukda-datamanagement-nvivo.pdf)

Wang, C. et Burris, M. A. (1997). Photovoice: Concept, methodology, and use for participatory needs assessment. *Health Education & Behavior, 24*(3), 369-387. https://doi.org/ 10.1177/109019819702400309 (https://doi.org/10.1177/109019819702400309).

## À propos de l'auteur

### Dr. Joel T. Minion

Joel T. Minion, PhD MLIS MA BA (spécialisé) est un chercheur qualitatif en santé, un bibliothécaire, un gestionnaire de données et un éducateur ayant de l'expérience en gestion de données de recherche (GDR) au Canada et en Europe. Il est actuellement chercheur scientifique au sein du programme Translating Research in Elder Care (TREC) de la Faculty of Nursing de l'Université de l'Alberta où il est responsable de la planification de l'héritage des données longitudinales du TREC. Joel était auparavant responsable de la recherche qualitative pour la Health Technology Assessment Unit de l'Université de Calgary au sein du O'Brien Institute for Public Health et avant cela, il était associé principal de recherche au centre de recherche PEALS (Policy, Ethics and Life Sciences) de l'Université Newcastle au Royaume-Uni. Il est titulaire d'un doctorat en informatique de la santé de l'Université de Sheffield et d'un diplôme MLIS de l'Université Western. Depuis 2010, Joel est activement impliqué dans la gestion des données de recherche qualitative et dans les efforts en cours pour les intégrer dans des cadres de GDR plus larges.

15.

# LA GESTION DES DONNÉES QUANTITATIVES EN SCIENCES SOCIALES

Dr. Alisa Beth Rod et Dr. Biru Zhou

**Objectifs d'apprentissage**

À la fin de ce chapitre, vous pourrez :

1. Définir les différents types de données quantitatives en sciences sociales.
2. Décrire les différentes façons de mettre en œuvre les pratiques de gestion des données de recherche dans le cadre de travaux en sciences sociales avec des données quantitatives.
3. Comprendre comment les pratiques de gestion des données de recherche peuvent contribuer à atténuer la crise de la reproductibilité et à faciliter le dépôt de données quantitatives en sciences sociales en vue de leur réutilisation.

## Introduction

La première étape dans la gestion des **données de recherche** quantitatives en **sciences sociales** est de revoir la conception générale de la recherche et d'identifier où les pratiques de gestion des données de recherche (GDR) peuvent être appliquées pour faciliter la recherche et renforcer les résultats de la recherche. La plupart des recherches quantitatives en sciences sociales suivent la conception des études scientifiques. Ces conceptions aident les chercheuses et chercheurs à générer des questions, à formuler des hypothèses et des prédictions concrètes, à concevoir le projet de recherche, à recueillir et analyser les données de recherche et à documenter les résultats pour communiquer les conclusions au public. Pour contextualiser la GDR dans le monde de la recherche quantitative en sciences sociales, il est important de bien comprendre le processus et le flux de travail de ce genre de projets de recherche. La prochaine section donnera un aperçu des études

quantitatives en sciences sociales, ce qui servira de contexte pour les sections suivantes qui traiteront de la gestion des données quantitatives en sciences sociales.

## Aperçu des recherches quantitatives en sciences sociales

Les chercheuses et chercheurs qui font de la recherche quantitative en sciences sociales utilisent deux approches générales fondamentales qui peuvent avoir des incidences sur la collecte et la gestion des données. Une de ces approches est la **conception descriptive** qui vise à explorer un phénomène ou une observation pour décrire un effet (de Vaus, 2001). Les recherches descriptives courantes comprennent des études menées par des gouvernements (p. ex., les niveaux de revenu des ménages, l'utilisation des bibliothèques municipales, les plaintes liées au bruit, le trafic à proximité des centres urbains au fil du temps, etc.). L'objectif des recherches descriptives est de décrire des phénomènes sociaux, économiques ou politiques sans s'attarder sur les causes de ces phénomènes. Des questions de recherche qui utilisent la conception descriptive peuvent comprendre :

- Quel est le niveau de pauvreté des communautés rurales?
- Le niveau d'inégalité sociale est-il en croissance ou en déclin à travers Montréal?
- À Toronto, à quel endroit les gens sont-ils plus susceptibles d'être arrêtés et reconnus coupables d'une infraction?
- Qui est le plus susceptible d'être arrêté et reconnu coupable d'une infraction en Alberta?

Une autre approche utilisée par les chercheuses et chercheurs pour étudier des phénomènes sociaux est la **conception explicative** qui vise à expliquer un phénomène ou une observation afin de comprendre un effet (de Vaus, 2001). Les études explicatives se préoccupent de comprendre la ou les cause(s) d'un phénomène social, économique ou politique. Les études explicatives sont la suite logique des recherches descriptives établies. Par exemple, si une étude descriptive établit que le taux d'expulsion d'un quartier particulier est nettement plus élevé que celui de tous les autres quartiers, on pourrait vouloir mener une étude explicative pour mieux comprendre les raisons ou les causes de cet écart. Des questions de recherches qui utilisent la conception explicative peuvent comprendre :

- Pourquoi le taux d'expulsion de la Ville Y est-il plus élevé que toute autre ville canadienne?
- Pourquoi les autobus scolaires de la Communauté Z sont-ils considérablement en retard?
- Pourquoi le taux de pauvreté de la Communauté X est-il le plus élevé du Manitoba?

Peu importe l'approche utilisée dans l'étude, la première étape du processus de recherche est de formuler une

question ou un ensemble de questions de recherche. La question de recherche exprime l'objectif de l'étude sous forme interrogative. La liste suivante énumère certains exemples de structure pour une question de recherche potentielle (avec x, y et z qui remplacent les concepts) :

- Quel est le lien entre x et y?
- De quelle façon l'emplacement de x a-t-il un effet sur y?
- Quels facteurs structurels ou démographiques peuvent prédire x, y et z?
- Pourquoi x a-t-il un effet sur y?

Voici quelques exemples de ces mêmes questions en y intégrant des concepts sociaux réels :

- Quel est le lien entre la pauvreté et l'éducation?
- De quelle façon l'emplacement des bibliothèques municipales a-t-il un effet sur la cohésion d'une communauté?
- Quels facteurs structurels ou démographiques peuvent prédire le chômage, l'insécurité économique et les besoins en logements sociaux?
- Pourquoi la personnalité a-t-elle un effet sur la susceptibilité à l'effet de cadrage?

La question de recherche encadrera les étapes suivantes dans la conception et dans la mise en œuvre d'une étude quantitative en sciences sociales, comme indiqué dans le tableau ci-dessous. Cliquez sur les titres pour explorer les différentes étapes d'un processus typique de recherche quantitative en sciences sociales :

## Processus de recherche quantitative en sciences sociales

*Un élément interactif H5P a été exclu de cette version du texte. Vous pouvez le consulter en ligne ici : https://ecampusontario.pressbooks.pub/gdrcanada/?p=62#h5p-1 (https://ecampusontario.pressbooks.pub/gdrcanada/?p=62#h5p-1)*

De bonnes pratiques de GDR sont pertinentes pour toutes les étapes d'un projet typique de recherche quantitative en sciences sociales, de la planification jusqu'à la publication des résultats de la recherche. Les **plans de gestion des données** sont des outils importants qui aident les chercheuses et chercheurs à réfléchir au traitement de leurs données de recherche tout au long des différentes étapes du processus de recherche.

Dans la section suivante, nous partagerons les considérations de GDR qui s'appliquent particulièrement au travail avec les données quantitatives en sciences sociales.

## La gestion des données de recherche quantitatives en sciences sociales : fichiers, formats et documentation

Les données quantitatives en sciences sociales ne comportent pas de différences inhérentes aux autres types de données quantitatives sauf en ce qui concerne leur(s) source(s) et l'objet des données. Les données quantitatives correspondent à des données numériques qui sont mesurées selon une échelle d'**intervalles** ou de **rapport**, ou des **variables catégoriques** qui sont **codées avec des variables factices** ou converties selon une échelle ordinale. La méthode la plus courante de collecte de données quantitatives originales en sciences sociales se fait par le biais d'outils de sondage.

> Les bonnes pratiques de GDR en sciences sociales nécessitent que les chercheuses et chercheurs documentent l'entièreté du processus d'enquête. Quand viendra le temps de partager ou d'archiver les données du sondage, vous pourrez joindre le jeu de données final aux questions du sondage et aux informations sur les personnes participantes et sur la façon dont la recherche a été menée.

Un outil de sondage, ou questionnaire se rapporte à une série de questions posées à une certaine population. Le but est de mesurer un ou plusieurs concepts. Un questionnaire de sondage peut inclure des items ou des questions qui **opérationnalisent** les différents concepts – c'est-à-dire qui transforment des concepts abstraits en variables et en indicateurs mesurables et quantifiables.

En plus des données de sondage, plusieurs chercheuses et chercheurs en sciences sociales dépendent de données administratives. Les **données administratives** sont des données recueillies par des organismes ou des agences gouvernementales à des fins administratives (c'est-à-dire, qui n'ont pas d'objectifs de recherche, mais qui visent plutôt à administrer ou à évaluer des services, des produits ou des biens). Des exemples de données administratives peuvent comprendre des statistiques d'état civil (p. ex., les taux de natalité et de mortalité), des dossiers de ressources humaines, des renseignements fiscaux personnels ou municipaux, des budgets, des emplacements de services publics et des bénéficiaires de programmes de services sociaux. Il est important de noter que les données administratives qui ne sont pas accessibles au public sont généralement

régies par des licences ou des contrats qui peuvent avoir un impact sur le partage et/ou le dépôt des données. Cette question a été examinée plus en détail au chapitre 13, « Les données sensibles. »

Dans votre pratique de GDR, vous devez tenir compte des licences sur les jeux de données lorsque vous planifiez la manière dont ils pourraient être partagés ou déposés à la fin du projet. Par exemple, certains contrats ou licences peuvent déterminer si, à la fin de la recherche, le jeu de données utilisé peut être partagé pour un processus d'évaluation par les pairs à des fins de vérifications des conclusions ou s'il peut être déposé à des fins de réutilisation par d'autres chercheuses ou chercheurs. Rappelez-vous ce que vous avez appris au sujet des licences et du partage des données dans le chapitre 12, « Planification de la gestion des données pour les processus de travail en science ouverte. »

Que les données soient issues d'enquêtes originales ou de sources administratives, les spécialistes en sciences sociales quantitatives recueillent et entreposent leurs données, la plupart du temps, dans un **format tabulaire**.

> Songer aux formats de préservation pour vos fichiers ou à leur durabilité dans le temps est une bonne pratique de GDR. Les formats courants pour la préservation des fichiers en format tabulaire sont les CSV ou TAB, qui sont tous deux des formats ouverts qui ne dépendent pas de logiciels propriétaires et qui sont accessibles par le biais d'une variété de programmes différents (p. ex., Stat, SAS, SPSS, Excel). Le stockage des données dans des **formats non propriétaires** ou la création d'une sauvegarde de toutes les données dans un de ces formats constitue une bonne pratique de GDR qui permet d'assurer la durabilité et l'**interopérabilité** de vos données pour une utilisation future. (Pour en savoir plus sur les formats, voir le chapitre 9, « Un aperçu du fascinant monde des formats de fichiers et des métadonnées. ») Toutefois, les chercheuses et chercheurs utilisent souvent Microsoft Excel pour recueillir et stocker leurs données tabulaires. Étant donné l'omniprésence de ce logiciel dans le paysage de recherche et de l'industrie, son utilisation n'est généralement pas problématique pour la réutilisation éventuelle des données. Le guide du Data Curation Network sur la curation des données de Microsoft Excel (https://github.com/DataCurationNetwork/data-primers/blob/master/Excel%20Data%20Curation%20Primer/Excel%20Data%20Curation%20Primer.md) est une ressource utile (en anglais uniquement).

En règle générale, les données tabulaires sont organisées de telle façon que chaque rangée représente une observation (p. ex., un participant, un quartier, un immeuble, une année) et chaque colonne représente une

variable (p. ex., l'information qui varie à travers les différentes observations). Nous discuterons de formats alternatifs de données tabulaires (p. ex., long versus large) dans la section suivante.

Il existe plusieurs bonnes pratiques en lien avec l'organisation d'un jeu de données tabulaires. L'une d'entre elles est l'élimination des espaces dans les variables, les fichiers et les noms des observations puisque les ordinateurs peinent à interpréter les espaces vides lors de l'automatisation des tâches. Une autre bonne pratique en matière d'attribution de noms est de limiter la longueur du nom des variables; l'utilisation de huit caractères ou moins empêche que le nom soit coupé ou abrégé par les logiciels d'analyse des données. Définir ainsi le nom des variables, vous permettra d'améliorer l'interopérabilité et la réutilisation éventuelle des données par d'autres logiciels.

Il arrive souvent que le nettoyage des données soit nécessaire avant d'analyser, partager ou déposer des données; ce sujet est abordé dans les chapitres 7 (« Le nettoyage de données dans le processus de gestion des données de recherche » ) et 8 (« Nouvelles aventures en nettoyage des données » ). En nettoyant vos données, vous allez aussi vouloir créer une documentation pour les accompagner, y compris une version codée des noms de vos variables et/ou de vos observations ainsi qu'un **guide de codification** connexe dans un document distinct. Les espaces dans les noms de fichiers ou dans les en-têtes de tableaux peuvent causer certains logiciels ou applications à planter ou peuvent entraîner des erreurs lors de l'ouverture ou de l'utilisation des fichiers. Par exemple, dans un environnement de ligne de commandes, les espaces sont utilisés comme **séparateurs**. Pour éviter les espaces vides, utilisez la **notation chameau** (ChaqueMotCommenceParUneMajuscule) ou le trait de soulignement (entre_les_mots) afin de créer des codes qui peuvent être lus par ordinateur.

Prenons par exemple le cas d'une chercheuse qui mène une enquête auprès de la communauté étudiante au premier cycle pour en savoir plus sur les coûts associés au matériel de cours. Le sondage comprend un des items suivants : « Au cours du semestre précédent, étiez-vous inscrit à des cours qui impliquaient des dépenses liées à des déplacements à l'intérieur de la grande région de Calgary? » Ce ne serait pas utile d'inscrire intégralement cette question dans la colonne du tableau. La chercheuse peut donc créer une version codée ou abrégée telle que « FraisTransport » qui remplacera la question intégrale dans l'en-tête de colonne et le nom de la variable dans le jeu de données. Pour garder la trace de ces remplacements ou codes, la meilleure pratique prône la création d'un guide de codification sous forme de document textuel distinct qui fait le lien entre les codes abrégés et les questions originales complètes du questionnaire.

En plus de faire le lien entre les codes et les noms complets des variables ou les items de questionnaire, un guide de codification peut aussi contenir des informations sur les données manquantes et les étiquettes ou valeurs de l'étendue des réponses pour une question particulière. Par exemple, si les réponses potentielles d'une question sont « oui », « non » et « je ne sais pas », la chercheuse pourrait utiliser des codes numériques avec des étiquettes de valeurs pour faire une analyse quantitative des réponses. Le guide de

codification peut contenir ces informations en indiquant que « oui » est codé comme 3, « non » est codé comme 2 et « je ne sais pas » comme 1.

Le tableau suivant représente un exemple des informations pouvant se retrouver dans le guide de codification de ce sondage :

**Tableau 1: Exemple d'un guide de codification associé à un sondage**

| **Code de la variable** | **Étiquette de la variable (Question originale)** | **Options de réponse** |
|---|---|---|
| FraisTransport | Au cours du semestre précédent, étiez-vous inscrit à des cours qui impliquaient des dépenses liées à des déplacements à l'intérieur de la grande région de Calgary? | 3 = oui<br>2 = non<br>1 = je ne sais pas |
| FraisManuels | Au cours du semestre précédent, étiez-vous inscrit à des cours qui impliquaient des dépenses liées à l'achat de manuels scolaires? | 3 = oui<br>2 = non<br>1 = je ne sais pas |
| Soucis | Vous est-il arrivé d'exprimer à un professeur des soucis en lien avec votre capacité à assumer les coûts du matériel requis pour leur cours? | 3 = oui<br>2 = non<br>1 = je préfère ne pas répondre |

Lorsque plusieurs variables ont les mêmes options de réponses – telles que « FraisTransport » et « FraisManuels » dans l'exemple ci-dessus – il est important d'uniformiser les valeurs pour les choix de réponses de ces variables afin d'éviter toute confusion au cours de la phase d'analyse du projet.

Il n'est pas rare que des laboratoires ou des équipes de recherche mènent simultanément plusieurs projets de recherches sur des sujets semblables en utilisant des mesures semblables. Prenons comme exemple deux études semblables qui sont menées en parallèle sur l'impact de la violence en milieu de travail sur des membres du personnel atteints de symptômes du trouble de stress post-traumatique (TSPT). L'une des recherches peut explorer l'intimidation en milieu de travail comme cause des symptômes du TSPT chez les membres du personnel et l'autre peut se pencher sur la violence physique de la part de la clientèle comme cause des symptômes du TSPT chez les membres du personnel. Dans ce cas-ci, les deux études mesurent les symptômes du TSPT. Pour améliorer l'interopérabilité à l'intérieur de l'équipe de recherche, il est important de maintenir à travers les deux études une uniformité dans l'attribution des noms et des codes des mesures du TSPT. Le guide de codification – qui fait partie de la documentation associée au jeu de données, idéalement accompagné d'un fichier **LISEZ-MOI** et/ou de **métadonnées** – est essentiel quand une chercheuse ou un chercheur prévoit de partager ou déposer son jeu de données auprès de la communauté de recherche ou de l'ouvrir au public. Il serait impossible d'utiliser le jeu de données sans connaître les définitions de chacune des variables (pour de plus amples exemples, consultez la ressource *What is a Codebook (https://www.icpsr.umich.e*

*du/web/ICPSR/cms/1983#:~:text=A%20codebook%20describes%20the%20contents,layout%20of%20a%20data%20collection)* du Inter-university Consortium for Political and Social Research (ICPSR), qui donne une description sommaire d'un guide de codification et présente quelques exemples de structures typiques).

Nommer les variables et les fichiers ainsi que définir des versions quantitatives de constructions sociales ou comportementales abstraites peut être complexe. Un des éléments clés dans la GDR pour les disciplines quantitatives, dont les sciences sociales, implique l'établissement de conventions pour le nommage des fichiers et la hiérarchie des répertoires de fichiers en utilisant un plan de gestion des données (PGD). Un PGD est un important outil de gestion de projet pour la documentation des conventions de nommage de fichiers, surtout lorsqu'il est question de données quantitatives qui peuvent comporter plusieurs versions différentes d'un jeu de données en format tabulaire avec des fichiers de code ou des scripts qui peuvent être nécessaires au nettoyage ou à l'analyse des jeux de données.

Les conventions de nommage des fichiers en sciences sociales quantitatives ne diffèrent pas forcément de celles des autres disciplines. Il est nécessaire d'y inclure suffisamment d'informations pour permettre d'identifier un fichier de façon précise et de bien distinguer les différentes versions d'un même jeu de données. Par exemple, il pourrait être important d'inclure « *raw* » (ou « brut ») dans le nom d'un fichier qui contient des données recueillies avant le nettoyage ou l'analyse. Une bonne pratique implique la création d'une copie du fichier de données brutes, avant toute intervention sur les données, comme fichier de travail et conserver celle-ci en tant que version authentique des données. La copie de travail du fichier de données devrait porter un nom qui la distingue clairement du fichier de données brutes, en plus des autres versions potentielles du jeu de données (p. ex., une version du jeu de données nettoyé ou une version du jeu de données nettoyé qui intègre des variables calculées à partir des données brutes). Au fil d'un projet, plusieurs fichiers peuvent être créés pour un même jeu de données. Un PGD peut être utilisé pour prévoir la création des différents types de fichiers et l'attribution de noms uniques qui permet à ces fichiers d'être bien identifiés. L'ICPSR (https://www.icpsr.umich.edu/web/pages/about/) basé à l'Université du Michigan aux États-Unis, généralement considéré comme étant le dépôt de données en sciences sociales le plus connu, a produit un modèle de PGD (https://www.icpsr.umich.edu/web/pages/datamanagement/dmp/plan.html) (en anglais uniquement) pour les sciences sociales qui comporte des conseils associés aux types de données que les chercheuses et chercheurs en sciences sociales ont généralement à recueillir et à gérer.

Il y a des considérations supplémentaires à prendre en compte dans la gestion de projets quantitatifs en sciences sociales lorsqu'il s'agit d'**études longitudinales**. L'enquête longitudinale est une méthode courante en sciences sociales où les chercheuses et chercheurs vont recueillir ou comparer des données des mêmes personnes participantes sur une période de plusieurs années. Les défis de ce genre d'enquête sont au niveau de la fusion des données d'une personne participante à une période particulière avec celles de la même personne à une autre période ainsi que de la préservation de l'intégrité de ces données sur toute la période de l'enquête et à travers les différentes itérations des jeux de données. Pour ajouter à la complexité, certaines personnes

participantes peuvent délaisser l'enquête au fil du temps – il y aura un certain degré d'abandon et par conséquent des variations dans le nombre de personnes participantes au fil des années.

La GDR comprend les pratiques liées à l'établissement d'un flux de travail ou d'un processus permettant de suivre la façon dont les fichiers sont fusionnés et les changements entre les différentes versions d'un jeu de données. La GDR concerne également les décisions quant au choix des versions du fichier à partager ou à déposer à long terme. Les chercheuses et chercheurs devraient-ils déposer chacune des vagues (p. ex., chaque jeu de données pour une période de temps particulière) comme un jeu de données distinct avec des instructions sur la façon de fusionner les fichiers? Ou devraient-ils partager un seul jeu de données fusionnées qui comprend plusieurs années? Il n'y a pas de bonnes ou de mauvaises réponses à ces questions. La GDR permet d'assurer la prise d'une décision, quelle qu'elle soit, idéalement en fonction de la version du jeu de données nécessaire à la reproductibilité des conclusions publiées ou des normes générales de la discipline, pourvu que la documentation soit recueillie et rendue accessible selon l'option choisie par les chercheuses ou chercheurs.

## Des enjeux de GDR associés aux outils et logiciels numériques pour la collecte de données quantitatives en sciences sociales

La recherche par sondage est une méthode courante et peu coûteuse qui est utilisée pour les recherches aussi bien qualitatives que quantitatives en sciences sociales. La plupart des sondages sont de nature non expérimentale. Ils sont utilisés pour décrire et évaluer la prévalence d'un phénomène et/ou pour identifier des liens particuliers entre différents facteurs.

Les renseignements recueillis en sciences sociales par le biais de sondages en ligne peuvent être de nature sensible et contenir des renseignements personnels (p. ex., l'âge, le genre, l'ethnicité, l'adresse courriel, l'adresse IP) et/ou des renseignements personnels de santé (p. ex., des diagnostics antérieurs autodéclarés de problèmes de santé). Comme établi par l'*Énoncé de politique des trois conseils sur l'éthique de la recherche avec des êtres humains* (EPTC 2), (https://ethics.gc.ca/fra/policy-politique_tcps2-eptc2_2022.html) chaque chercheuse et chercheur a le devoir de protéger ses données de recherche et les renseignements des personnes participantes contre les accès non autorisés et illégaux. À cet effet, l'établissement du niveau de sensibilité des données de recherche et des options qui s'imposent pour le stockage, la collecte et l'analyse des données actives constitue un autre aspect clé de la GDR pour tout projet qui implique des êtres humains. Pour en savoir plus, consultez le chapitre 13, « Les données sensibles. »

Par contre, la plupart d'entre nous ne sont pas des spécialistes en cybersécurité. Il est excessivement difficile de vérifier si un fournisseur agit conformément aux lois et règlements en vigueur, s'il dispose de mesures de

contrôle de sécurité externes certifiées ou si les données sont chiffrées alors qu'elles sont en transit et au repos. Lorsque possible, l'utilisation d'outils de sondage approuvés par l'établissement ou sous licence institutionnelle peut épargner aux chercheuses et chercheurs de nombreux maux de tête causés par la conformité avec les politiques institutionnelles ou gouvernementales de cybersécurité. En préparant le PGD pour un projet quantitatif en sciences sociales, vous avez la chance de décrire les méthodes de collecte des données ainsi que les outils ou logiciels que vous prévoyez utiliser dans le processus. Il s'agit d'un aspect important dans l'étape de la planification et celui-ci vient confirmer l'utilité d'un PGD dans le contexte de la recherche quantitative en sciences sociales.

Par exemple, si vous aviez à utiliser un outil de sondage en ligne géré par un tiers externe (probablement un service **infonuagique**), il est important d'enquêter afin de déterminer l'emplacement physique du serveur principal et des serveurs des sous-traitants. Bien que plusieurs outils de sondage infonuagiques soient fiables et sécuritaires, les pratiques de leurs sous-traitants ou leur emplacement physique (p. ex., si le serveur est situé à l'extérieur du Canada) pourraient mettre à risque la sécurité de vos données en raison de leur non-conformité aux lois et règlements canadiens en matière de protection de la vie privée. Si le serveur qui héberge la plateforme de sondage en ligne se trouve aux États-Unis, les données qui y sont stockées sont assujetties au Patriot Act des États-Unis. De plus, certaines ententes de financement particulières pourraient empêcher le stockage des données de recherche à l'extérieur du Canada. Ce genre de considération peut être examiné et résolu d'avance en utilisant un PGD.

## La curation des données quantitatives en sciences sociales à des fins de reproductibilité

La dernière étape d'un projet de recherche typique en sciences sociales quantitatives implique la prise de décisions en lien avec le dépôt (c'est-à-dire, la publication) et/ou l'archivage des données qui sous-tendent les publications issues de l'étude. Les normes disciplinaires en sciences sociales associées au partage ouvert des données peuvent varier selon les disciplines ou les champs d'études particuliers, mais le partage devient tranquillement une pratique courante. De plus, les bailleurs de fonds comme les **trois organismes subventionnaires** du Canada et les publications savantes d'une variété de domaines de sciences sociales exigent de plus en plus l'accès ou le dépôt des données de recherche dans un dépôt public. Toutefois, le moteur derrière cette pression à publier les données de recherche, y compris toute la documentation ou les métadonnées associées, est la crise de la reproductibilité (Turkyilmaz-van der Velden *et al.*, 2020).

La crise de la reproductibilité se rapporte à l'incapacité des chercheuses et chercheurs à répéter ou à reproduire les conclusions des recherches publiées. La répétition est une méthode clé pour assurer la validité ou l'intégrité des conclusions de recherche. Dans la plupart des cas, la raison pour laquelle une étude ne peut être vérifiée par répétition est attribuable à un problème avec les données originales, l'indisponibilité des données ou le

manque de détails dans la description des étapes entreprises pour l'analyse des données, ce qui empêche de produire les mêmes résultats (Baker, 2016). Les sciences sociales quantitatives n'ont pas été épargnées par la crise de la reproductibilité et plusieurs rétractations très médiatisées, en raison de problèmes ou de fraudes liés aux données sous-jacentes de la publication, ont unifié les efforts qui visent à soutenir une plus grande transparence dans les pratiques d'accès aux données (Figueiredo *et al.*, 2019). Par exemple, deux politologues ont mené une étude charnière, du moins en apparence, sur les convictions politiques qui a ensuite été publiée en 2015 dans la revue Science. Toutefois, au cours des cinq mois suivants, deux étudiants de troisième cycle ayant fait une demande d'accès aux données à des fins de répétition, ont découvert des preuves de fraude délibérée; la publication a ensuite été rétractée (Konnikova, 2015). Retraction Watch (https://retractionwatch.com/2010/08/03/why-write-a-blog-about-retractions/) et PubPeer (https://pubpeer.com/static/about) sont deux sites Web populaires qui, par production participative, font le suivi des rétractations et des préoccupations relatives aux données sous-jacentes des recherches publiées dans les publications savantes. Ainsi, la communauté savante se responsabilise pour produire des recherches qui peuvent être reproduites.

En plus du ICPSR, les chercheuses et chercheurs en sciences sociales quantitatives peuvent faire appel à plusieurs dépôts de données publiques pour publier leurs données. Ces dépôts répondent aux normes disciplinaires en matière de transparence et de reproductibilité des recherches, ainsi qu'aux mandats des organismes subventionnaires et des publications savantes qui exigent que les données de recherche soient Faciles à trouver, Accessibles, Interopérables et Réutilisables (https://www.frdr-dfdr.ca/docs/fr/principes_fair/) (FAIR). Une instance Borealis (https://borealisdata.ca/fr/), basée sur le **logiciel ouvert** Dataverse, est disponible dans la plupart des établissements canadiens en tant que dépôt institutionnel des données. Cette offre s'insère dans un cadre plus large de ressources d'infrastructures en gestion des données de recherche fournies par consortium. Les chercheuses et chercheurs affiliés à ces établissements peuvent déposer leurs jeux de données dans leur dépôt institutionnel de Dataverse. Bien que la plateforme de dépôt soit ouverte à toutes disciplines, le logiciel Dataverse a d'abord été développé pour les données quantitatives en sciences sociales, il est donc bien adapté aux types de fichiers couramment produits par les chercheuses et chercheurs en sciences sociales quantitatives, notamment les petits fichiers tabulaires et les fichiers script qui y sont associés.

Déposer ses données dans un dépôt public représente une étape vers un plus grand accès aux données, mais la pratique ne suffit pas pour assurer la reproductibilité d'une étude ou le respect des principes FAIR. Des étapes de curation supplémentaires devraient être prises, généralement par des bibliothécaires ou d'autres professionnelles ou professionnels de l'information qui peuvent agir comme intermédiaires dans le dépôt des données en convertissant les fichiers d'un format propriétaire – tels que les fichiers SPSS ou STATA – à un format ouvert – tels que R ou CSV. De plus, une documentation est nécessaire pour permettre la réutilisation d'un jeu de données quantitatives ou la reproductibilité des résultats. La documentation d'un jeu de données quantitatives en sciences sociales peut inclure la description de l'étude, le guide de codification, les métadonnées sur la collecte des données (p. ex., les systèmes de pondération utilisés pour les données de sondage, les périodes de collecte des données, les logiciels utilisés pour recueillir et analyser les données, etc.),

les scripts ou codes nécessaires pour nettoyer les données ou reproduire les éléments relatifs à une publication, ainsi que la licence pour la réutilisation ou les conditions d'utilisation des données. Les personnes qui effectuent la curation doivent s'assurer que les données quantitatives en sciences sociales et tout outil de collecte des données (p. ex., des outils de sondage) détiennent les licences appropriées. Dans le contexte des sciences sociales quantitatives, les outils de collecte des données d'un projet de recherche peuvent avoir tout autant, sinon plus, de valeur que les résultats provenant des données de cette recherche. Les chercheuses et chercheurs qui utilisent des données administratives (des données municipales ouvertes, des données de Statistiques Canada, etc.) doivent s'assurer que les licences gouvernementales ouvertes permettent le dépôt des jeux de données dérivés et quelles sont les exigences en matière d'attribution pour les sources originales des données.

Le schéma de métadonnées le plus couramment utilisé pour les données en sciences sociales est le **Data Documentation Initiative (DDI)**, qui comprend des champs tels que la taille de l'échantillon, la couverture géographique, l'unité d'analyse (p. ex., un ménage, un individu, etc.) et plusieurs autres champs qui s'appliquent aux sciences sociales. Généralement, les dépôts de données conçus pour héberger des jeux de données en sciences sociales intègrent des champs du DDI dans l'interface de dépôt des données et peuvent ensuite produire automatiquement un fichier de métadonnées lisible par machine (p. ex., XML) comme partie intégrante du processus de téléversement.

> Les bonnes pratiques de GDR en sciences sociales comprennent la conservation d'informations précises et détaillées sur l'étude, dont les mesures utilisées pour la collecte des données, les abréviations ou codes utilisés pour le **nettoyage** et la préparation des données, le script ou le code pour l'analyse des données ainsi que les métadonnées particulières (p. ex., la taille de l'échantillon, la pondération de l'enquête, le code des valeurs factices, etc.). En fournissant des informations complètes et précises sur le projet dans les bons champs de l'interface du dépôt de données, vous augmentez non seulement la découvrabilité et l'impact du projet, mais améliorez également les possibilités de réutilisation des données pour un usage secondaire par d'autres chercheuses et chercheurs.

# Conclusion

Règle générale, la gestion des données de recherche quantitative en sciences sociales implique des processus et considérations similaires aux pratiques de GDR  appliqués aux données spécifiques des autres disciplines. Les

sujets qui sont propres au cycle de gestion des données quantitatives en sciences sociales se rapportent aux outils logiciels particuliers utilisés pour la collecte des données (p. ex., l'utilisation de plateformes infonuagiques de sondages) et à la production ultérieure de multiples fichiers tabulaires au cours du processus de collecte, de nettoyage et d'analyse des données. Les principaux aspects pratiques dans la gestion des données quantitatives en sciences sociales impliquent généralement : le suivi des différentes versions des jeux de données tabulaires par le biais de conventions de nommage de fichiers uniformément appliquées; des noms de fichiers et de variables qui utilisent du texte ou des abréviations lisibles par machine; l'utilisation d'un outil de collecte des données qui permet la personnalisation du formatage des sondages et le maintien d'une documentation exhaustive (p. ex., un guide de codification et des métadonnées) pour s'assurer que les données respectent le plus possible les principes **FAIR**.

### Questions de réflexion

1. Pourquoi est-il important d'établir un PGD pour les données d'enquête quantitative en sciences sociales?
2. Comment le choix de la conception de la recherche et de la méthode de collecte des données est-il lié aux aspects de la GDR dans un projet de recherche quantitative en sciences sociales?

### Éléments clés à retenir

- Les conceptions descriptives visent à explorer un phénomène ou une observation afin d'y décrire un effet, tandis que les conceptions exploratoires visent à expliquer un phénomène ou une observation afin d'y comprendre un effet. Avant même le début du projet de recherche quantitative en sciences sociales, un PGD peut être utile pour établir les conventions de nommage des fichiers, la hiérarchie des répertoires, la préparation des métadonnées et de la documentation pertinente ainsi que la marche à suivre pour le dépôt éventuel des données .
- Les plateformes de sondages les plus couramment utilisées en sciences sociales sont les produits logiciels infonuagiques. En utilisant ce type de plateforme, vous devez tenir compte

des implications en matière de cybersécurité et de protection de la vie privée des personnes participantes. Au cours de la phase de collecte des données, gardez à l'esprit comment les tableurs seront versionnés et nommés en vue de leur réutilisation.

- La crise de la reproductibilité se rapporte à l'incapacité des chercheuses et chercheurs à répéter ou reproduire les conclusions des recherches publiées. Si la vérification d'une étude par répétition n'est pas possible, c'est généralement en raison d'un problème avec les données originales, de l'indisponibilité des données ou de la description insuffisante des étapes entreprises pour l'analyse des données, ce qui empêche de produire les mêmes résultats. Ceci a un impact direct sur l'accès aux données qui sous-tendent les publications en sciences sociales quantitatives, généralement par le biais de dépôts de données publics.

# Lectures et ressources supplémentaires

## Alliance de recherche numérique du Canada

- Exemples de PGD en sciences sociales:
  - *Plan de gestion des données pour personnes, places, politiques et perspectives* (https://doi.org/10.5281/zenodo.4116582)
  - *Plan de gestion des données pour utilisation des sites Web de profils universitaires* (https://zenodo.org/record/4116569#.ZCC1zhWZPao)

## Consortium of European Social Science Data Archives (CESSDA)

- *Data Management Expert Guide* (https://www.cessda.eu/Training/DMEG)

## Data Curation Network

- Microsoft Excel (https://github.com/DataCurationNetwork/data-primers/blob/master/Excel%20Data%20Curation%20Primer/Excel%20Data%20Curation%20Primer.md)
- SPSS (https://github.com/DataCurationNetwork/data-primers/tree/master/SPSS%20Data%20Curation%20Primer)

## ICPSR

- *What is a Codebook (https://www.icpsr.umich.edu/web/ICPSR/cms/1983)*
- *Guide to Social Science Data Preparation and Archiving* (https://www.icpsr.umich.edu/web/pages/deposit/guide/)
- *Sample Data Management Plan for Depositing Data with ICPSR (https://www.icpsr.umich.edu/web/pages/datamanagement/dmp/plan.html)*

Pour des exemples en lien avec l'application de la GDR dans le contexte des sciences sociales, consultez Emmerlhainz, C. (2020). *Tutorials on Ethnographic Data Management*. Data in the Disciplines IMLS Grant. https://library.lclark.edu/dataworkshops/ethnography-modules (https://library.lclark.edu/dataworkshops/ethnography-modules)

# Bibliographie

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature, 533*, 452-454. https://doi.org/10.1038/533452a (https://doi.org/10.1038/533452a)

de Vaus, D. (2001). *Research design in social research*. Sage Publications

Figueiredo, D., Lins, R., Domingos, A., Janz, N. et Silva, L. (2019). Seven reasons why: A user's guide to transparency and reproducibility. *Brazilian Political Science Review, 13*(2). https://doi.org/10.1590/1981-3821201900020001 (https://doi.org/10.1590/1981-3821201900020001)

Konnikova, M. (2015, 22 mai). How a gay-marriage study went wrong. *The New Yorker*. https://www.newyorker.com/science/maria-konnikova/how-a-gay-marriage-study-went-wrong (https://www.newyorker.com/science/maria-konnikova/how-a-gay-marriage-study-went-wrong)

Turkyilmaz-van der Velden, Y., Dintzner, N. et Teperek, M. (2020). Reproducibility starts from you today. *Patterns, 1*(6), 1-6. https://doi.org/10.1016/j.patter.2020.100099 (https://doi.org/10.1016/j.patter.2020.100099)

## À propos des auteurs

Dr. Alisa Beth Rod

Alisa Beth Rod est spécialiste de la gestion des données de recherche à la bibliothèque de l'Université McGill. Alisa détient une maîtrise et un doctorat en sciences politiques de l'Université de Californie, Santa Barbara, et un baccalauréat en bioéthique de l'American Jewish University. Avant de rejoindre McGill, Alisa a été méthodologiste d'enquête chez Ithaka S+R, puis directrice associée de l'Empirical Reasoning Center au Collège Barnard de l'Université Columbia. Elle possède une expérience approfondie en collecte et utilisation de données d'êtres humains dans le contexte de la recherche par sondage, des méthodes qualitatives et des systèmes d'information géographique (SIG).

## Dr. Biru Zhou

Biru Zhou est conseillère principale (gestion des données de recherche) au bureau du vice-recteur (recherche et innovation) de l'Université McGill. Biru est titulaire d'une maîtrise et d'un doctorat en psychologie de l'Université Concordia. Après avoir terminé sa formation postdoctorale à l'École de santé publique de l'Université de Montréal, elle s'est jointe à l'Université McGill en 2016. Elle possède une vaste expérience dans la conception et la réalisation d'études interculturelles impliquant des données humaines sensibles recueillies par le biais de sondages en ligne et d'expériences en laboratoire.

16.

# LES DONNÉES DE RECHERCHE GÉOSPATIALES AU CANADA: UN SURVOL DES PROJETS RÉGIONAUX

Martin Chandler; Kara Handren; Stéfano Biondo; Amber Leahey; Sarah Rutley; et Rhys Stevens

**Objectifs d'apprentissage**

À la fin de ce chapitre, vous pourrez :

- Comprendre l'état actuel de l'infrastructure et des services de données de recherche géospatiales à travers les différentes régions du Canada.
- Expliquer quelques-unes des considérations uniques relatives à la gestion des données géospatiales.
- Fournir des exemples et des modèles de gestion des données géospatiales.
- Reconnaître l'avenir de la gestion des données de recherche géospatiales au Canada.

## Introduction

Les bibliothèques au Canada offrent une variété de services pour la découverte, l'accès et la préservation des données de recherche géospatiales. Les infrastructures et services ont été développés au niveau régional, principalement par les établissements universitaires, pour soutenir la gestion des collections et des ressources de données géospatiales, créant ainsi une mosaïque de services de données de recherche à travers le pays. Ce chapitre fera un survol des approches et des principaux projets d'infrastructures dans la gestion des données de recherches géospatiales au Canada.

Les données spatiales/géospatiales (ci-après désignées « géospatiales ») n'ont pas toujours été perçues comme des données auxquelles une attention particulière doit être accordée du point de vue de la **gestion des données de recherche** (GDR). Toutefois, en raison des aspects uniques de leur création, utilisation et accès, les données géospatiales nécessitent des considérations particulières de gestion qui se distinguent des autres domaines de la GDR.

Règle générale, la curation des données géospatiales tombe sous la responsabilité des bibliothécaires ou gestionnaires de données géospatiales qui possèdent une expertise en la matière, puisque ces deux groupes sont mieux outillés pour répondre aux défis associés aux données géospatiales. Ce chapitre vise justement à préciser les défis qui sont propres à la GDR géospatiales; à énumérer les nombreux projets régionaux actuellement en cours ou en développement qui cherchent à répondre aux défis en matière de préservation et d'accès aux données de recherche géospatiales et à discuter des tendances futures pour la GDR géospatiales au Canada.

## Les données géospatiales et les SIG

Que sont les données géospatiales? Et qu'est-ce qui distingue les données de recherche géospatiales des données de recherche dans leur ensemble? Toutes données relatives à des objets ou à des événements associés à un lieu sont des données géospatiales. Elles peuvent inclure des instances où le lieu est statique (dans un lieu défini sur une période donnée comme un immeuble ou un tremblement de terre) ou dynamique (qui manifeste un changement ou un mouvement sur une période donnée tel que la croissance urbaine ou les effets d'une sécheresse sur les nappes phréatiques environnantes). Les données géospatiales combinent des informations de localisation avec les caractéristiques d'un objet, d'un événement ou d'un concept (les données descriptives) et aussi parfois, mais pas toujours, avec des informations temporelles (Stock et Guesgen, 2016).

Les données géospatiales dépendent souvent de l'utilisation d'un système d'information géographique (SIG) tel que QGIS, ArcGIS ou Google Earth. Ce système offre de nombreux moyens et méthodes de développer, d'utiliser et d'exporter des données géospatiales, y compris la création et le partage de jeux de données. Les données de recherche géospatiales combinent ou relient des points de données spatiales (ou caractéristiques) avec d'autres données sources et variables pour faciliter l'utilisation des données. Ces variables peuvent souvent inclure des données de nature géographique telles que les données de recensement au niveau des secteurs de recensement ou des codes postaux.

Les considérations de GDR géospatiales dépendent beaucoup de l'exportation des données vers

différents formats (l'**interopérabilité** des formats), de la possibilité d'afficher et de réutiliser les cartes statiques, de l'utilisation et de la réutilisation des données statistiques et géographiques, de la réutilisation d'applications de données interactives et de l'utilisation de fonctionnalités et de composantes cartographiques.

Étant donné la nature de la GDR géospatiales, de son utilisation dans les SIG et de la façon dont les données y sont traitées, une connaissance préalable de la gestion des données est souvent nécessaire. Une introduction à l'utilisation des données géospatiales est disponible dans le guide *Learn QGIS* de Anita Graser ou des ouvrages de Esri Press, *Getting to know ArcGIS* ou sa série *GIS Tutorial for…* Bien qu'il soit possible de créer des données dans un SIG, l'outil est plus souvent utilisé pour faire des liens entre des données géospatiales et d'autres types de données (p. ex., des **données tabulaires** avec des données géospatiales préexistantes).

D'autres chapitres de ce manuel abordent des sujets de GDR plus généraux, dont la gestion des fichiers. Avec les projets décrits ci-dessous, la création de données géospatiales et la gestion des données de recherche géospatiales sont mises de l'avant. Ce chapitre abordera principalement les différents projets régionaux qui ont été entrepris ou qui sont en cours dans les bibliothèques universitaires canadiennes pour gérer et préserver les données de recherche géospatiales au pays.Parmi ces projets, l'accent sera mis sur ceux qui visent à rendre les données de recherche géospatiales repérables, accessibles et réutilisables pour une grande variété de publics et de personnes susceptibles d'utiliser les données.

La gestion et la réutilisation des données de recherche géospatiales nécessitent une réflexion sur les espaces physiques à partir desquels les données ont été recueillies ou auxquels elles se rapportent. Il y a une tendance vers la découverte géospatiale qui intègre des **cartes de base** avec une recherche utilisant les mots-clés. Un affichage géographique et un aperçu des jeux de données sont souvent offerts (consultez, par exemple, Scholars GeoPortal (https://geo2.scholarsportal.info/) de l'OCUL ou CarrefourGéo (https://geohub-fr.lio.gov.on.ca/?locale=fr) de l'Information sur les terres de l'Ontario). Les données sont ensuite affichées dans un format réduit directement sur la carte de base ou elles sont représentées par une étendue géographique (bounding box) qui illustre leur ampleur géographique. Il est important de noter que la gestion des données de recherche géospatiales nécessite une infrastructure robuste; il en sera question dans quelques-uns des projets régionaux décrits plus loin dans le chapitre. Puisque cette infrastructure est généralement plus coûteuse, la gestion de données géospatiales pour le stockage à long terme et la découverte implique souvent des projets consortiaux plutôt qu'individuels.

# Les formes de données géospatiales

Bien que de nombreuses formes de données puissent comprendre des éléments géospatiaux (p. ex., une variable de ville, de division de recensement ou d'adresse), les données géospatiales peuvent aussi comprendre des formats distincts sous forme de **données matricielles** ou **vectorielles**. Les données matricielles (ou raster) sont des  matrices de cellules organisées en rangées et en colonnes, dont chaque cellule comporte une information et une représentation visuelle. Par exemple, une carte ou un dessin numérisé ainsi qu'une image satellite constituent des données matricielles (Esri, 2016).


Figure 1. Données matricielles d'une carte numérisée : Bellin, 1764.

Les données vectorielles sont une représentation de caractéristiques ou de phénomènes réels dans un SIG avec

des données sous-jacentes qui permettent de faire des liens entre la(les) caractéristique(s) et d'autres formes de données. Les données vectorielles peuvent être divisées en point, ligne ou polygone. Les points de données sont des endroits uniques dans l'espace (p. ex., l'emplacement d'un arbre); les lignes de données ou polylignes sont deux ou plusieurs de ces points, des sommets, dont le premier et le dernier ne sont pas égaux, montrant une ligne ou une série de lignes (p. ex., une route) et les données en polygone sont trois ou plus sommets dont le dernier est équivalent au premier, formant ainsi une figure fermée (p. ex., les frontières d'une propriété, d'une région ou d'une province) (QGIS Documentation, s.d.).

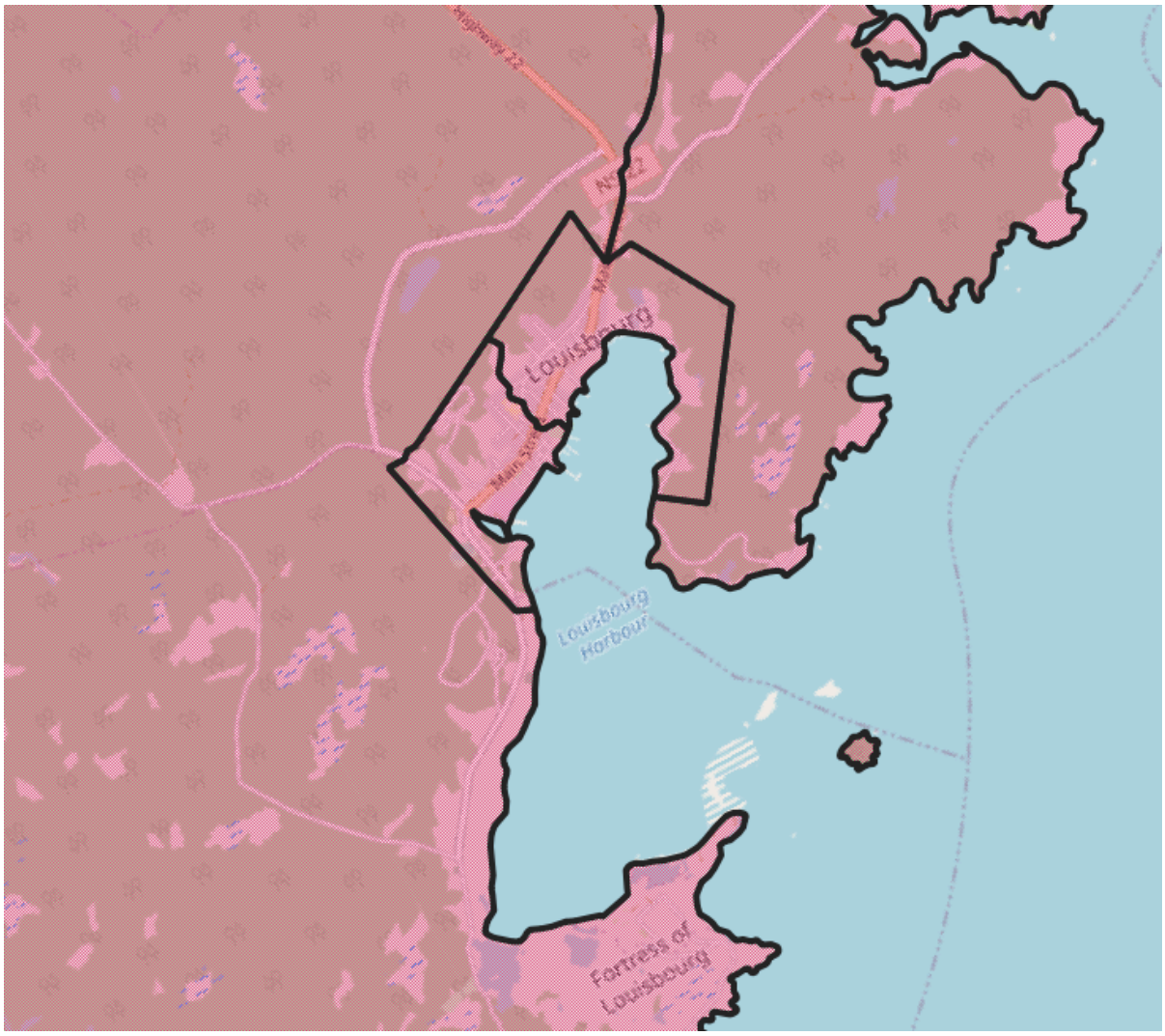



Figure 2. Des données en polygone: Statistiques Canada (2019) et des personnes contribuant à OpenStreetMap (2023).

Les données géospatiales tabulaires existent souvent dans un tableau ou un format où les valeurs sont séparées par des virgules (CSV). Elles peuvent être aussi simples qu'une adresse ou le nom d'un lieu géographique (p. ex., Unama'ki) ou aussi complètes qu'un ensemble de points, d'étendues, d'identifiants spatiaux et de hiérarchies de noms ou d'identifiants géographiques.

## Les données géospatiales comme interaction

Puisque les données géospatiales impliquent une représentation de l'espace, elles sont rarement créées en tant que jeu de données unique. Elles dépendent plutôt d'interactions avec d'autres jeux de données spatiales dont les données spatiales sous-jacentes qui servent à les localiser à l'intérieur d'un SIG et/ou qui impliquent le développement de données spatiales supplémentaires pour commencer, faire avancer ou conclure une analyse des données en question. La GDR peut nécessiter une planification des interactions avec des données abstraites. Mais la GDR géospatiales nécessite une attention particulière à la planification des interactions entre les données aussi bien abstraites que physiques, aux différents modes et méthodes nécessaires à ces interactions et à la façon dont le logiciel d'analyse traitera ces interactions. L'utilisation d'un SIG implique en soi une planification soigneuse de la gestion des données puisque le logiciel monte les données à partir d'un endroit numérique plutôt que de les copier dans le logiciel. En faisant la sauvegarde d'un projet dans un SIG, l'emplacement des données est sauvegardé. Ainsi, le déplacement d'un jeu de données peut rendre le projet intraitable à moins que la personne responsable corrige l'emplacement du jeu de données.

Il est important de préciser que la création de données géospatiales est aussi bien une fin en soi qu'un développement vers d'autres fins, dont l'analyse, la visualisation ou l'analyse préalable avant la conception d'un projet. Les données géospatiales peuvent être créées en guise de résultats de recherche ou comme appui pour la préparation, l'analyse ou la visualisation d'une autre source de données. Elles représentent donc une fin et un intermédiaire. Autrement dit, elles servent comme résultat de recherche (comme peut l'être tout autre jeu de données), comme outil d'analyse (avec SPSS, NVivo, Voyant, etc.) et comme outil de présentation des données (avec Tableau, ggplot, etc.). De plus, tandis qu'un jeu de données numériques peut se présenter comme un fichier unique à utiliser, le jeu de données géospatiales nécessite des données géospatiales d'appui, des projections de cartes (p. ex., la variété de moyens utilisés pour représenter un globe tridimensionnel dans une représentation en deux dimensions) et des systèmes de référence des coordonnées (p. ex., les différents systèmes qui déterminent où et comment un jeu de données géospatiales doit être affiché sur une carte).

## Les données en tant qu'objet ou en tant que processus

Enfin, en raison de la nature interconnectée et interactive des données géospatiales en recherche, il faut tenir compte de la GDR géospatiales aussi bien en matière de données produites par la recherche qu'en termes de données utilisées au cours du processus de recherche. Par exemple, une chercheuse peut avoir besoin d'une partie de fichier des limites de recensement de Statistiques Canada. Dans son analyse, elle pourrait alors extraire une partie du fichier des limites. Ce faisant, ces données constituent un produit de recherche au même titre que l'extraction des données de recensement constituerait des données de recherche. L'extrait du fichier des limites peut alors représenter une étape préalable à l'analyse et peut être modifié en utilisant différents systèmes de référence des coordonnées et/ou différentes projections (p. ex., une projection conforme de

Lambert modifiée en une projection Web Mercator). La frontière entre les données préparées pour une utilisation en recherche et les données créées en raison d'une utilisation en recherche devient alors plus floue pour les données géospatiales. Ainsi, la GDR géospatiales inclura – du moins pour les besoins de ce chapitre – la gestion aussi bien des jeux de données préparés que des données issues de la recherche. La suite de ce chapitre présente des projets réalisés dans diverses régions du Canada qui répondent à l'un ou l'autre de ces trois objectifs :

1. Contribuer actuellement à la GDR géospatiales;
2. Contribuer dans le futur à la GDR géospatiales;
3. Identifier les difficultés liées à la GDR géospatiales.

# Les projets géospatiaux régionaux

Tels que noté plus tôt, les besoins en gestion des données géospatiales impliquent que l'accès et la préservation soient surtout réalisés par le biais de solutions en consortium plutôt que par des établissements individuels. Nous discuterons ci-dessous de la variété de solutions de différents consortiums régionaux pour la gestion des données de recherche géospatiales.

## Les provinces de l'Atlantique

En date de 2023, les provinces de l'Atlantique n'ont toujours pas accès à des méthodes partagées ou consortiales pour le stockage et la livraison des données, malgré le rôle de premier plan des bibliothèques de la Nouvelle-Écosse en matière de systèmes partagés (Marshall, 1999, p.134). Toutefois, les bibliothécaires de données des établissements universitaires ont discuté de cette problématique et l'ont identifiée comme étant un besoin. De plus, des discussions ont été entamées avec des organisations ayant des systèmes consortiaux, particulièrement celui de Scholars GeoPortal en Ontario. Il y a un certain optimisme vis-à-vis de l'établissement d'un système national qui serait géré soit par un système partagé en consortium, soit par les bibliothèques universitaires associées à l'Alliance de recherche numérique du Canada. Ces discussions préliminaires demeurent informelles, mais il importe de soulever qu'elles ont lieu (IDD-Atlantique, communications personnelles, fév-mars 2022).

Comme il n'existe pas de systèmes partagés au niveau provincial ou régional, la mise en œuvre de la GDR géospatiales est entièrement entre les mains des établissements locaux dans les situations où les données de recherche géospatiales ont été reconnues. Chaque établissement a développé une approche de la gestion des données de recherche qui lui est propre, déterminée surtout en fonction des capacités de l'établissement ou de la bibliothèque à soutenir leur communauté. De la même manière, chaque établissement a développé son

approche en matière de GDR géospatiales. Souvent, surtout dans le cas d'établissements plus petits, le traitement de ces questions se fait selon les besoins (p. ex., si la bibliothèque reçoit une demande d'un membre du corps professoral, le personnel cherchera des solutions appropriées selon ce qui est réalisable, souvent sous la forme d'un dépôt Dataverse ou d'un jeu de données hébergé localement). Bien que de tels systèmes ad hoc ne constituent pas des solutions idéales pour le stockage, l'utilisation et la découverte de données de recherche géospatiales, ils demeurent les meilleures options quand les ressources sont limitées (IDD-Atlantique, communications personnelles, fév-mars 2022).

Plusieurs établissements ont choisi d'utiliser des instances de Borealis (https://borealisdata.ca/fr/) (anciennement Dataverse de Scholars Portal) en tant que dépôt de données pour héberger les données créées par ses chercheuses et chercheurs. (Consultez Lunaris (s.d.) pour la liste des dépôts de données des établissements et des plateformes d'hébergement.) Lunaris (anciennement le service de découverte du DFDR) est distinct de Borealis, mais moissonne les dépôts des établissements pour offrir un outil qui permet d'accéder aux données et de naviguer les dépôts locaux). Les instances de Dataverse offrent une bonne découvrabilité. Toutefois, Dataverse n'offre pas d'outil robuste d'affichage géospatial ou de plateforme de découverte. L'outil Geodisy a pallié en partie ce manque (consultez ubc-library (2022) et autres références dans ce chapitre), mais il a été remplacé par Lunaris. Le système dans sa forme actuelle ne permet pas l'affichage des données et le découpage de zones précises; il ne permet qu'un affichage de base de la couverture des données. Il y a d'importantes lacunes au niveau de la recherche et de l'utilisation de données géospatiales ainsi que du stockage et des services de données de recherche géospatiales.

Le centre du SIG de l'Université Dalhousie est le système le plus développé de la région des provinces atlantiques. Construit à partir du ArcGIS Hub d'Esri, le portail est en cours de développement et vise à donner accès à tous les jeux de données détenus par ou avec une licence de l'Université. Ainsi, il permet les recherches géospatiales et des méthodes de visualisation, en plus du découpage préliminaire avant le téléchargement. Toutefois, comme il héberge des jeux de données sous licence, le portail est uniquement réservé à la communauté  de Dalhousie et n'est pas accessible aux autres établissements. Cette situation ne peut que décevoir la communauté externe qui voudrait accéder à ces données.

## Le Québec

Au Québec, chaque bibliothèque universitaire assurait la gestion et la diffusion des données géospatiales de façon indépendante, et ce de manière plus ou moins automatisée jusqu'en 2019. En 2015, une entente historique entre le Bureau de coopération interuniversitaire (BCI) et le ministère de l'Énergie et des Ressources naturelles (MERN) a ouvert la porte à un nouveau mode de gestion et de diffusion des données géospatiales au sein du réseau universitaire québécois.

Jusqu'en 2015, toutes les universités québécoises devaient acheter les données gouvernementales

individuellement et ne pouvaient se les prêter entre elles en raison des contrats de licence. Du jour au lendemain, grâce à l'entente BCI-MERN, les universités ont pu utiliser et se partager plus de 250 **couches** qui représentent 50 téraoctets (To). Comment gérer et partager cette masse de données? Ce ne sont pas toutes les universités qui possèdent une plateforme adéquate pour organiser et diffuser ces données géospatiales au profit de l'enseignement et de la recherche.

Dans une vision de collaboration interuniversitaire et de mutualisation des processus et des ressources, la Bibliothèque de l'Université Laval s'est montrée ouverte à partager son expertise et son savoir-faire dans le domaine géospatial via la création d'une plateforme partagée, gérée par l'Université Laval et accessible aux bibliothèques participantes. Leur solution intégrerait l'ensemble des fonctionnalités nécessaires à la découverte, la visualisation et l'extraction des données géospatiales, ainsi que leur chargement dans un environnement sécuritaire et performant.

Le résultat est Géoindex, une infrastructure unique accessible aux 18 établissements universitaires québécois via 18 portes d'entrée paramétrées par chacun d'eux selon leurs préférences. Grâce à son moteur de recherche spatial et textuel performant, cette plateforme permet de facilement découvrir, de visualiser et d'extraire des données géospatiales et des photographies aériennes pour appuyer l'enseignement et la recherche. Géoindex se décline en deux modules : le module Géospatial et le module Géophoto, qui sont décrits ci-dessous.

L'entente BCI-MERN a servi de levier pour développer Géoindex, mais cette nouvelle plateforme permet d'héberger et diffuser d'autres données géospatiales provenant de sources diverses gérées selon différentes licences. On retrouve donc dans Géoindex des données sous licences provenant de l'entente, dont des données LiDAR qui offrent aux chercheuses et chercheurs de nouvelles interprétations du territoire. Mais il comprend aussi des données issues de projets de recherche comme celles de L'Atlas des vulnérabilités, qui illustre entre autres l'indice de sensibilité face aux vagues de chaleur, ou encore les données bathymétriques de l'Arctique recueillies par le brise-glace scientifique Amundsen.  Chaque couche d'information est décrite selon un profil de **métadonnées** (Profil UL) qui répond aux critères du Profil nord-américain (PNA) de la norme ISO 19115. L'utilisation du Répertoire de vedettes-matière de l'Université Laval (RVM) est mise à profit pour la standardisation des descriptions des sujets utilisés.

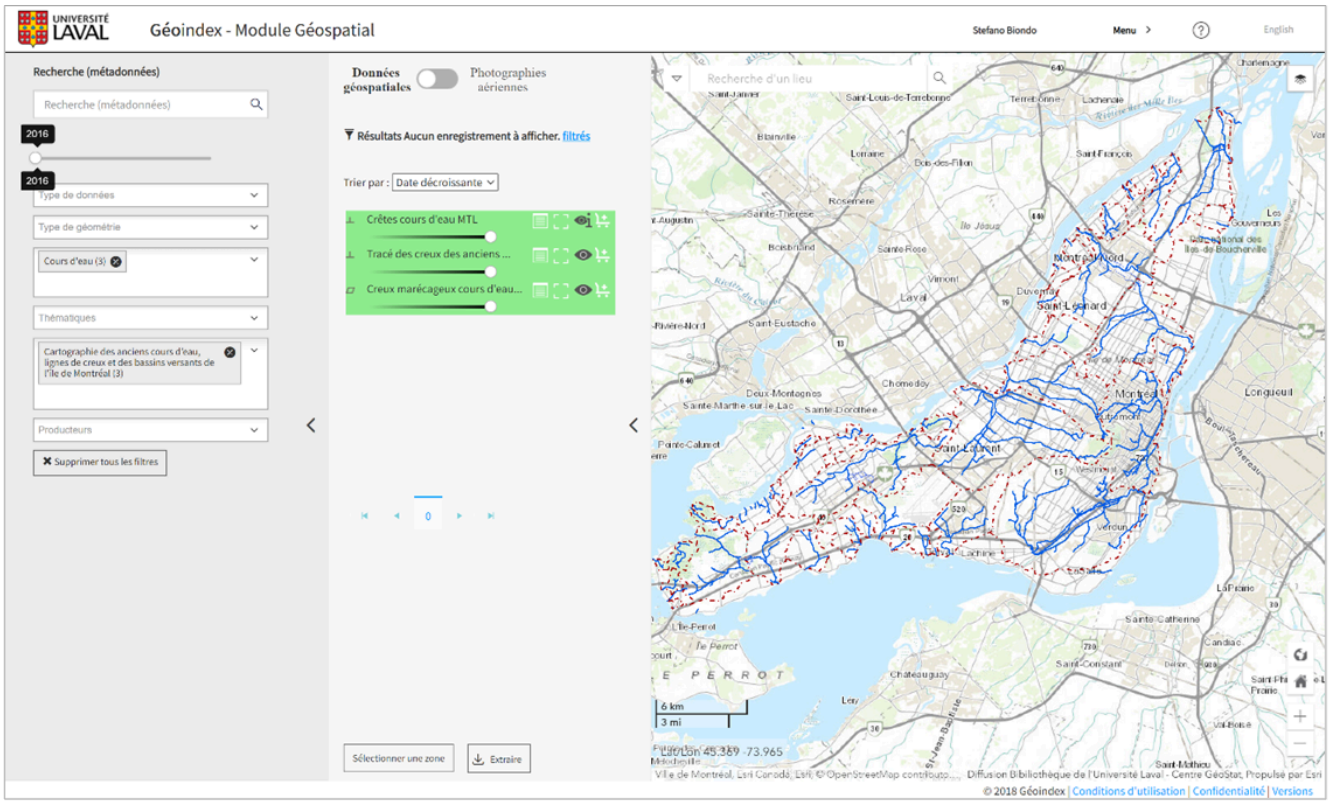



Figure 3. Exemple de données de recherche géospatiales : La cartographie des anciens cours d'eau de Montréal, réalisée par une chercheuse de l'Université de Montréal.

Les données sont accessibles à l'ensemble du réseau universitaire, mais certaines sont aussi ouvertes et accessibles au grand public, dont plus de 250 cartes topographiques datant de 1909 à 2000. Géoindex permet aussi de mettre en valeur des documents historiques provenant des collections des bibliothèques, comme les cartes topographiques, mais aussi des documents encore plus anciens comme cette carte relatant la première expédition de John Franklin dans le Grand Nord canadien en 1819, que la bibliothèque de l'Université Laval a numérisée et géoréférencée afin de lui donner une seconde vie.

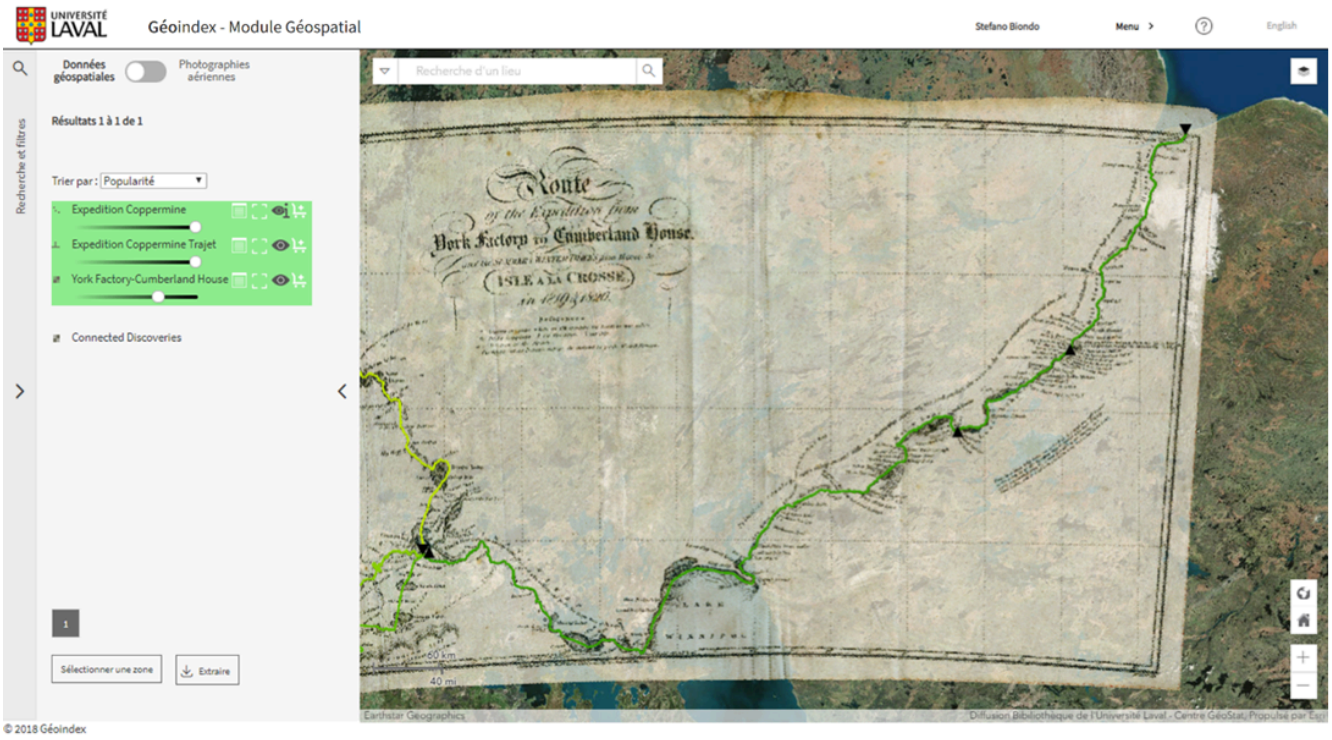



Figure 4. Exemple de données géospatiales pouvant lancer un projet de recherche: Carte historique géoréférencée et trajet vectorisé de l'expédition Coppermine menée par Sir John Franklin entre 1819-1822.

Toujours dans une perspective de soutien à l'enseignement et à la recherche en facilitant la découverte d'information géographique, le module Géophoto dédié au repérage des photographies aériennes intégré à Géoindex, a été bonifié au cours de l'année 2022. En basculant vers ce module, les usagers peuvent consulter l'ensemble des inventaires de photographies aériennes détenues par les universités québécoises, soit plus de 1 200 000 photographies aériennes datant du 20$^{e}$ siècle. Il s'agit d'une information primaire, ou donnée brute, d'une grande importance pour comprendre le territoire tel qu'il était à un moment précis. Une entente signée de nouveau entre le BCI et le MERN permettra également d'ajouter d'ici 2026 plus d'un million de photographies aériennes numérisées par le MERN. En février 2023, on dénombrait 400 000 exemplaires numérisés disponibles dans le module Géophoto.

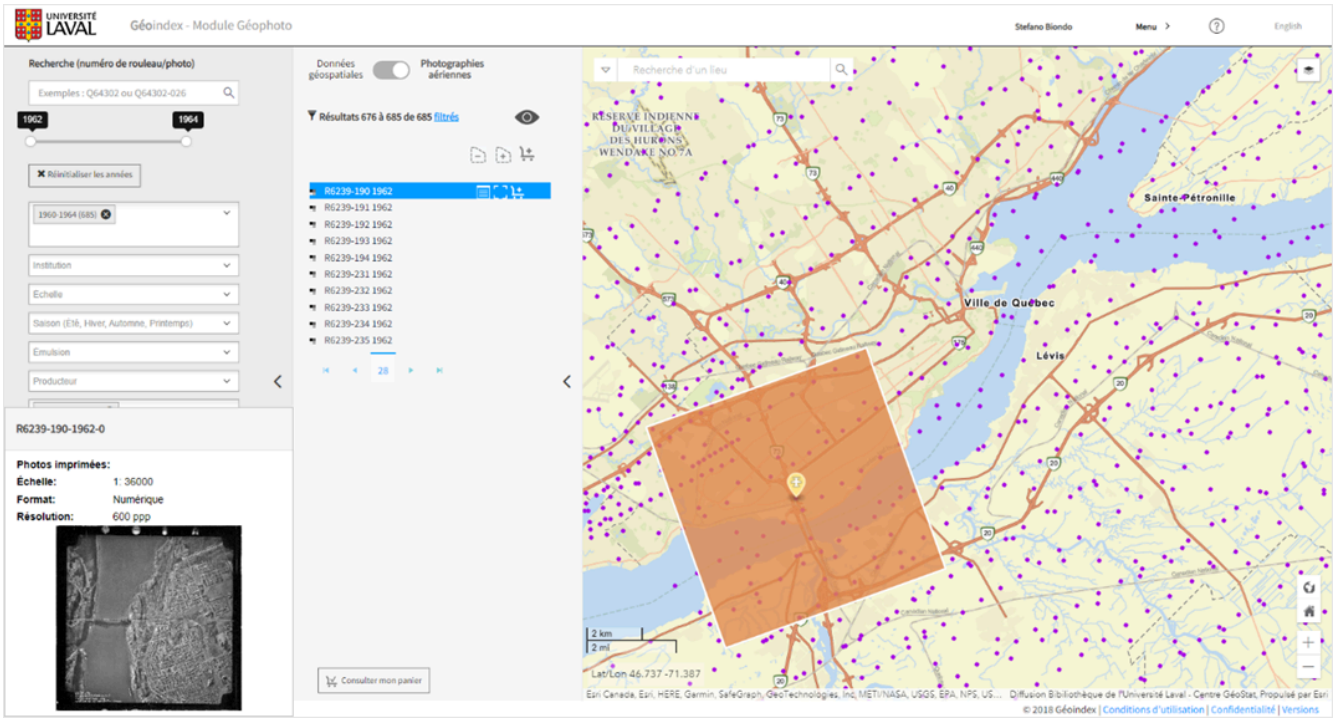



Figure 5. Le module Géophoto de la plateforme GéoIndex qui permet de consulter par le biais d'un accès unique l'ensemble des collections de photographies aériennes des universités québécoises.

Bien que la plateforme Géoindex puisse héberger et diffuser des données géospatiales issues de projets de recherche, elle n'a pas été conçue spécifiquement pour ce type de données. Par exemple, il n'y a pas d'émission de **DOI** et les métadonnées de ce géocatalogue ne sont pas exposées sur le Web, elles ne peuvent donc pas être moissonnées par d'autres moteurs de recherche. Il est prévu, lors d'une future mise à jour, de rendre les métadonnées se trouvant dans Géoindex ouvertes et accessibles aux autres moteurs de recherche.

Pour l'instant, les cas de données géospatiales issues spécifiquement de projets de recherche sont peu fréquents dans Géoindex. Toutefois, ses fonctionnalités de découverte, de visualisation et d'extraction feront probablement augmenter ce nombre au cours des prochaines années sans toutefois remplacer les dépôts de données de recherche traditionnels comme Dataverse. Géoindex doit être perçu comme un complément aux dépôts traditionnels avec des liens entre eux pour faciliter la découverte et la consultation.

## L'Ontario

Les bibliothèques en Ontario collaborent depuis longtemps à la mise en place de systèmes de découverte et de gestion pour des collections partagées, coordonnée par le biais de l'Ontario Council of University Librairies (OCUL) (https://ocul.on.ca/). Tel que précisé dans le chapitre 4, « Historique et paysage canadien de la gestion des données de recherche, » l'OCUL a été établi en 1967 et représente un consortium des vingt et une bibliothèques universitaires dans la province de l'Ontario. L'organisme s'implique dans les activités collectives

telles que l'achat, le stockage et la fourniture de  ressources et de services aux bibliothèques. L'infrastructure derrière ces systèmes partagés est soutenue par Scholars Portal, le fournisseur d'infrastructure numérique de l'OCUL qui regroupe des bibliothécaires, des personnes responsables de l'administration des systèmes et d'autres du développement. Le personnel de Scholars Portal est à l'emploi des bibliothèques de l'Université de Toronto. Cette infrastructure provinciale, gérée par le consortium, héberge une variété de collections partagées. Elle a été impliquée dans l'établissement, le maintien et le soutien d'une panoplie de plateformes d'accès pour la collection et la livraison des données ainsi que dans le soutien à la clientèle. Ces initiatives comprennent des collections de publications, telles que Scholars Portal Périodiques et Scholars Portal Livres, ainsi que des plateformes axées sur les microdonnées et les données géospatiales, dont Scholars GeoPortal, Odesi (http://odesi.ca) et Borealis. Une variété de collections sous licence partagée, de collections numériques ouvertes et de collections d'archives sont hébergées et accessibles aux chercheuses et chercheurs universitaires des établissements membres participants.

L'OCUL Geo Community (anciennement le Map Group de l'OCUL) a joué un rôle déterminant dans le développement de Scholars GeoPortal (http://geo2.scholarsportal.info/) en 2012. Scholars GeoPortal est un outil Web de découverte des données qui fournit un accès aux données commerciales et à celles sous licence, aux collections nationales de données brutes, aux données des gouvernements régionaux, aux **données ouvertes** et aux données d'imagerie matricielles (notamment les projets, acquisitions et cartes numérisées issus des gouvernements). L'application est une construction personnalisée qui utilise une combinaison des technologies d'Esri et d'autres logiciels déjà en usage chez Scholars Portal. Elle exploite ArcGIS Server comme base de données et serveur en arrière-plan et elle utilise les outils API fournit par Esri pour la visualisation et le téléchargement des données stockées dans ces serveurs par le biais d'un SIG frontal personnalisé. Ce SIG sert également de catalogue partagé et d'outil de découverte des données et est soutenu par un robuste éditeur de métadonnées qui produit des métadonnées conformes à la norme ISO 19115 stockées dans une base de données MarkLogic XML. Actuellement, un nouveau projet de développement est en cours pour remettre à jour le GeoPortal afin de garantir la pérennité de la plateforme et d'assurer qu'elle continue de répondre aux besoins de la communauté. Dans le cadre de ce travail de refonte, des intégrations à Borealis sont explorées (le sujet est abordé  au chapitre 4 dans un contexte national et régional).

Les bibliothèques de l'OCUL ont travaillé à faciliter l'accès aux données géospatiales rendues disponibles grâce au développement d'infrastructures et de licences partagées. Elles ont également activement participé à des projets spéciaux et des initiatives, aussi bien en Ontario que dans le contexte canadien plus large. Le projet des cartes topographiques historiques a mené à la numérisation de plus de 1000 cartes topographiques aux échelles de 1:25 000 et 1:63 360, couvrant les années 1906 à 1977. Les bibliothèques travaillent actuellement sur un projet plus important qui vise à réutiliser ces processus de travail sur la collection de cartes à 1:50 000 du Système national de référence cartographique (SNRC) et d'intégrer ces cartes au GeoPortal et à Borealis, fournissant ainsi une plus grande **intégration** de la collection avec l'infrastructure de données de recherche

nationale du Canada (p. ex., Lunaris). À ce jour, plus de 6000 cartes de la collection à 1:50 000 ont ainsi été rendues disponibles.

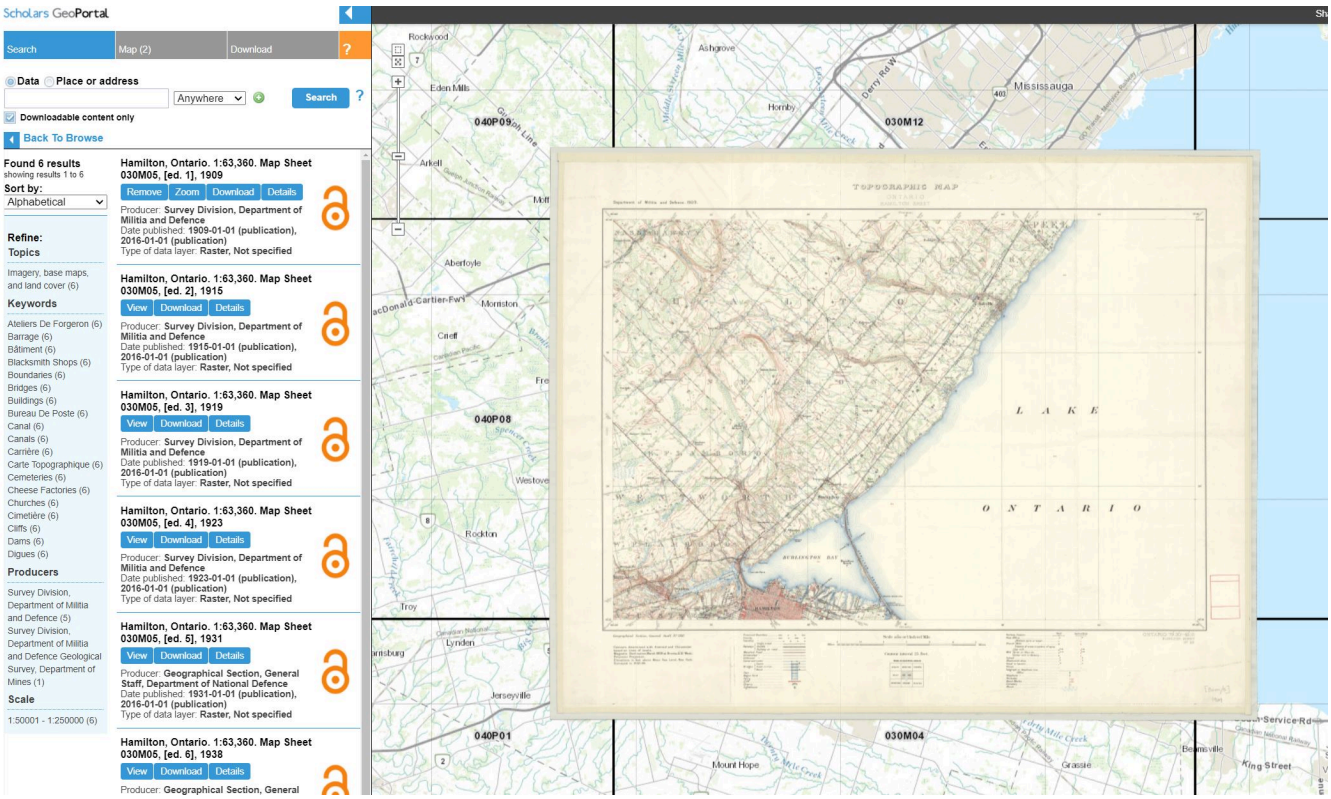



Figure 6. Une carte du SNRC de Hamilton, Ontario (Fiche 030M05), telle qu'affichée dans GeoPortal.

L'Ontario Library Research Cloud (OLRC) est une collaboration entre les bibliothèques universitaires de l'Ontario visant à construire un réseau de stockage infonuagique à grande capacité et géographiquement distribué en utilisant des technologies à code **source libre**. L'OLRC est conçu pour héberger d'importants volumes de contenus numériques permettant la préservation à long terme économique et durable ainsi que le soutien aux outils de recherche en matière d'exploration de données et de textes (*text and data mining*). Cette ressource est actuellement exploitée par de nombreux établissements de l'OCUL pour la préservation de leurs données géospatiales afin d'assurer un accès à long terme. Permafrost s'appuie sur l'OLRC en soutenant les processus de travail pour la création de **paquets d'informations archivés** (AIP) en utilisant une instance d'Archivematica qui est gérée et appuyée par le consortium. Archivematica est une suite d'outils en source libre qui a été développée par Artefactual pour appuyer le versement et la préservation des objets numériques. Dans certains cas, Permafrost est connecté aux dépôts. L'instance de Islandora de la bibliothèque de l'Université McMaster, qui comprend plus de 12 000 cartes, plans et photos aériennes de la collection de cartes de Lloyd Reeds, représente un bon exemple de l'importance de cette infrastructure. Les données sont copiées de façon automatique et régulière dans l'OLRC et stockées en tant que AIP dans leurs archives numériques.

En raison de l'augmentation continue du volume des données, Scholars Portal a identifié le besoin de fournir de nouvelles solutions techniques pour soutenir le transfert de jeux de données volumineux au sein des services de données des bibliothèques universitaires. Cette recherche de solutions numériques est devenue d'autant plus pressante au cours de la pandémie de COVID-19; les restrictions en matière de contact rendaient impossibles les processus de travail existants dans un environnement à distance. Scholars Portal a développé une solution grâce à l'utilisation de Globus, un outil de transfert des données qui soutient les processus de travail pour les transferts de fichiers lourds et le stockage directement dans les environnements de recherche. L'OCUL explore actuellement une intégration plus approfondie dans le cadre de la refonte de Scholars GeoPortal.

Des métadonnées normalisées sont également essentielles pour faciliter l'accès à la recherche et la découverte de données géospatiales. Au cours du développement de GeoPortal, l'OCUL a réalisé d'importantes transformations en recommandant et en adoptant la norme ISO 19115 et des vedettes-matière canadiennes issues des agences gouvernementales fédérales et provinciales. Ces normes pour la création de jeux de données et de métadonnées au niveau des séries ont mené à une amélioration de la découverte, des capacités de recherche et de l'accès aux collections. L'expertise de Scholars Portal à fournir des formations sur les métadonnées géospatiales a également permis une plus grande compréhension de l'importance des normes pour les métadonnées géospatiales à travers la communauté de l'OCUL. Ces normes ont été appliquées aussi bien aux collections locales qu'aux projets spéciaux.

## Les Prairies

Le Council of Prairie and Pacific University Libraries (COPPUL) (https://coppul.ca/) est une association des bibliothèques universitaires des provinces de l'Ouest canadien qui comprend douze membres des Prairies – l'Alberta, la Saskatchewan et le Manitoba – qui sont énumérés dans le tableau 1. La capacité du personnel de ces bibliothèques de répondre aux besoins pour les services spécialisés en données géospatiales varie de façon considérable. Nombreuses sont les bibliothèques qui n'offrent aucun service géospatial ou de SIG, tandis que les plus grands établissements universitaires (p. ex., Calgary, Alberta, Manitoba) sont en mesure d'offrir un plus large éventail de services de données géospatiales. Les bibliothèques du COPPUL desservent des populations étudiantes et professorales très disparates en grosseur et soutiennent différents programmes académiques avec des exigences en GDR très variées. Ainsi, il existe d'importantes variations dans les types de services offerts par ces bibliothèques. De façon plus précise, ces services impliquent (1) de fournir un accès aux données géospatiales offertes par des agences externes; (2) de créer des produits liés au domaine géospatial et aux SIG utiles à la production de nouvelles recherches; et (3) de gérer les données géospatiales qui ont été produites par les chercheuses et chercheurs de leurs établissements respectifs dans le cadre de leurs activités de recherche.

**Tableau 1: Les activités de gestion des données de recherche géospatiales dans les bibliothèques des Prairies membres du COPPUL.**

| Université | Province | LibGuide de données géospatiales / SIG | Catalogue de données géospatiales/SIG | Dépôt Dataverse pour la GDR | Disponibilité des jeux de données géospatiales en GDR |
|---|---|---|---|---|---|
| Athabasca | AB | ❌ | ❌ | ❌ | ❌ |
| Concordia | AB | ❌ | ❌ | ❌ | ❌ |
| MacEwan | AB | ✔ | ✔ | ✔ | ❌ |
| Mount Royal | AB | ✔ | ❌ | ✔ | ❌ |
| Alberta | AB | ✔ | ❌ | ✔ | ✔ |
| Calgary | AB | ✔ | ✔ | ✔ | ✔ |
| Lethbridge | AB | ✔ | ❌ | ❌ | ❌ |
| Régina | SK | ✔ | ❌ | ✔ | ✔ |
| Saskatchewan | SK | ✔ | ❌ | ❌ | ✔ |
| Brandon | MB | ❌ | ❌ | ✔ | ❌ |
| Manitoba | MB | ✔ | ✔ | ✔ | ✔ |
| Winnipeg | MB | ✔ | ❌ | ✔ | ❌ |

Les bibliothèques membres du COPPUL ont été activement impliquées dans la création de produits en lien avec le domaine géospatial et les SIG pour aider leurs clientèles à repérer et utiliser des jeux de données géospatiales présents dans leurs collections. Les types les plus courants de matériel géospatial inclus dans ces produits comprennent des cartes historiques, des cartes topographiques, des images aériennes, des modèles numériques d'altitude (MNA) et des documents sur le climat et l'environnement. Voici quelques exemples d'initiatives particulières :

- Spatial & Numeric Data Services (SANDS) (https://sands.ucalgary.ca/) de la bibliothèque de l'Université de Calgary a participé au développement de nombreuses applications cartographiques (https://sands.ucalgary.ca/App.php) (en anglais uniquement) qui donnent accès à des cartes historiques rares (p. ex., des cartes sériées des régions des Prairies canadiennes (Three-Mile Sectional Maps of the Canadian Prairies), des plans des cantons de l'Alberta, des plans d'assurance contre les incendies de Calgary). Les cartes originales ont été numérisées et géoréférencées pour permettre la visualisation d'emplacements géographiques sur une carte Web Esri qui peut être téléchargée.
- La sixième phase du Shared Print Archive Network (SPAN) (https://coppul.ca/collections/phase-6-working-group/) (site en anglais uniquement) du COPPUL a été mandatée pour identifier des cartes topographiques et historiques de l'Ouest canadien (à l'échelle approximative de 1:25 000 et 1 :63 360)

pour la préservation et la recherche. L'identification de ces cartes ouvre la voie à de nouvelles possibilités de numérisation et de visualisation semblables aux cartes topographiques disponibles dans SANDS (https://sands.ucalgary.ca/App/CalgaryTopoMaps/) et Scholars GeoPortal de l'Ontario (http://geo1.scholarsportal.info/#_lang=fr&layersInfo_BingMapsRoad_opacity:1;;&basemap=Bing%20Maps%20Road&extent_xmin:-14574524.69897848&ymin:5189271.366866516&xmax:-5240646.301021519&ymax:12762040.633133484&spatialReference_wkid:102100).

- La collection de photographies aériennes du sud de l'Alberta (https://digitallibrary.uleth.ca/digital/collection/p22022coll2/custom/sa_aerial_map) (site en anglais uniquement) affiche les emplacements géographiques de photos aériennes verticales téléchargeables en utilisant une carte en ligne de Leaflet et le logiciel de bibliothèque numérique **CONTENTdm**. Le service de bibliothèque,archives et collections spéciales de l'Université de la Saskatchewan a créé une carte en ligne semblable qui localise les **photographies obliques** de la collection de photos aériennes de Howdy McPhai (http://mcphail.library.usask.ca/browsemap)l (site en anglais uniquement).

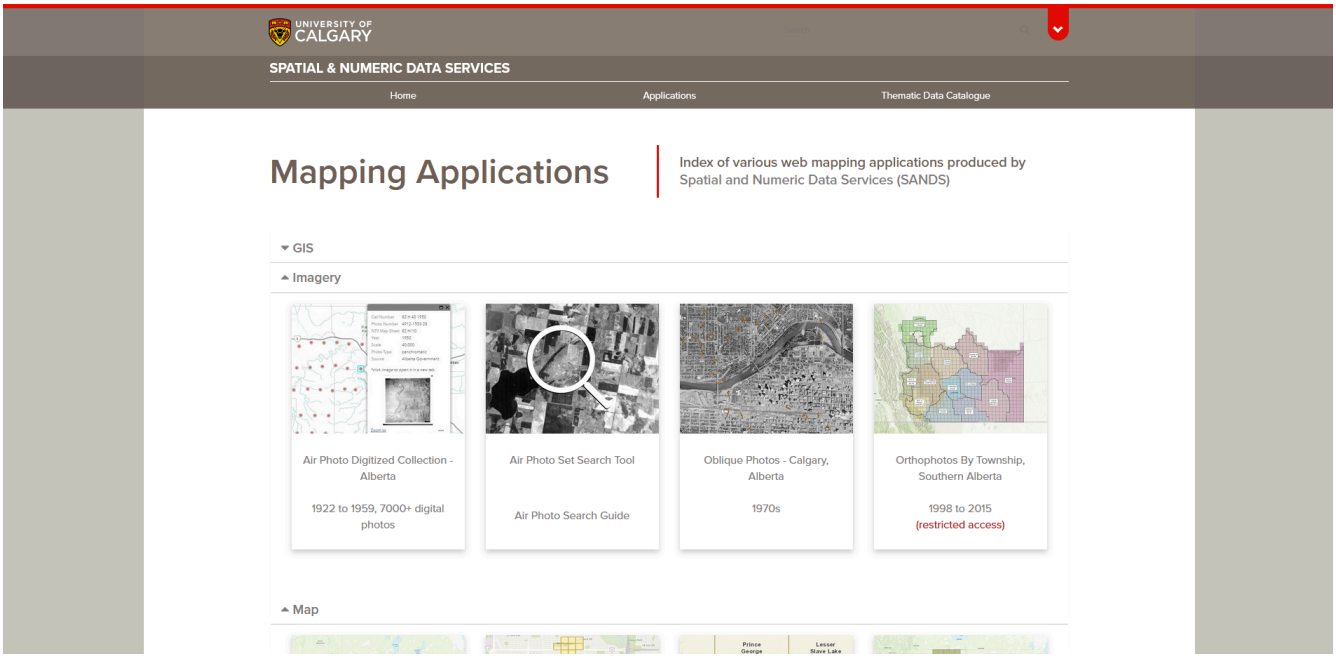



Figure 7. Applications de cartographie, Spatial & Numeric Data Services (SANDS), les bibliothèques et ressources culturelles de l'Université de Calgary.

Les bibliothèques membres du COPPUL se sont impliquées, à des degrés différents, dans la gestion et la curation des données (y compris les données géospatiales) produites par les chercheuses et chercheurs de leur université respective. Actuellement, huit des douze bibliothèques membres du COPPUL utilisent des dépôts Dataverse pour héberger et partager leurs jeux de données au nom des membres de leur communauté savante (voir le tableau 1). Sept bibliothèques utilisent Borealis (https://borealisdata.ca/fr/) comme service d'hébergement externe, tandis que l'Université du Manitoba gère sa propre instance de Dataverse. Le nombre

global de jeux de données déposés et disponibles dans les dépôts Dataverse des Prairies (un total de 1099 en date de mars 2022) est relativement modeste, mais continue de croître.

Les universités des Prairies publient aussi leurs jeux de données dans des dépôts de données propres à certains domaines (p. ex., Dryad (https://datadryad.org/stash) pour les biosciences) ou dans le DFDR (https://www.frdr-dfdr.ca/repo/?locale=fr) du Canada qui a été créé en partenariat avec l'Université de la Saskatchewan et plusieurs autres universités canadiennes. Des recherches peuvent être effectuées dans le DFDR par le biais de Lunaris qui fournit une fonctionnalité importante optimisée par Geodisy qui permet aux gens qui l'utilisent de faire une « recherche sur la carte (https://www.lunaris.ca/fr?search_field=all_fields&bbox=-144.316406%2034.597042%20-59.941406%2072.501722)» pour explorer et localiser, par le biais d'une carte en ligne, les jeux de données issus de régions particulières du Canada.

En mai 2022, les bibliothèques de l'Université du Manitoba ont lancé leur dépôt GISHub pour les données géospatiales. Le projet a été conçu à l'origine comme solution de stockage local sécurisé pour les données géospatiales, mais il a éventuellement été élargi pour y inclure des outils disponibles avec une licence d'exploitation Esri. Il vise à permettre la découverte et l'accès aux données aussi bien propriétaires qu'ouvertes de chercheuses et chercheurs, en plus de fournir un environnement local sécurisé pour une utilisation active des jeux de données géospatiales.

Pour les établissements sans instance de Dataverse, les données de recherche géospatiales créées localement peuvent être partagées dans le DFDR ou ailleurs. Par exemple, bien qu'il ne soit pas un dépôt de données, le dépôt institutionnel de l'Université de la Saskatchewan, HARVEST (https://harvest.usask.ca/), héberge un petit nombre de jeux de données de recherche géospatiales. Au fur et à mesure que les bibliothèques du COPPUL mettent en œuvre leurs stratégies de GDR pour répondre aux exigences de la **Politique des trois organismes sur la gestion des données de recherche**, il est raisonnable de s'attendre à une plus grande uniformité au niveau du comment, quand et où les données de recherche géospatiales seront partagées.

## La Colombie-Britannique

L'écosystème des données de recherche géospatiales en Colombie-Britannique est déterminé en fonction des services offerts par les établissements universitaires et organismes publics de la province. Les politiques de la Colombie-Britannique en matière de partage des données ont permis aux utilisatrices et utilisateurs de faire des recherches et d'accéder à une grande variété de données ouvertes en utilisant le BC Data Catalogue (https://catalogue.data.gov.bc.ca/) ainsi que plusieurs autres plateformes spécialisées conçues pour obtenir des données géospatiales à l'échelle de la province, notamment ParcelMap BC (https://ltsa.ca/products-services/parcelmap-bc/) produit par BC Land Title and Survey ainsi que LidarBC (https://governmentofbc.maps.arcgis.com/apps/MapSeries/index.html?appid=d06b37979b0c4709b7fcf2a1ed458e03). À un niveau plus granulaire, plusieurs districts et municipalités régionales de la Colombie-Britannique ont rendu leurs données

disponibles par le biais de plateformes plus locales de découverte des données, telles que le catalogue de données ouvertes de la ville de Surrey (https://data.surrey.ca/) et le portail de données ouvertes du Grand Vancouver (https://open-data-portal-metrovancouver.hub.arcgis.com/) (sites en anglais uniquement).

Dans le milieu universitaire de la Colombie-Britannique, les établissements postsecondaires utilisent des politiques indépendantes de collecte de données géospatiales fondées sur les exigences locales en matière d'administration, d'enseignement et de recherche. Les bibliothèques de quatre établissements sont les principaux propriétaires des collections de données géospatiales qui font partie du Abacus Data Network (https://abacus.library.ubc.ca/): l'Université Simon-Fraser, l'Université de la Colombie-Britannique (UBC), l'Université du nord de la Colombie-Britannique et l'Université de Victoria. L'infrastructure nécessaire au fonctionnement d'Abacus est maintenue par la bibliothèque de UBC. À chacune des universités membres d'Abacus est confié un sous-ensemble du réseau; la communauté de chacun des établissements est authentifiée pour n'utiliser que les données ayant les licences d'utilisation propres à son campus. Il s'agit donc d'une solution locale pour le développement des collections et la curation des données.

Entre 20% à 30% des données stockées dans Abacus sont des données géospatiales. Toutefois, le logiciel sous-jacent qui prend en charge Abacus – Dataverse – n'est pas conçu pour offrir du soutien spécialisé pour le repérage et l'utilisation des données géospatiales. Consciente de cette situation, la bibliothèque de UBC a créé un logiciel intermédiaire qui permet à Dataverse de se connecter à une pile géospécifique de logiciels libres, dont GeoServer et GeoBlacklight. Appelé Geodisy (Phase 1), (https://researchcommons.library.ubc.ca/projects/geodisy-phase-1/) ce projet a été subventionné par CANARIE entre octobre 2018 et mars 2020. À ce moment-là, une deuxième phase du projet a été entamée sous le financement de la Nouvelle organisation d'infrastructure de recherche numérique (NOIRN, désormais l'Alliance de recherche numérique du Canada ou simplement « l'Alliance ») et est administrée par le service de découverte canadien, Lunaris. Le service est maintenant utilisé pour exploiter la recherche sur la carte Geodisy (https://www.lunaris.ca/fr?search_field=all_fields&bbox=-144.316406%2034.597042%20-59.941406%2072.501722) de Lunaris.

## Les orientations futures

La gestion actuelle des données de recherche géospatiales dépend de solutions régionales, développées en fonction des besoins, avec l'aide de bibliothécaires qui travaillent à anticiper les besoins futurs. Les restrictions en matière de temps et de charge de travail impliquent que le domaine n'évolue qu'en réaction à la GDR dans son ensemble. Le domaine géospatial comporte des besoins particuliers qui nécessitent toujours des solutions créatives de gestion des données pour assurer leur utilisation actuelle et future. Les solutions à la majorité des problèmes tendent vers des initiatives partagées ou en consortium et elles semblent continuer à aller en ce sens pour l'avenir, menant possiblement à un dépôt national de données de recherche géospatiales. Des discussions concertées sur les métadonnées géospatiales seront nécessaires ainsi qu'un travail approfondi sur les

plateformes d'accès géospatiales; ces solutions seront vraisemblablement développées par le biais de méthodes régionales.

Si les défis actuels et des solutions envisagées ont été examinés, il importe de noter que certaines lacunes dans le contenu sont liées aux données biaisées. Des ateliers de cartographie autochtone, présentés par le Firelight Group, ont encouragé le développement des SIG et des données géospatiales auprès de nations autochtones. Le travail se poursuit dans les milieux universitaires sur les relations colons-autochtones, mais le développement du domaine reste lent. Les données géospatiales souffrent aussi des mêmes biais systémiques envers les communautés noires et non blanches dans la création et l'utilisation générale des données; là aussi, les avancées sont lentes. Au niveau linguistique, le Québec a fait preuve de leadership en matière d'accès multilingue aux données en s'engageant dans la traduction bilingue de métadonnées. Toutefois, les autres provinces tardent à créer et diffuser des métadonnées qui ne sont pas en anglais. Pour terminer, l'environnement canadien a longtemps favorisé le sud et bien qu'on ait tenté de faire appel à des expertises nordiques en matière de GDR géospatiales, le Grand Nord canadien demeure largement sous-exploré.

Il peut sembler banal de décrire le domaine des données de recherche géospatiales comme étant à la fois émergent et développé. Toutefois, un effort concerté est fait pour développer le travail déjà accompli et pour aligner la GDR géospatiales avec les besoins des chercheuses et chercheurs et des bibliothèques à travers le pays. Le travail se poursuit, particulièrement grâce à l'Alliance de recherche numérique du Canada (https://alliancecan.ca/fr/services/gestion-des-donnees-de-recherche/reseau-dexperts) et aux consortiums universitaires mentionnés plus tôt.

**Questions de réflexion**

1. Qu'est-ce qui fait l'unicité des données géospatiales et quels sont les impacts sur les considérations en gestion des données de recherche?
2. La gestion des données de recherche géospatiales est-elle mieux prise en charge par des établissements locaux, par des consortiums régionaux ou par l'entremise d'investissements en infrastructure à l'échelle nationale? Quels sont les avantages et les inconvénients de chacune des méthodes?
3. La gestion des données de recherche nécessite une infrastructure qui la soutient. Quelles infrastructures existent à l'heure actuelle? Quelles sont les lacunes à combler pour améliorer la préservation, l'accès et l'utilisation des données de recherche géospatiales?

**Éléments clés à retenir**

- Les données géospatiales impliquent une interaction complexe de jeux de données, mais nécessitent surtout de réfléchir aux données dans leur rapport à l'espace.
- La gestion individuelle des données géospatiales est étroitement liée à la gestion des données de recherche et des ressources existent déjà pour en apprendre davantage sur le sujet.
- Dans tout le pays, des projets régionaux tentent de gérer la préservation et l'accès aux données de recherche géospatiale dans le cadre plus large des données géospatiales.
- Les établissements postsecondaires dirigent ces projets régionaux en fonction de leur disponibilité.

## Lectures et ressources supplémentaires

L'Alliance de recherche numérique du Canada a de nombreuses ressources sur la gestion des données et les meilleures pratiques, ainsi que des groupes de discussion sur ces domaines. Pour plus d'informations, consultez le Réseau d'experts de l'Alliance numérique du Canada (https://alliancecan.ca/fr/services/gestion-des-donnees-de-recherche/reseau-dexperts) et le *Guide des pratiques exemplaires sur les métadonnées de Dataverse Nord (https://zenodo.org/record/5668962#.Ypi6LdrMKUk)*.

Un livre blanc a été rédigé par la NOIRN (maintenant intégrée à l'Alliance) faisant état des besoins actuels et futurs de l'infrastructure des données géospatiales au Canada. Ce document précise quelques-uns des besoins et particularités relatives aux données géospatiales :

> Brodeur, J., Handren, K., Berish, F., Chandler, M., Fortin, M., Leahey, A. et Stevens, R. (2020). *Enabling broad reuse of Canada's geospatial data and digitized cartographic materials. A response to the NDRIO Call for White Papers on Canada's Future DRI*. https://alliancecan.ca/sites/default/files/2022-03/final-enabling-broad-reuse-of-canadas-geospatial-data-and-digitized-cartographic-materials.pdf (https://alliancecan.ca/sites/default/files/2022-03/final-enabling-broad-reuse-of-canadas-geospatial-data-and-digitized-cartographic-materials.pdf)

Pour une introduction aux SIG, consultez le matériel de formation disponible du QGIS. (https://www.qgis.org/fr/site/forusers/trainingmaterial/index.html)

# Bibliographie

Bellin, J. (1764). *Port de Louisbourg*. J.N. Bellin.

Esri. (2016). *Que sont les données raster?* ArcMap. https://desktop.arcgis.com/fr/arcmap/10.3/manage-data/raster-and-images/what-is-raster-data.htm (https://desktop.arcgis.com/fr/arcmap/10.3/manage-data/raster-and-images/what-is-raster-data.htm)

Lunaris. (s.d.). *Dépôts sources*. https://www.lunaris.ca/fr/source_repositories (https://www.lunaris.ca/fr/source_repositories)

Marshall, P. (1999). Novanet, Inc.–Nova Scotia, Canada. *Information Technology and Libraries, 18*(3), 130-134. https://www.proquest.com/docview/215830105 (https://www.proquest.com/docview/215830105)

OpenStreetMap. (2023). *Planet OSM* [Jeu de données]. https://planet.openstreetmap.org (https://planet.openstreetmap.org)

QGIS Documentation. (s.d.). *Vector data: Overview*. https://docs.qgis.org/2.8/en/docs/gentle_gis_introduction/vector_data.html (https://docs.qgis.org/2.8/en/docs/gentle_gis_introduction/vector_data.html)

Statistiques Canada. (2019). *Fichiers des limites du recensement de 2016: Aires de diffusion*. https://www12.statcan.gc.ca/census-recensement/alternative_alternatif.cfm?l=fra&dispext=zip&teng=lda_000b16a_e.zip&k=%20%20%20%2090414&loc=http://www12.statcan.gc.ca/census-recensement/2011/geo/bound-limit/files-fichiers/2016/lda_000b16a_e.zip (https://www12.statcan.gc.ca/census-recensement/alternative_alternatif.cfm?l=fra&dispext=zip&teng=lda_000b16a_e.zip&k=%20%20%20%2090414&loc=http://www12.statcan.gc.ca/census-recensement/2011/geo/bound-limit/files-fichiers/2016/lda_000b16a_e.zip)

Stock, K. et  Hans G. (2016). Geospatial Reasoning With Open Data. Dans R. Layton et P. A. Walter (dir.), *Automating Open Source Intelligence* (p. 171–204). Syngress. https://doi.org/10.1016/B978-0-12-802916-9.00010-5 (https://doi.org/10.1016/B978-0-12-802916-9.00010-5)

ubc-library. (2022). *Geodisy*. Github. https://github.com/ubc-library/geodisy (https://github.com/ubc-library/geodisy)

# À propos des auteurs

## Martin Chandler

Martin Chandler est le bibliothécaire des services de données à l'Université du Cap-Breton. Il soutient la découverte et l'utilisation des données et des données géospatiales ainsi que les intersections créatives dans les arts et les sciences sociales.

## Kara Handren

Kara Handren est bibliothécaire de données à la bibliothèque des cartes et données de l'Université de Toronto. Elle soutient la découverte et l'analyse de données notamment par le biais de  l'exploration de textes et de données et en s'occupant des systèmes d'information géographique.

## Stéfano Biondo

Titulaire d'un baccalauréat en géographie de l'Université du Québec à Montréal et d'une maîtrise en sciences de l'information de l'Université de Montréal, Stéfano Biondo a développé une expertise en gestion et en diffusion des données géospatiales au sein des bibliothèques universitaires. À l'origine de la création du Centre GéoStat de la bibliothèque de l'Université Laval, où il occupe la fonction de cartothécaire depuis 2005, il participe à l'acquisition, à la conservation et à la mise en valeur des collections cartographiques et géospatiales.

## Amber Leahey

Amber Leahey est bibliothécaire de données et des systèmes d'information géographique (SIG) ainsi que directrice des services pour Borealis, le dépôt Dataverse canadien, un dépôt de données national sécurisé et bilingue fourni en partenariat avec les bibliothèques universitaires et les établissements de recherche à travers le Canada. Dans son rôle, elle soutient les bibliothèques, les établissements et les chercheuses et chercheurs dans la gestion, le partage, la préservation et la réutilisation des données grâce au développement continu des services de soutien en lien avec les données et la recherche à Scholars Portal et aux bibliothèques de l'Université de Toronto. Elle est titulaire d'une maîtrise en bibliothéconomie et en sciences de l'information de l'Université de Toronto.

## Sarah Rutley

Sarah Rutley est bibliothécaire des données et des systèmes d'information géographique (SIG) à l'Université de la Saskatchewan. Ses recherches portent sur la gestion, la découverte et l'accessibilité des données.

## Rhys Stevens

Rhys Stevens est bibliothécaire universitaire (bibliothécaire III) à la bibliothèque de l'Université de Lethbridge en Alberta. Il est bibliothécaire et spécialiste de l'information pour l'Alberta Gambling Research Institute et également bibliothécaire responsable de la géographie, l'archéologie, l'anthropologie, les cartes, les documents gouvernementaux et les données spatiales/numériques.

PARTIE V

# PERSPECTIVES SUR LA GESTION DES DONNÉES DE RECHERCHE

17.

# GESTION DES DONNÉES DE RECHERCHE ET MOUVEMENT DE LA SCIENCE OUVERTE: POSITIONS ET ENJEUX

Cynthia Lisée et Édith Robert

**Objectifs d'apprentissage**

À la fin de ce chapitre, vous pourrez:

- Connaître les écoles de pensée influençant les pratiques de science ouverte
- Catégoriser les grands domaines d'activité de la science ouverte
- Caractériser la présence de pratiques de gestion des données de recherche dans la science ouverte
- Remettre en question le discours dominant sur la science ouverte

**Évaluation préliminaire**

Selon vous, quelle place la GDR occupe-t-elle dans les pratiques de science ouverte ?

## Introduction

Le mouvement international en faveur de l'ouverture de la science n'est pas étranger à l'intérêt accordé par nos

décideuses et décideurs politiques aux pratiques de gestion des données de recherche (GDR) et participe à l'élaboration de nouvelles pratiques dans la conduite de la recherche. Cette effervescence autour des données de recherche invite tout un chacun à valoriser ses résultats de recherche. Il est de mise de clarifier la place de la GDR dans ce grand mouvement de la science ouverte et de soulever au passage quelques enjeux. À cette fin, en première partie nous résumerons les diverses écoles de pensée qui façonnent la science ouverte et mettrons en évidence les grands axes et principes d'élaboration de pratiques de science ouverte en signalant leur lien avec la GDR. La deuxième partie exposera quelques vertus qu'on attribue à la science ouverte tout en les liant auxdites écoles de pensée. Nous poursuivrons la contextualisation de la GDR. La dernière partie cherchera à dépasser le discours dominant résolument optimiste sur les bienfaits de la science ouverte. Nous proposons de faire deux petits pas de côté pour explorer un territoire de controverses : 1) ce que nous enseigne l'expérience historique du **libre accès** à la publication 2) ce que l'angle de la recherche qualitative révèle de la pertinence du discours positif sur le partage de données de recherche. Nous conclurons ce chapitre par une invitation aux praticiennes et praticiens de la GDR à considérer comment les composantes de ce chapitre viennent éclairer les pratiques professionnelles courantes et comment ces composantes peuvent offrir de nouvelles perspectives pour revisiter ces pratiques.

## Positionnement de la GDR dans la science ouverte

À la suite d'une analyse conceptuelle sur un corpus de 75 études sélectionnées rigoureusement, Vicente-Saez et Martinez-Fuentes (2018, p. 434) proposent la définition suivante pour la science ouverte :

« La science ouverte est un savoir transparent et accessible qui est partagé et développé au travers des réseaux de collaboration[1] » [traduction].

La qualité de transparence fait référence à la façon de présenter les résultats de l'activité scientifique d'une manière qui favorise sa réutilisation et couvre toutes les phases d'un processus de recherche scientifique. C'est un peu comme si la production des connaissances devait être menée de façon à permettre sa vérification, sa reproductibilité et un contrôle qualité par les pairs.

Le caractère accessible des connaissances implique leur diffusion rapide auprès de tous les publics, gratuitement, normalement sur le Web. Les produits de connaissance concernés sont variés : articles, opinions scientifiques, données, communication dans les conférences, manuels, codes. Le caractère repérable de ces produits constitue également une facette qui caractérise l'accessibilité.

La notion de partage est à considérer sous l'éclairage de la transparence et de l'accès : le partage englobe tant les

1. *"Open Science is transparent and accessible knowledge that is shared and developed through collaborative networks"*

étapes intermédiaires de la recherche scientifique que l'étape de publication. L'accès et la transparence soutiennent le partage. Ainsi, l'accès concerne diverses modalités, comme les personnes qui auront accès aux contenus, selon quel modèle de sécurité, par consultation in situ ou transfert de fichiers, etc. La transparence est relative à la mise à la disposition du public approprié des contenus pour des fins de reddition de compte, de validation de la recherche (p. ex., publication du protocole de recherche) et de partage de connaissances (p. ex., prépublication, rapport d'évaluation).

Finalement, le caractère collaboratif de la science ouverte engage principalement le recours à des technologies pour faciliter la collaboration entre scientifiques, mais embrasse aussi le dialogue entre nations, disciplines et rôles.

Ces clarifications ayant été établies, nous retenons cette définition de la science ouverte comme socle commun pour comprendre comment la GDR s'inscrit dans la science ouverte.

## Écoles de pensée de la science ouverte

Les pratiques de la science ouverte couvrent un large spectre de pratiques influencées par plusieurs écoles de pensée que Fecher et Friesike (2014) ont proposées pour comprendre les divers points de vue des parties prenantes : la communauté de recherche, les gens en politique, les organismes de financement, les éditeurs commerciaux, la population citoyenne. Bien que leur analyse de la documentation remonte déjà à une dizaine d'années, ce découpage est toujours d'actualité dans le domaine si l'on prend en considération les nombreuses citations dont elle fait encore l'objet. Ils résument ainsi les développements de la science ouverte en cinq écoles de pensée.

### École publique

L'école publique milite pour que la science soit accessible aux citoyennes et citoyens et que les responsables de la recherche entretiennent une conversation, voire une collaboration, avec la population. L'interaction citoyenne est possible à deux niveaux : soit par la vulgarisation du produit final afin qu'il soit compréhensible par toutes et tous, soit en rendant le processus de recherche accessible en y intégrant la citoyenne et le citoyen.

### École démocratique

L'école démocratique milite pour que les produits de la recherche comme les articles, les livres, les données de recherche, les codes logiciels soient accessibles librement et gratuitement par toutes et tous.

## École pragmatique

L'école pragmatique veut que la science soit plus efficace et mise sur le développement du travail collaboratif.

## École des infrastructures

L'école des infrastructures concentre ses efforts sur l'amélioration des technologies, si possible, **non propriétaires**, et de leur **interopérabilité** pour mieux soutenir la recherche. C'est l'idée que, grâce aux technologies, la science progressera différemment.

## École des indicateurs

Finalement, les adeptes de l'école des indicateurs cherchent à évaluer l'impact de la recherche selon d'autres normes qui s'éloignent des dérives bibliométriques actuelles (Gingras, 2014) et en tenant compte du contexte numérique dans lequel s'insère la recherche.

**Tableau 1. Exemple de pratiques de GDR selon les écoles de pensée.**

| **École de pensée** | **Exemples d'activité de GDR** |
|---|---|
| École publique | • Planification de la collecte de données par des citoyennes et citoyens. Consulter des cas de science citoyenne dans la plateforme Zooniverse (https://www.zooniverse.org/);<br>• Documentation du contexte de production des données pour permettre leur réutilisation, voire une documentation compréhensible pour des personnes utilisatrices provenant d'horizons plus variés que les chercheuses et chercheurs d'origine;<br>• Visualisation des données ou infographie qui facilite la compréhension des résultats pour les personnes qui décident ou la population. Exemple : infographie (https://sporevidencealliance.ca/wp-content/uploads/2021/10/COVIDEND_MBMC_rapidreview_VE_infographic_final.pdf) (en anglais uniquement) d'une synthèse de connaissances sur le déclin en efficacité des vaccins contre la Covid (SPOR Evidence Alliance, 2021) [2]. |
| École démocratique | • Publication de **données ouvertes** par divers paliers de gouvernement; par extension, l'ouverture complète de certaines données de recherche pour permettre aux entreprises et aux citoyennes et citoyens de réaliser des innovations ou de mieux s'informer;<br>• **Politique des trois organismes sur la gestion des données de recherche** |

2. D'autres infographies sont offertes sur le site COVID-END, *Scan Evidence Products*, https://www.mcmasterforum.org/networks/covid-end/covid-end-evidence-syntheses/scan-evidence-products (https://www.mcmasterforum.org/networks/covid-end/covid-end-evidence-syntheses/scan-evidence-products)

| | |
|---|---|
| | qui encourage le dépôt des données de recherche dans le respect des **principes FAIR**. La composante « A » (accessible) est comprise selon un spectre d'ouverture : du plus ouvert (données ouvertes) à l'accès restreint et protégé (protocoles d'entente);<br>• Intégration d'une déclaration d'accessibilité des données dans un article scientifique. Consultez, notamment, les modèles (https://authorservices.taylorandfrancis.com/data-sharing/share-your-data/data-availability-statements/) (en anglais uniquement) de Taylor & Francis. |
| École pragmatique | • Vérification de la possibilité de réutiliser des données avant de décider d'en produire de nouvelles;<br>• Reconnaissance des contributions se qualifiant pour l'auteurité sur un jeu de données et remerciement des personnes ayant contribuées qui ne se qualifient pas pour la propriété intellectuelle;<br>• Dépôt des données de recherche ou publication de leurs **métadonnées** pour faire connaître l'existence de ces données et favoriser l'établissement de nouvelles collaborations. |
| École des infrastructures | • Développement d'infrastructures de dépôt de données interopérables, gérées et soutenues par des intérêts et des fonds publics (p. ex., Borealis, le dépôt Dataverse canadien (https://borealisdata.ca/fr/));<br>• Préférence pour le recours aux formats de fichier ouverts;<br>• Recours aux puissances de calcul distribuées et autres services infonuagiques. Consultez l'offre de service (https://alliancecan.ca/fr/services/calcul-informatique-de-pointe) de l'Alliance de recherche numérique du Canada;<br>• Utilisation de **cahier de laboratoire électronique** qui facilite la collaboration et le partage des objets de recherche. Consultez le *Report of the Working Group on Electronic Lab Notebooks (https://www.enssib.fr/bibliotheque-numerique/notices/71035-report-of-the-working-group-on-electronic-lab-notebooks)*. |
| École des indicateurs | • Introduction de statistiques d'utilisation des jeux de données dans les plateformes;<br>• Citations aux jeux de données. |

Chaque école de pensée offre sa propre conjecture sur les développements de la science, ce qui mène à des travaux répondant à une variété d'objectifs. L'ensemble de ces travaux induisent des pratiques dans la conduite et l'administration de la recherche et forment ce que l'on appelle le mouvement de la science ouverte. La section suivante présente une catégorisation de ces différents travaux, ou domaines d'activités.

## Domaines d'activités de la science ouverte

Le portail Foster Open Science (https://www.fosteropenscience.eu/) est une plateforme d'apprentissage en ligne couvrant l'ensemble des thématiques relatives à la science ouverte. Il est destiné aux personnes qui souhaitent intégrer des pratiques de science ouverte à leurs processus de travail. Il est le fruit du projet

européen *Fostering the practical implementation of Open Science in Horizon 2020 and beyond* financé par Horizon 2020 de 2017 à 2019. Dans sa section *What is Open Science ? Introduction (https://web.archive.org/web/20181229190240/https:/www.fosteropenscience.eu/content/what-open-science-introduction)*, on trouve une représentation des facettes d'activités de la science ouverte conçue par Gema Bueno de la Fuente (s.d.). La science ouverte dissémine ses principes d'ouverture, de transparence, de partage et de collaboration dans des domaines d'activités couvrant l'ensemble de la démarche de recherche, de sa conception à sa diffusion. Le tableau ci-dessous résume les faits saillants de ces facettes de la science ouverte auxquels nous ajoutons la facette « Protocole de recherche ouvert » au début de la phase de conception pour mieux rendre compte des développements récents. Pour chaque facette, nous proposons l'école de pensée qui semble davantage orienter ce domaine d'activité. Nous fournissons également quelques actions de GDR pour illustrer qu'elle est bien présente dans toutes les sphères de la science ouverte.

**Tableau 2. Ubiquité des pratiques dans les domaines d'activités de la science ouverte**

| Phases de la recherche | Domaine | École de pensée | Résumé | Exemple d'action en GDR |
|---|---|---|---|---|
| Conception | Protocole ouvert | Pragmatique | Publication de la méthodologie avant de commencer la collecte de données dans un registre comme OSF Registries (http s://osf.io/registries) | Transparence pour mieux contrôler la manipulation des données au sein d'une équipe |
| | Journal de bord ouvert | Pragmatique | Gestion de tous les fichiers de données afin d'assurer la reproductibilité d'une démarche de recherche | Facilite la gestion de l'accès sécuritaire aux données en phase active |
| | Données ouvertes | Démocratique | Partage dans le respect des principes FAIR | Choix d'un dépôt FAIR |
| | Révision par les pairs ouverte | Pragmatique | On renonce complètement ou partiellement à l'anonymat des gens qui font l'évaluation et la rédaction | Accessibilité de le jeu de données pour les personnes qui font la révision avec documentation adéquate |
| | Libre accès à la publication | Démocratique | Accès immédiat, gratuit, et sans barrière technique avec attribution de licences d'utilisation | Introduire dans la publication une déclaration d'accessibilité des données |
| Diffusion | Code source ouvert | Infrastructure | Les logiciels de recherche financés par les fonds publics de même que ceux utilitaires à la recherche doivent favoriser l'autonomie technologique de l'entreprise scientifique en utilisant et en produisant du code **source ouvert** | Codes produits pour traiter et analyser les données devraient être inclus dans le partage des données |
| | Réseaux | Pragmatique | Favorise le réseautage et la | Jeux de données publiés |

| | | | | |
|---|---|---|---|---|
| | sociaux universitaires | | valorisation des résultats de la recherche | deviennent aussi des résultats de recherche à valoriser dans ses réseaux |
| | Science citoyenne | Publique | Collaboration entre les responsables de la recherche et le public en faisant participer ces derniers, possiblement à toutes les étapes de la démarche de recherche | Outre la formation des citoyennes et citoyens aux pratiques GDR, leur apport devient en soi une nouvelle source de données à prendre en compte dans la gestion du projet |
| | Ressources éducatives libres (REL) | Publique | L'accès ouvert aux connaissances scientifiques passe aussi par des pratiques éducatives qui offrent un contenu auquel tous ont accès | Les données ouvertes deviennent une REL lorsqu'utilisées en contexte pédagogique (Atenas et Havemann, 2015) |

Comme on le comprend désormais grâce aux chapitres précédents, les pratiques de GDR sont utiles tout au long des phases d'un projet de recherche. Il est intéressant de relever que les domaines d'activités spécifiques de la science ouverte sont tous interpellés également par des actions de GDR. Par ailleurs, on constate qu'aucune pratique en émergence n'est directement associée à l'école des indicateurs alors que la majorité d'entre elles sont plus fortement influencées par l'école pragmatique (quatre catégories sur neuf). Rappelons que l'école pragmatique vise essentiellement à rendre la science plus efficace, notamment en favorisant la collaboration.

## Les vertus de la GDR en contexte

On prête aux pratiques de science ouverte de nombreuses vertus, notamment ce qu'illustre la *Feuille de route pour la science ouverte (https://www.ic.gc.ca/eic/site/063.nsf/fra/h_97992.html#4)* du Canada. Le tableau ci-dessous accompagne d'une question quelques-uns des avantages des pratiques de science ouverte souvent mis de l'avant par les gens qui en font la promotion. Chaque question est une invitation à prendre le temps de réfléchir.

**Tableau 3. Les vertus de la science ouverte remises en question par le prisme de la GDR.**

| **L'ouverture de la science...** | **École de pensée dominante** | **Question sur le contexte de la GDR** |
|---|---|---|
| Facilite la reddition de compte | Pragmatique | La gouvernance autour de la politique des trois organismes en gestion de données de recherche permet-elle les suivis nécessaires à cette reddition ? |
| Accroît la reproductibilité des résultats | Pragmatique | Comment se conçoit la reproductibilité des résultats dans le cas de la recherche qualitative ? |
| Augmente la confiance du public à l'égard de la science | Publique | Comment contribuer à la littératie en matière de données (*data literacy*) des citoyennes et |

| | | |
|---|---|---|
| | | citoyens ? |
| Réduit le dédoublement des efforts | Pragmatique | Comment valoriser la reproductibilité ? |
| Accélère l'innovation | Pragmatique | Quels types de données pour quels types d'innovation ? |
| Valorise la diversité des systèmes de connaissance | Publique | Comment concrètement prendre en compte les savoirs marginalisés (p. ex., principes PCAP®) ? |
| Crée des synergies internationales et nationales | Pragmatique | Comment maintenir les spécificités locales à travers le besoin d'harmonisation ? |

Il faut éviter de voir dans les pratiques de science ouverte une panacée à des problématiques qui existent depuis toujours. Même si ces pratiques et les activités de GDR connexes participent à une certaine redéfinition des façons de faire et ouvrent la voie vers de nouvelles solutions, les réalités structurelles à la source de certains problèmes ne sont pas nécessairement prises en compte; par conséquent, ces derniers ne peuvent pas être réellement endigués. Voici quelques exemples de réflexion suscitée par les questions du tableau 3.

1. Faciliter la reddition de compte : d'après la *Feuille de route pour la science ouverte* du Canada, « [le] libre accès aux résultats de la recherche scientifique permet une plus grande reddition de comptes aux contribuables et aux bailleurs de fonds de la recherche » (Bureau du conseiller scientifique en chef du Canada, 2020). En revanche, la reddition de compte nécessite une politique de GDR forte des paliers gouvernementaux, ce dont on peut douter étant donné le peu de suivi de demandes effectué dans le cas de la politique fédérale sur le libre accès (Paquet *et al*., 2022).
2. Augmenter la confiance des citoyennes et citoyens en la science : cette confiance ne peut s'établir qu'en fournissant simplement plus de données (et plus d'articles). Il faut aussi travailler au rehaussement de la littératie de données (et informationnelle) de la population. Le risque de mettre en œuvre des pratiques de GDR en vase clos, sans arrimage avec les objectifs de science ouverte et des enjeux de littératie, est encore bien présent et pourrait nuire au potentiel d'amélioration de cette confiance.
3. Accélérer l'innovation : la science ouverte prône une pratique du partage et de réutilisation des données qui peuvent soutenir l'innovation. Voilà une proposition louable, particulièrement si elle comprend l'innovation sociale qui répondrait, selon nous, aux plus grands besoins de la société, et qui bénéficierait de données probantes alimentant les décisionnaires. Cependant, il existe des défis méthodologiques et épistémologiques à la production des **données probantes** en sciences humaines et sociales ainsi que dans le développement d'infrastructures permettant leur exploitation par les gens qui prennent les décisions. Le Canada, ainsi qu'une douzaine d'autres pays, travaille à établir des mécanismes et des flux d'information qui rendraient accessibles les données probantes aux décideuses et décideurs (Commission mondiale sur les données probantes pour relever les défis sociétaux, 2023).

# Au-delà du discours optimiste sur l'ouverture

Mirowski (2018) croit que la science ouverte prendrait ses racines dans le présent régime néolibéral de la science. Il postule que la reconfiguration de nos institutions et de la nature des connaissances est due aux impératifs du marché plutôt qu'à de réels nouveaux problèmes dans la conduite de la recherche. Pour celles et ceux qui sont moins familiers avec ce courant politique, nous vous suggérons l'article de McKeown (2022) qui énumère quelques caractéristiques de l'université néolibérale.

Dans cette section, nous proposons d'accepter d'un œil plus critique ces nouveaux développements en nous penchant sur deux situations. La première cherche à dégager un enseignement de l'évolution historique de la publication en libre accès; la deuxième, à relever des défis relatifs au partage de données en contexte de recherche qualitative, particulièrement s'il est compris selon des normes étrangères à ce type de recherche.

## Ce que nous enseigne le libre accès à la publication

Les éditeurs commerciaux ont joué un rôle non négligeable dans l'évolution des pratiques de communication savante des dernières décennies. De fait, le contexte de la récente pandémie de Covid a permis de démontrer le rôle qu'ils pouvaient jouer dans l'accès libre à la connaissance. Au cours de l'année 2020, une impressionnante augmentation de l'accessibilité aux publications scientifiques sur les coronavirus a été rapportée comparativement aux deux décennies précédentes, et ce, grâce à la coopération des éditeurs commerciaux (Belli *et al*., 2020). Il reste toutefois à voir si cette ouverture se maintiendra, car cette forte croissance du libre accès a été de type bronze, c'est-à-dire qu'une bonne partie de ces articles n'ont pas de licence garantissant la pérennité de ce libre accès. L'accessibilité à ce matériel est encore dépendante de la bonne volonté des éditeurs commerciaux.

**Tableau 4. Les types de libre accès.**

| Type de libre accès | Définition | Gratuit pour le lectorat | Gratuit pour les autrices ou auteurs |
|---|---|---|---|
| Diamant | Publication dans une revue dont le contenu est en libre accès immédiat selon divers modèles d'affaires. | X | X |
| | Initiative de publication en libre accès immédiat contrôlée par les milieux universitaires et financée par des fonds publics, des dons. | X | X |

| | | | |
|---|---|---|---|
| Hybride | Certains articles sont mis en libre accès sur paiement de frais de traitement d'article (FTA); d'autres exigent un abonnement. Les revues totalement financées par des FTA sont associées à la voie dorée. | X | Cela dépend si l'autrice ou l'auteur choisit de publier en accès libre en payant un FTA. |
| Bronze | Article rendu librement accessible sur décision de la maison d'édition, mais sans licence pérennisant cette décision d'ouverture. | X (possiblement temporaire) | X |
| Voie verte | Autoarchivage d'une des versions du manuscrit dans un dépôt. | X | X |

Plusieurs organismes de financement se sont rassemblés pour exercer une pression dans le but d'inciter fortement les éditeurs commerciaux de revues à transitionner leur modèle d'affaires vers le libre accès. Lors de la 14e conférence de Berlin sur le libre accès, en 2018 (Max Planck Digital Library), des organisations de 37 pays réparties sur les 5 continents ont fait une déclaration commune de soutien au Plan S[3]. Il s'agit d'une stratégie soutenue par un consortium d'organismes de financement, cOAlition S, qui vise à faire du libre accès aux publications une réalité. Les Fonds de recherche du Québec (FRQ) est l'un des premiers organismes de financement nord-américains à avoir rejoint cOAlition S en 2021. Ce vaste mouvement évolutif dans l'écosystème de l'économie de l'information savante signifie qu'à terme, les bibliothèques ne géreront que très peu d'abonnements. Selon toute vraisemblance, ils seront remplacés par des ententes financières avec les éditeurs commerciaux, et ce, jusqu'à payer les **frais de traitement d'articles** des chercheuses et chercheurs provenant de leurs établissements respectifs.

Fort est de constater que les éditeurs commerciaux de revues savantes tirent encore largement leur épingle du jeu, car selon une étude de 2017 rapportée par Zhang et ses collègues (2022), les frais de traitement d'article évoluent plutôt à la hausse, davantage que l'indice des prix à la consommation. Un écho familier rappelant comment l'évolution des coûts d'abonnement avait pris à la gorge les bibliothèques universitaires partout dans le monde. Case départ : comme à l'époque où les fonds publics servaient à payer des abonnements dont le coût était devenu intenable pour les bibliothèques universitaires, ces fonds vont maintenant en large partie se retrouver dans les poches des éditeurs commerciaux qui contrôlent cette augmentation – au détriment de l'essor des modèles de type diamant. Ces derniers permettent « aux scientifiques de publier en accès ouvert et sans frais » (Institut Pasteur, 2021) et sont plus cohérents avec les principes de science ouverte, car les revues sont financées par des fonds publics, des fonds d'universités, ou par des fondations. Les modèles de type

3. cOAlition S définit ainsi le Plan S (https://www.coalition-s.org/) : "le Plan S est une initiative lancée en septembre 2018 soutenant la publication en libre accès. Il repose sur cOAlition S, un consortium international d'organisations de financement de la recherche d'organisations de recherche. Le Plan S exige que les publications scientifiques qui résultent de recherches financées par des subventions publiques soient publiées en libre accès dans des revues ou sur des plateformes se conformant à certaines exigences" [traduction] (Coalition S, s.d.).

diamant sont aussi en parfaite harmonie avec les motivations initiales des premières initiatives de libre accès comme la Déclaration de Bethesda (https://www.ouvrirlascience.fr/declaration-de-bethesda-pour-ledition-en-libre-acces/) et la Déclaration de Berlin (https://openaccess.mpg.de/Berlin-Declaration) (site en anglais uniquement) en 2003 : redonner aux communautés de recherche le pouvoir sur la diffusion de leurs produits de connaissance.

De nombreuses parties prenantes participent à l'économie de l'information savante; certaines ont des intérêts corporatistes plus axés sur le profit que sur le soutien à la conduite même de la recherche. Si l'on considère l'état de balbutiement actuel du partage de données en tant qu'acte de communication savante, c'est-à-dire avec ses propres « coutumes » de publications, on peut se demander si des forces économiques similaires ne cherchent pas à en configurer les normes et usages et à en contrôler les infrastructures. L'expérience du libre accès aux publications renseignera-t-elle les nouvelles pratiques en matière de partage de données ?

## L'angle de la recherche qualitative sur le partage des données

Les chercheuses et chercheurs œuvrant dans le domaine de la recherche qualitative s'interrogent sur l'impact de la science ouverte sur les conditions de production des savoirs dans leur discipline. Ce questionnement découle à la fois de la définition souvent attribuée aux données de recherche ainsi que par la tendance observée dans plusieurs pays de favoriser la libre circulation des données de recherche. Par exemple, les *Principes et les lignes directrice de l'OCDE pour l'accès aux données de la recherche financée sur fonds publics* indiquent que :

> L'accès ouvert aux données de la recherche financée sur fonds publics et leur partage contribuent non seulement à maximiser l'impact des nouvelles technologies et des nouveaux réseaux numériques sur le potentiel de recherche, mais permettent aussi un retour plus important sur l'investissement public dans la recherche. (OCDE, 2007, p. 12)

Les organismes de financement qui encouragent le partage de données en ont souvent une définition sommaire. Les **trois organismes subventionnaires** définissent les données de recherche comme étant « des faits, des mesures, des enregistrements ou des observations recueillies par des chercheurs et d'autres personnes, assortis d'une interprétation minimale de leur contexte. » (Gouvernement du Canada, s.d.)

Nous présentons quelques explications quant à la manière dont la définition et le partage des données soulèvent des préoccupations.

### Question du contexte et reproductibilité

Comme indiqué dans le tableau 3, une des vertus attribuées à l'ouverture de la science est qu'elle accroît la

reproductibilité. Cependant, une des inquiétudes soulevées par de nombreuses parties prenantes du milieu de la recherche qualitative est l'importance du contexte avant de pouvoir envisager toute possibilité de reproductibilité des résultats d'un projet de recherche. Dans la perspective positiviste du secteur des sciences de la nature ou biomédicales, les données sont généralement considérées comme indépendantes du contexte, comme en fait foi la définition ci-dessus. En contrepartie, dans la recherche qualitative, qui utilise souvent une perspective constructiviste, le contexte est indissociable de la problématique de recherche (Hesse, 2019, p. 566). Ainsi, la question de la reproductibilité des résultats de la recherche ne se pose pas de la même façon qu'en sciences pures. Avec une telle toile de fond, comment des données issues de projets de recherche qualitatifs pourront-elles être partagées et réutilisées? Sera-t-il concrètement possible de tenir compte du contexte de production des données ?

## Mythe de la donnée brute et de la donnée neutre

En contexte de recherche qualitative, il importe d'être conscient que les données partagées auront préalablement fait l'objet d'un travail d'interprétation. Un jeu de données, peu importe la discipline, est une construction qui ne peut s'abstraire des sujets humains qui la conçoivent. Avant d'être déposé dans une plateforme, le jeu de données a fait l'objet de délibérations, de négociations et de décisions d'inclusion et d'exclusion bien ancrées dans des discours dominants, dans des réalités historico-socioéconomiques. Par conséquent, il est impossible de prétendre à la neutralité des données partagées (Neff *et al*., 2017). La documentation des jeux de données révèle particulièrement toute son importance tout en laissant ouverte la question des connaissances tacites pouvant échapper à cet effort de documentation, connaissances précieuses pour bien comprendre un jeu de données. Dans cette perspective, le partage des données se présente comme un exercice éminemment complexe.

## Type de recherche favorisé et hiérarchisation des méthodologies

Selon l'OCDE, pour qu'un jeu de données soit partageable, il doit correspondre à certains critères, dont celui d'être idéalement numériquement exploitable (OCDE, 2007). Ce caractère numérique des données peut tendre à une valorisation de l'utilisation du Big Data, puisque ces données massives, produites rapidement et dans divers formats, se multiplient et sont faciles d'accès. Cette valorisation de la recherche sur de gros corpus de données soulève le risque que les méthodologies qualitatives soient subordonnées aux méthodologies quantitatives. De plus, un glissement pourrait s'opérer pour que les techniques privilégiées par les analyses qualitatives servent uniquement à confirmer les résultats apportés par les méthodes quantitatives (Hesse *et al*., 2019). Finalement, Hesse *et al*. rapporte aussi la crainte que les recherches utilisant de petits échantillons soient moins reconnues que celles qui utilisent de gros corpus (2019).

# Conclusion : ouverture sur l'ouverture

Nous avons vu comment les activités de la GDR imprègnent les pratiques de science ouverte et avons abordé comment le discours dominant, enthousiaste et résolument optimiste sur l'adoption de ces propositions visant à ouvrir la science fait l'économie de la complexité du réel. L'espace accordé pour ce chapitre et l'objectif général de formation de ce manuel limitent le traitement approfondi des fondements idéologiques de cet appel à ouvrir la science. Il est toutefois intéressant de noter que les préoccupations suscitées par ces nouvelles pratiques ont donné naissance à un nouveau domaine d'étude, soit les études de données critiques (*critical data studies*). Ce domaine de recherche récent propose des pistes de solutions favorables à des pratiques de GDR prenant en compte des particularités disciplinaires. Plus particulièrement, puisque la conduite de la recherche qualitative commence à changer avant même que d'importants consensus disciplinaires n'aient émergé, nous sommes d'avis qu'éclairer la pratique professionnelle en GDR par ce courant des études de données critiques outillera ses praticiennes et praticiens à éviter d'acculturer ces communautés de recherche.

Une approche critique ou sociopolitique pour lire les développements de la science ouverte permettrait de faire plus facilement un pas de côté pour éclairer différemment le discours enthousiaste autour du mouvement de la science ouverte et de ses pratiques. Nous sommes ravies de conclure ce chapitre avec une invitation aux praticiennes et praticiens de la GDR à responsabiliser leurs pratiques professionnelles en creusant les discours, en écoutant les diverses voix qui s'expriment dans ce vaste mouvement de la science ouverte pour ainsi tenter d'élaborer des réponses potentielles aux questions suivantes : quel(s) systèmes économiques et politiques produisent les structures sociales, valeurs, normes, idéologies, biens et produits financiers? Pour qui? Avec quelle technologie? Pourquoi celle-là? Où se trouvent les infrastructures de science ouverte ? Dans les faits, qui bénéficie de l'ouverture de la science ?

**Questions de réflexion**

1. Comparez la définition de la science ouverte du portail Foster Open Science avec celle proposée dans ce chapitre. Quelles différences et ressemblances pouvez-vous relever ?
   Définition de Foster Open Science :

   > La science ouverte est la pratique de la science de manière à ce que d'autres puissent collaborer et contribuer, où les données de recherche, les notes de laboratoire et d'autres processus de recherche sont librement accessibles, dans des conditions qui permettent la

réutilisation, la redistribution et la reproduction de la recherche et de ses données et méthodes sous-jacentes[4] [traduction].

2. Par quelle(s) école(s) de pensée pensez-vous que la GDR est surtout traversée ?
3. Vrai ou faux : Considérant comment s'est développé la publication en libre accès, il n'y a aucune raison de craindre que quelques entreprises avec des intérêts commerciaux bâtissent un oligopole sur des produits facilitant l'exploitation des données de recherche.
4. Pourquoi la question de la reproductibilité des résultats de la recherche ne se pose pas de la même façon en recherche qualitative qu'en contexte de recherche en sciences pures ?
5. Quel nouveau domaine de recherche vous permettrait de repérer des points de vue plus critiques sur les pratiques de GDR?

Voir le solutionnaire pour les réponses.

**Éléments clés à retenir**

- Cinq écoles de pensée façonnent les pratiques de science ouverte : école publique, école démocratique, école pragmatique, école des infrastructures, école des indicateurs.
- Les pratiques de science ouverte peuvent se catégoriser en neuf grands secteurs d'activités touchant l'ensemble des étapes d'un projet de recherche, de sa conception à sa diffusion : ouverture des protocoles de recherche, utilisation de journaux de bord électroniques, données ouvertes, ouverture du processus de révision par les pairs, libre accès à la publication, code source ouvert, réseaux sociaux scientifiques, science citoyenne et ressources éducatives libres.
- Les fonds publics sont encore largement attribués aux éditeurs commerciaux dans la configuration du libre accès à la publication et la question demeure, à savoir si les

4. "*Open Science is the practice of science in such a way that others can collaborate and contribute, where research data, lab notes and other research processes are freely available, under terms that enable reuse, redistribution and reproduction of the research and its underlying data and methods.*" https://www.fosteropenscience.eu/foster-taxonomy/open-science-definition (https://www.fosteropenscience.eu/foster-taxonomy/open-science-definition)

infrastructures de science ouverte, actuelles et futures, pouvaient être soumises au même risque oligopolistique.
- Les pratiques d'ouverture et de partage des données de recherche présentent des enjeux épistémologiques particuliers dans les domaines des sciences humaines et sociales et dans les méthodologies de recherche qualitative : la complexité de partager des données de recherche qualitative, la hiérarchisation des méthodologies de recherche, l'impossible neutralité des données.

## Lectures et ressources supplémentaires

- Portail Foster Open Science (https://www.fosteropenscience.eu/), une plateforme d'apprentissage en ligne couvrant l'ensemble des thématiques relatives à la science ouverte
- Iwasiński, Łukasz. (2020). Theoretical Bases of Critical Data Studies. *Teoretyczne podstawy critical data studies, 115A*(1A), 96-109.

## Bibliographie

Belli, S., Mugnaini, R., Baltà, J. et Abadal, E. (2020). Coronavirus mapping in scientific publications: When science advances rapidly and collectively, is access to this knowledge open to society? *Scientometrics, 124*(3), 2661-2685. https://doi.org/10.1007/s11192-020-03590-7 (https://doi.org/10.1007/s11192-020-03590-7)

Bueno de la Fuente, G. (s.d.). *What is Open Science? Introduction.* Foster. https://web.archive.org/web/20181229190240/https:/www.fosteropenscience.eu/content/what-open-science-introduction (https://web.archive.org/web/20181229190240/https:/www.fosteropenscience.eu/content/what-open-science-introduction)

Bureau du conseiller scientifique en chef du Canada. (2020). *Feuille de route pour la science ouverte*. Gouvernement du Canada. https://science.gc.ca/site/science/fr/bureau-conseillere-scientifique-chef/science-ouverte/feuille-route-pour-science-ouverte (https://science.gc.ca/site/science/fr/bureau-conseillere-scientifique-chef/science-ouverte/feuille-route-pour-science-ouverte)

Coalition S. (s.d.) *About Plan S*. https://www.coalition-s.org/ (https://www.coalition-s.org/)

Commission mondiale sur les données probantes pour relever les défis sociétaux (2023). *Renforcer les systèmes*

*nationaux d'appui aux preuves*. https://www.mcmasterforum.org/networks/evidence-commission/domestic-evidence-support-systems (https://www.mcmasterforum.org/networks/evidence-commission/domestic-evidence-support-systems)

Fecher, B. et Friesike, S. (2014). Open Science: One Term, Five Schools of Thought. Dans Bartling S. et Friesike S. (dir.) *Opening Science: The Evolving Guide on How the Internet is Changing Research, Collaboration and Scholarly Publishing* (pp. 17-47). Springer. https://doi.org/10.1007/978-3-319-00026-8_2 (https://doi.org/10.1007/978-3-319-00026-8_2)

Gingras, Y. (2014). *Les dérives de l'évaluation de la recherche. Du bon usage de la bibliométrie*. Raisons d'agir.

Gouvernement du Canada. (s.d.). *Politique des trois organismes sur la gestion des données de recherche – Foire aux questions*. https://science.gc.ca/site/science/fr/financement-interorganismes-recherche/politiques-lignes-directrices/gestion-donnees-recherche/politique-trois-organismes-gestion-donnees-recherche-foire-aux-questions#1a (https://science.gc.ca/site/science/fr/financement-interorganismes-recherche/politiques-lignes-directrices/gestion-donnees-recherche/politique-trois-organismes-gestion-donnees-recherche-foire-aux-questions#1a)

Hesse, A., Glenna, L., Hinrichs, C., Chiles, R. et Sachs, C. (2019). Qualitative Research Ethics in the Big Data Era. *American Behavioral Scientist, 63*(5), 560–583. https://doi.org/10.1177/0002764218805806 (https://doi.org/10.1177/0002764218805806)

Institut Pasteur. (2021, 23 avril). La voie diamant de l'Open Access. *Open science: évolutions, enjeux et pratiques*. https://openscience.pasteur.fr/2021/04/23/la-voie-diamant-de-lopen-access/ (https://openscience.pasteur.fr/2021/04/23/la-voie-diamant-de-lopen-access/)

McKeown, M. (2022). The View from Below: How the Neoliberal Academy Is Shaping Contemporary Political Theory. *Society, 59*(2), 99-109. https://doi.org/10.1007/s12115-022-00705-z (https://doi.org/10.1007/s12115-022-00705-z)

Mirowski, P. (2018). The future(s) of open science. *Social Studies of Science, 48*(2), 171203. https://doi.org/10.1177/0306312718772086 (https://doi.org/10.1177/0306312718772086)

Neff G., Tanweer A., Fiore-Gartland B. et Osburn L. (2017). Critique and Contribute: A Practice-Based Framework for Improving Critical Data Studies and Data Science. *Big Data. 5*(2), 85-97. https://doi.org/10.1089/big.2016.0050 (https://doi.org/10.1089/big.2016.0050)

OCDE – Organisation de coopération et de développement économique. (2007). *Principes et lignes directrices pour l'accès aux données de la recherche financée sur fonds publics*. https://www.oecd.org/fr/science/inno/38500823.pdf (https://www.oecd.org/fr/science/inno/38500823.pdf)

Paquet, V., van Bellen, S. et Larivière, V. (2022). Measuring the prevalence of open access in Canada: A national comparison. *The Canadian Journal of Information and Library Science / La Revue canadienne des sciences de l'information et de bibliothéconomie, 45*(1), 1–21. https://doi.org/10.5206/cjilsrcsib.v45i1.14149 (https://doi.org/10.5206/cjilsrcsib.v45i1.14149)

SPOR Evidence Alliance. (2021). *Vaccine Effectiveness Over Time in Vaccinated Individuals: A Living Review*. https://sporevidencealliance.ca/wp-content/uploads/2021/10/COVIDEND_MBMC_rapidreview_VE_infographic_final.pdf (https://sporevidencealliance.ca/wp-content/uploads/2021/10/COVIDEND_MBMC_rapidreview_VE_infographic_final.pdf)

Vicente-Saez, R. et Martinez-Fuentes, C. (2018). Open Science now: A systematic literature review for an integrated definition. *Journal of business research, 88*, 428-436. https://doi.org/10.1016/j.jbusres.2017.12.043 (https://doi.org/10.1016/j.jbusres.2017.12.043)

Zhang, L., Wei, Y., Huang, Y. et Sivertsen, G. (2022). Should open access lead to closed research? The trends towards paying to perform research. *Scientometrics, 127*, 7653–7679. https://doi.org/10.1007/s11192-022-04407-5 (https://doi.org/10.1007/s11192-022-04407-5)

## À propos des auteurs

### Cynthia Lisée

Cynthia Lisée est impliquée dans le dossier GDR de l'Université du Québec à Montréal (UQAM) depuis 2018 et elle agit à titre de bibliothécaire en soutien à la recherche depuis 2020. Elle participe ou a participé à divers groupes de travail GDR (Alliance de recherche numérique, Partenariat des bibliothèques universitaires du Québec et UQAM). À l'UQAM, elle fait notamment partie de l'équipe pilotant la mise en œuvre de la stratégie institutionnelle GDR. Le soutien aux revues savantes fait également partie de son portefeuille de dossiers. Elle détient un baccalauréat en physique et a exploré la géologie. Elle détient également d'autres diplômes universitaires et scolarité en sciences humaines et sociales, notamment la maîtrise en sciences de l'information de l'Université de Montréal. ORCID : 0000-0003-3883-3676 (https://orcid.org/0000-0003-3883-3676)

### Édith Robert

Édith Robert est bibliothécaire à l'École des sciences de la gestion à l'Université du Québec à Montréal. Titulaire d'une maîtrise en sociologie, elle a travaillé de nombreuses années dans différents centres de recherche. Cumulant ses intérêts pour la profession de bibliothécaire universitaire avec les approches théoriques développées par la sociologie des sciences, elle s'intéresse aux enjeux de la communication savante

et à la question des savoirs «marginalisés», dans le développement des collections. Elle est également membre du service conseil de l'ACFAS et enseigne au Collège Rosemont dans le programme Techniques de recherche et gestion de données.

18.

# UNE PERSPECTIVE PRATIQUE SUR LE DOMAINE ÉVOLUTIF DE LA GESTION DES DONNÉES DE RECHERCHE

Dr. Joel T. Minion

**Objectifs d'apprentissage**

À la fin de ce chapitre, vous pourrez :

1. Comprendre les facteurs au cœur du développement de la gestion des données de recherche.
2. Définir les rôles et responsabilités de divers groupes qui participent à la gestion des données de recherche.
3. Comprendre dans quelle mesure la gestion des données de recherche, les méthodes de recherche et les types de données continuent d'évoluer, et ce, tant ici qu'à l'international.
4. Formuler une stratégie élémentaire pour gérer un jeu de données de recherche particulier.

## Introduction

Comme vous le savez, la gestion et la supervision systématiques des **données de recherche** sont en train de devenir des compétences essentielles pour les chercheuses et chercheurs des établissements d'enseignement supérieur au Canada et de partout dans le monde, tout comme pour les bibliothécaires et autres spécialistes des données qui les soutiennent. Si les progrès au chapitre de la gestion des divers types de données profitent à la recherche en général à long terme, ce changement ne cesse de soulever des inquiétudes pratiques chez les personnes responsables de la **gestion des données de recherche** (GDR). Un des principaux défis de la GDR : il s'agit d'un domaine de pratique émergent. Les attentes en matière de gestion des données varient selon le

type de données, le domaine d'étude, l'établissement, la source de financement et le ressort territorial. Les initiatives pour gérer et partager les données génomiques, par exemple, sont plus avancées que celles relatives aux données ethnographiques. De même, chaque pays n'accorde pas la même urgence à mettre en œuvre des stratégies pour faire progresser la GDR.

Par conséquent, les chercheuses et chercheurs et autres spécialistes ayant besoin de gérer des données ne disposent pas toujours d'une voie claire à suivre avec des indications fiables; dans certains cas, il s'agit même de défricher le terrain. Tous les travaux en GDR ont ceci en commun qu'ils doivent réfléchir de manière critique aux tâches à réaliser. Aucune approche unique ne fonctionnera à grande échelle. Ce chapitre a pour but de vous aider à développer une perspective critique à l'égard de la gestion des données de recherche, quel que soit votre rôle dans le processus. Comme vous le constaterez, la capacité à réfléchir à ces défis en lien avec la GDR exige un ensemble de compétences dans plusieurs domaines : se familiariser avec les éléments complexes des données de recherche; mettre en application les approches actuelles dans de nouveaux contextes; déterminer où et quand chercher des réseaux de pratique externes pour obtenir du soutien et affiner votre ingéniosité et créativité.

Il est essentiel de retenir que le travail de gestion des données relève à la fois de l'art et de la science. S'il existe plusieurs principes et pratiques pour vous guider, la GDR n'est souvent rien de plus que des chercheuses et chercheurs qui commencent en GDR, dont le temps est limité, qui tentent de traiter leurs données selon leur niveau de connaissance.

La réflexion repose sur trois questions.

- *Pourquoi insister sur la GDR?* Cette partie examine ce qui stimule les nouvelles exigences pour gérer les données de manière systématique et pourquoi la réponse est importante pour vous;
- *À qui la responsabilité?* Le travail de GDR comprend divers groupes. L'expertise et les responsabilités des gens ont une incidence sur la manière dont le travail est effectué et comment le soutien est offert;
- *Où se trouve l'avant-garde?* Parce que la GDR n'en est encore qu'à ses débuts, nos efforts doivent être encadrés par la pratique actuelle tout en restant à l'affût des changements en cours, au Canada et ailleurs.

Fort de ces questions, nous conclurons le chapitre avec des étapes pratiques à prendre en considération lorsque nous gérons des données de recherche, quel que soit le projet. Ensemble, questions et étapes devraient améliorer vos compétences en résolution de problèmes et optimiser votre capacité à réaliser des travaux en lien avec la GDR.

# Pourquoi insister sur la GDR?

Si vous êtes nouvelle ou nouveau dans le monde des données de recherche, vous aurez peut-être la surprise d'apprendre à quel point l'approche de la GDR, systématique et contrôlée de l'extérieur de la recherche, est novatrice. Généralement, les chercheuses et chercheurs, en collaboration avec leur établissement, sont responsables de la gestion des données de recherche : la manière de les organiser, de les archiver et la décision de les partager. Souvent, les données de recherche sont perçues comme étant la propriété de la chercheuse ou du chercheur qui les ont créées. Elles sont le produit d'un investissement substantiel en temps, en effort personnel, en formation et en perfectionnement professionnel de la part de la chercheuse ou du chercheur. Les données constituent la pierre angulaire de sa carrière et la base de publications scientifiques. Elles ont peut-être été partagées avec des collègues proches, mais il n'y a que peu d'incitatifs, et encore moins d'exigences, à organiser les données selon des normes externes ou à les rendre accessibles à autrui.

Dans ces circonstances, il y a eu très peu de motivation à adopter une approche plus systématique de la GDR. Qu'est-ce qui a changé? Jusqu'à un certain point, la culture a changé dans diverses communautés de recherche afin de reconnaître l'impact possible de la collaboration sur la progression des disciplines et la production de connaissances. Si cette évolution se poursuit (plus rapidement dans certains domaines que dans d'autres), elle n'explique pas, à elle seule, pourquoi des concepts comme les **principes FAIR** et des outils comme les **plans de gestion de données** (PGD) ont vu le jour. Deux facteurs ont été particulièrement percutants : 1) les attentes qui évoluent de la part des sources de financement, 2) les progrès technologiques et la puissance des données massives.

## Attentes en évolution

Pendant plus d'une décennie, les principaux organismes de financement (p. ex., les **trois organismes fédéraux de financement de la recherche au Canada** et des organismes semblables ailleurs dans le monde) ont commencé à optimiser la production de connaissances issues des recherches qu'ils soutiennent. Ils exigent des données bien organisées et (idéalement) ouvertes, et ce, pour plusieurs raisons. D'abord, des données bien gérées réduisent le dédoublement, car elles permettent à la communauté de recherche de cerner les études déjà réalisées sur un sujet. La GDR donne un accès plus complet aux données de recherche, bien au-delà de ce qui est inclus dans les articles ou les livres qu'une chercheuse ou un chercheur choisit de publier (ou peut publier). Ensuite, un meilleur accès aux données historiques signifie de meilleures occasions de réaliser une **analyse secondaire**, ce qui maximise les résultats de recherche pour chaque dollar, peso, livre, euro investi. Enfin, les organismes de financement conviennent qu'une meilleure gestion des données et une plus grande ouverture protègent la robustesse et la transparence de la recherche qu'ils appuient (Pinfield *et al*., 2014).

**Exercice : la montée des plans de gestion des données (PGD)**

Comme vous l'avez appris, l'obligation pour les chercheuses et chercheurs de soumettre des plans de gestion des données avec leurs demandes de subvention devient de plus en plus fréquente au Canada. Dans d'autres pays, certains organismes de financement exigent un PGD depuis plus d'une décennie. Effectuez des recherches en ligne pour trouver les références les plus récentes en matière de PGD (soit des exemples de plans, soit des appels à les rendre obligatoires). Après avoir examiné quelques exemples, réfléchissez de manière critique aux types de données, aux domaines d'étude, aux pays/organismes de financement en question. Que découvrez-vous au sujet de la montée des PGD?

## Progrès technologiques

Les progrès en matière de technologies informatiques stimulent, eux aussi, la GDR. Pensons notamment à la capacité à stocker des jeux de données massifs et de travailler avec eux, l'arrivée de l'infonuagique et du partage de données par Internet ainsi que la réduction des coûts des technologies informatiques. À l'origine, de telles améliorations constituaient les principaux avantages pour les domaines qui travaillent avec des données massives (p. ex., l'astronomie, la génomique, la cartographie géospatiale), ce qui explique, en partie, pourquoi la GDR a progressé plus rapidement dans certaines disciplines que dans d'autres (la nature des données en question – quantitative – constitue un autre facteur). De tels progrès technologiques ont façonné les possibilités dans d'autres domaines d'étude, comme la numérisation des ressources en sciences humaines et la capacité d'analyser les données des médias sociaux. Des améliorations apportées aux logiciels d'analyse ont également permis de relier des données de recherche de manière sécuritaire à d'autres types de données (p. ex., les dossiers médicaux, les sources météorologiques) pour créer d'encore plus gros jeux de données.

## Autres facteurs

Bien entendu, d'autres facteurs stimulent également la GDR. Une gestion plus cohérente des données rend le processus de recherche plus efficace et peut mener à des résultats plus solides. Comme mentionné dans le chapitre sur la GDR et la recherche qualitative, une meilleure organisation des données d'entrevues améliore l'analyse, car il devient possible d'établir plus facilement des liens dans de grands jeux de transcription. De plus

en plus, la recherche est transdisciplinaire, ce qui signifie qu'elle dépasse les limites et méthodologies épistémologiques, réunissant ainsi des groupes variés de chercheuses et chercheurs. La GDR soutient de tels efforts et facilite la collaboration. Enfin, quelques universitaires en fin de carrière veulent laisser à la postérité des produits fondés sur des données qui expliquent ces dernières au-delà de ce qui est inclus dans les **métadonnées** habituelles, comme pourquoi et comment une méthodologie ou une théorie en particulier a été choisie pour générer les données. Une pratique améliorée de GDR permet également aux chercheuses et chercheurs d'expérience de relier des données provenant d'études connexes, parfois à travers des décennies. Une meilleure gestion (particulièrement de la documentation) fait en sorte que l'utilisation future de telles données respecte ce qui se peut – et ne se peut pas – en matière d'analyse secondaire.

Reconnaître les moteurs de la GDR nous aide à comprendre pourquoi la gestion des données est importante et quels sont nos propres objectifs en matière de données. Si vous êtes chercheuse ou chercheur, votre priorité est-elle simplement de satisfaire aux exigences de GDR de votre organisme de financement? Ou cherchez-vous également à définir un programme de recherche exhaustif au fil du temps? Si vous êtes bibliothécaire de données et aidez les chercheuses et chercheurs à préparer leurs données pour les verser dans un dépôt, que devez-vous savoir au sujet des différentes attentes disciplinaires en GDR? Quel est le niveau de service que vous voulez offrir? Plusieurs raisons motivent la GDR et plusieurs degrés de gestion sont possibles. Par conséquent, il est essentiel que la stratégie de GDR et la finalité d'un projet précis soient harmonisées avec des facteurs contextuels plus vastes.

## Qui est responsable?

La volonté d'adopter davantage des approches systématiques en GDR s'accompagne d'une interrogation sur la responsabilité des gens impliqués. Qui organise les données? Comment? Qui décide des normes de métadonnées à respecter? Qui choisit le dépôt? La liste des tâches et des décisions est longue. En principe, la chercheuse ou le chercheur ayant le plus d'ancienneté est responsable de gérer les données, nommément la chercheuse principale ou le chercheur principal. Dans la pratique, la personne responsable délègue régulièrement les travaux en lien avec la GDR (p. ex., la collecte de données, le nettoyage, l'organisation, l'archivage) à d'autres membres de son équipe, surtout aux chercheuses ou chercheurs au postdoctorat et les associées ou associés de recherche. C'est là que s'effectue la plus grande partie de la gestion des données.

La délégation de la responsabilité s'accompagne d'au moins deux complications. La première, c'est que les personnes les plus près des données et qui, souvent, les connaissent le mieux, sont employées en vertu de contrats à court terme. Lorsqu'elles passent à autre chose (comme c'est souvent le cas), leurs connaissances partent avec elles à moins que des mesures aient été prises pour documenter celles-ci le mieux possible. Malheureusement, ceci ne se produit pas toujours, ce qui a des répercussions sur la gestion efficace et uniforme des données pour la durée d'une étude. La deuxième, c'est que selon l'expérience et la formation des

membres de l'équipe, il s'agit soit d'adeptes de la GDR qui requièrent peu de surveillance, soit de novices en matière de principes et de bonnes pratiques en GDR. Dans ce dernier cas, ces personnes auront besoin d'un suivi serré, d'une formation efficace et de soutien en gestion des données par des spécialistes ne faisant pas partie de l'équipe de recherche.

Sur la scène internationale, deux modèles sont apparus pour offrir des services de soutien en GDR : celui dirigé par les bibliothécaires puis celui mené par les chercheuses et chercheurs. Tous deux tentent d'améliorer les compétences des chercheuses et chercheurs à tous les paliers et à faciliter la gestion des données conformément aux attentes des organismes subventionnaires, aux exigences des périodiques, et à l'évolution des pratiques spécifiques à chaque discipline. Ce qui distingue principalement les deux modèles? La personne qui offre son soutien.

L'approche de la GDR menée par les bibliothécaires est courante en Amérique du Nord. La responsabilité des services de GDR incombe alors aux bibliothèques universitaires où les bibliothécaires de données, entre autres spécialistes, aident à former et à soutenir les chercheuses et chercheurs dans la gestion de leurs données de recherche et participent à la stratégie de GDR à l'échelle de l'établissement.

L'approche de la GDR menée par les chercheuses et chercheurs est courante en Europe où les services de GDR sont affectés aux nouvelles divisions créées à cette fin au sein des universités. Ces bureaux se trouvent dans la bibliothèque universitaire, sans nécessairement en faire partie. Ceci signifie que les services de GDR sont mis sur pied et gérés indépendamment des services de la bibliothèque. Généralement, les travaux de soutien à la GDR sont assurés par la communauté étudiante des cycles supérieurs (avec un doctorat ou, à tout le moins, une maîtrise fondée sur la recherche).

**Exercice : qui est embauché?**

Ces deux modèles sont très bien illustrés par les offres d'emploi publiées. En règle générale, les postes en GDR en Amérique du Nord exigent des compétences distinctes de celles exigées en Europe. Les trois listes de diffusion ci-dessous comprennent régulièrement des postes connexes à la GDR. Envisagez de vous y abonner pour suivre les discussions qui s'y déroulent et comparer les emplois afin d'examiner les compétences exigées des candidats (elles sont également profitables si vous vous intéressez davantage à la GDR en général).

Liste de diffusion CANLIB-DATA (principalement en anglais) (Canada et États-Unis) : https://researchdata.library.ubc.ca/learn/canlib-data-listserv/ (https://researchdata.library.ubc.ca/learn/canlib-data-listserv/)

Liste de diffusion RESEARCH-DATAMAN (en anglais uniquement) (R.-U./UE) : https://www.jiscmail.ac.uk/cgi-bin/webadmin?A0=RESEARCH-DATAMAN (https://www.jiscmail.ac.uk/cgi-bin/webadmin?A0=RESEARCH-DATAMAN)

Liste de diffusion Forum-DataQc (en français) (Canada) : https://mailman.uquebec.ca/mailman/listinfo/forum-dataqc (https://mailman.uquebec.ca/mailman/listinfo/forum-dataqc)

Chaque modèle a ses forces et ses limites. Les bibliothécaires universitaires ont de l'expertise en gestion de l'information et, en Amérique du Nord, ont généralement une qualification commune (c'est-à-dire, diplôme de MLIS avec agrément d'ALA). Par conséquent, les bibliothécaires ont une base commune en ce qui a trait aux principes et aux pratiques de gestion de l'information. Si quelques bibliothécaires universitaires effectuent de la recherche ou détiennent un doctorat, leur rôle professionnel principal est de soutenir la recherche, ce qui signifie une expérience limitée dans la réalisation de projets de recherche de plus grande envergure ou dans la collecte et l'analyse de données complexes.

Les chercheuses et chercheurs développent une expertise en données en progressant vers l'obtention de leur doctorat. Tout au long de leur carrière, les personnes qui font de la recherche passent des années ancrées dans des cultures de recherche particulières, à travailler directement avec des données. Du même souffle, ces mêmes chercheuses et chercheurs ont probablement moins de compétences au chapitre des manières normalisées de gérer et d'organiser l'information et les données. Il n'est pas rare dans le milieu de la recherche que des systèmes singuliers qui fonctionnent parfaitement pour un individu ou une équipe soient créés.

### Étude de cas : le programme en GDR à la TU Delft, aux Pays-Bas

TU Delft est la plus grande université technique aux Pays-Bas. Son programme en GDR compte parmi les plus avancés et les plus créatifs. Elle offre un contraste intéressant aux services élaborés par les universités canadiennes. Lancé en 2018, le programme de Delft repose sur deux principes centraux : 1) les chercheuses et chercheurs se trouvent au cœur de la science ouverte; 2) les **personnes responsables de l'intendance des données** peuvent avoir un rôle-conseil pour améliorer la culture et la pratique de GDR à l'échelle de l'établissement. D'entrée de jeu, le programme a pour objectif d'améliorer la culture de gestion des données et non sa conformité.

L'approche de Delft se distingue de plusieurs manières. D'abord, elle affecte une personne

responsable de l'intendance des données à chaque faculté; la GDR est présente là où s'effectue la recherche plutôt que d'attendre que les chercheuses et chercheurs fassent appel à ses services. Ainsi, les personnes responsables de l'intendance de données sont en bonne posture pour évaluer ce qui se produit sur le terrain. Ensuite, en règle générale, ces gens ont un doctorat, ce qui sous-entend des compétences avancées en recherche et, souvent, de l'expérience en la matière. Enfin, le programme a été mis sur pied en tant qu'initiative d'apprentissage actif qui investit temps et énergie à analyser ses services et à rapporter ses principales conclusions dans des revues et lors de congrès.

Pour comprendre l'approche de Delft, visitez leur site Web (https://www.tudelft.nl/en/library/research-data-management) (en anglais uniquement) et découvrez-en davantage sur le rôle de la personne responsable de l'intendance des données ainsi que sur les gens embauchés pour le jouer.

Plomp, E., Dintzner, N. J. R., Teperek, M. et Dunning, A. (2019). Cultural obstacles to research data management and sharing at TU Delft. *Insights, 32*(1). https://doi.org/10.1629/uksg.484 (https://doi.org/10.1629/uksg.484)

Les différents modèles et leur efficacité sont peu étudiés dans la littérature. C'est probablement le reflet d'une élaboration et d'une intégration des services de GDR encore en cours dans les structures et cultures universitaires. Cependant, les modèles indiquent que la GDR exige la participation de plusieurs groupes. Si la responsabilité ultime de la GDR repose sur les épaules des chercheuses principales ou chercheurs principaux, la gestion courante des données ainsi que la supervision et les services de soutien est l'affaire d'autres personnes. Il est important de souligner que les chercheuses et chercheurs, bibliothécaires et autres spécialistes des données apportent une expertise particulière et leurs perspectives quant à la manière dont les données peuvent être gérées et la forme que devrait prendre la GDR à l'avenir.

## Où se trouve l'avant-garde?

Bien que la GDR soit un phénomène relativement récent, il est important de rappeler que la recherche et les données évoluent. De nouveaux sujets à découvrir et de nouvelles technologies de recherche ne cessent de voir le jour (p. ex., la crise des opioïdes, l'édition de gènes), tout comme les nouveaux types de données et les manières de les analyser (p. ex., les données des médias sociaux, l'analyse augmentée). De tels progrès permettent la réalisation de recherches qu'il aurait été impossible de mener il y a à peine quelques années. Si le rythme de ce changement varie au fil du temps et des domaines d'étude, il a une incidence sur la manière dont nous abordons la GDR, y compris les services de soutien disponibles et la manière dont ils sont offerts. Cette

section met en évidence deux exemples de l'importance de réfléchir de manière critique à la pratique actuelle et de rester à l'affût des progrès réalisés ailleurs.

Le premier exemple porte sur l'utilisation fréquente en formation sur la GDR d'infographie du cycle de vie pour illustrer le processus de recherche et la gestion des données qu'elle contient (voir le chapitre 1). De telles images ont pour but de jeter une lumière sur les étapes standards en recherche, de la planification initiale à l'archivage et à la réutilisation des données. Les modèles de cycle de vie sont efficaces, car ils sont accessibles; toutefois, une telle représentation ne concorde pas toujours bien à la manière dont certaines formes de recherche se déroulent. Ceci peut mener à des compréhensions erronées du fonctionnement – réel ou potentiel – de la GDR. Par exemple, la plupart des données des **sciences sociales** sont recueillies de manière itérative, ce qui signifie que les chercheuses et chercheurs entreprennent une réflexion en temps réel et une modification méthodologique tout au long du processus de collecte de données. Par exemple, un sociologue peut ajouter de nouvelles questions ou de nouvelles personnes participantes peuvent se joindre à des groupes de discussion. De telles études ne se déroulent pas de la même manière que la recherche en laboratoire.

Malgré la circularité de l'image, les modèles de cycle de vie sont bizarrement linéaires et supposent un processus du début à la fin qui ne se superpose pas bien à certaines méthodologies. De tels modèles ne mettent pas en évidence l'importance des relations dans la collecte des données entre des études connexes ou sur des périodes données (p. ex., en recherche longitudinale). Ils peinent à représenter comment les données historiques sont utilisées de plus en plus pour générer de nouvelles données par le biais d'analyses secondaires et de liens entre les données; le résultat ainsi obtenu accroît la capacité d'effectuer davantage de recherche. Le défi en matière de pratique et de prestation de service en GDR consiste à suivre l'évolution des méthodes de recherche  et des types de données générées.

### Exercice : évaluer le cycle de vie de la recherche

En 2018, Cox et Tam ont publié un article qui remet en question l'utilisation des modèles de cycle de vie pour représenter le processus de recherche. Ils ont opposé l'utilité de ces modèles à leur propension à simplifier à l'excès les activités concernées. Les auteurs ont interpellé les chercheuses et chercheurs de divers domaines à participer davantage à l'élaboration de tels modèles. Lisez leur publication et réfléchissez à la manière dont les prestataires de GDR (p. ex., les bibliothécaires qui offrent une formation en GDR) peuvent mieux représenter les complexités de la recherche et les intégrer à la gestion des données.

Cox, A. M. et Tam, W. W. T. (2018). A critical analysis of lifecycle models of the research process and research data management. *Aslib Journal of Information Management, 70*(2), 142-157. https://doi.org/10.1108/AJIM-11-2017-0251 (https://doi.org/10.1108/AJIM-11-2017-0251)

Le deuxième exemple démontre la nécessité de garder un œil sur l'avant-garde en matière de GDR en ce qui a trait aux structures administratives et de gouvernance. Un refrain fréquemment entendu au chapitre du partage des données, surtout lorsqu'il est question des dépôts, dit que les données doivent être les plus ouvertes possibles, mais aussi limitées que nécessaires. Comment cela se traduit-il, concrètement? Habituellement, les approches actuelles comprennent un **accès ouvert**, l'imposition d'une période d'embargo (c'est-à-dire, l'accès est refusé pendant un certain temps), ou exigent peut-être une permission de la part de la chercheuse ou du chercheur d'origine. Quelles autres options sont possibles?

Dans certains domaines de recherche, des infrastructures ont été mises en place pour examiner les utilisations suggérées des données avant de publier ces dernières. De tels systèmes de gouvernance aident à veiller à la conformité aux restrictions éthiques d'origines, empêchent d'endommager la propriété intellectuelle de la chercheuse ou du chercheur original (ou de l'équipe de recherche originale) et préviennent les risques, comme la réidentification, pour les personnes qui ont participé à l'étude (Murtagh *et al.*, 2018). La révision accrue des données peut également aider les gens hors de la sphère de la recherche à accéder aux données (p. ex., les journalistes, les groupes politiques, les citoyennes et citoyens scientifiques) tout en faisant en sorte que le travail de la chercheuse ou du chercheur ne soit pas discrédité, intentionnellement ou non (Murtagh *et al.*, 2018).

Les **comités d'accès aux données** (CAD) représentent une forme de gouvernance. Présents surtout en Europe et aux États-Unis, il s'agit d'organismes décisionnaires indépendants ayant pour but de superviser l'accès aux jeux de données à des fins de recherche, un peu comme un comité d'éthique, mais à la conclusion de la recherche. Ils réglementent l'accès aux données déjà recueillies. Les CAD sont plus présents dans des domaines comme la recherche biomédicale humaine où la combinaison de données permet une analyse plus poussée d'un plus grand nombre d'échantillons. Par exemple, une équipe peut souhaiter regrouper les données provenant de plusieurs **biobanques** internationales afin d'étudier le lien entre un variant génomique et un état de santé en particulier. Parce que de telles données sont grandement confidentielles, il y a très peu de chances qu'elles soient un jour accessibles librement. Certains CAD ont recours à des outils de prise de décision par ordinateur pour prendre des décisions en fonction du niveau de risque; d'autres s'en remettent aux examens de spécialistes dans le domaine. Des comités composés de personnes sont généralement préférés lorsqu'il s'agit de recherches d'avant-garde, où les données sont utilisées de manières novatrices, ou lorsque le domaine d'étude est particulièrement sensible.

**Étude de cas : METADAC**

De 2015 à 2020, j'ai fait partie d'une équipe qui a réalisé une ethnographie de METADAC (*Managing Ethico-social, Technical and Administrative issues in Data ACcess*), un comité sur l'accès aux données au Royaume-Uni. Ce comité a supervisé l'accès à des données génomiques et biosociales détenues par plusieurs études de cohortes longitudinales. Les membres du comité ont examiné les demandes de chercheuses et chercheurs de partout dans le monde qui suggéraient une recherche complexe sur le plan sociotechnologique et avant-gardiste sur le plan technologique (p. ex., relier des profils génétiques à des modes de scrutin). METADAC a cessé ses activités en décembre 2020 à la suite de changements à sa structure de financement; toutefois, son site Web est encore accessible pour y obtenir des détails au sujet de sa structure et des projets qu'il a approuvés (https://www.metadac.ac.uk (https://www.metadac.ac.uk)).

Murtagh, M. J., Blell, M. T., Butters, O. W., Cowley, L., Dove, E. S., Goodman, A., Griggs, R. L., Hall, A., Hallowell, N., Kumari, M., Mangino, M., Maughan, B., Mills, M. C., Minion, J. T., Murphy, T., Prior, G., Suderman, M., Ring, S. M., Rogers, N. T., ... Burton, P. R. (2018). Better governance, better access: practising responsible data sharing in the METADAC governance infrastructure. *Human Genomics, 12*(1), 1-12. https://doi.org/10.1186/s40246-018-0154-6 (https://doi.org/10.1186/s40246-018-0154-6)

Il est important d'être au courant de travaux importants, comme celui de Cox et Tam, ou de nouvelles infrastructures d'accès aux données, comme METADAC, car une telle connaissance aide à orienter la manière dont les données sont gérées et comment les services de soutien à la GDR sont organisés. L'objectif de la GDR évoluera à mesure que la gestion des données se développera pour englober davantage de disciplines et de types de données et en même temps que la nature de la recherche et des données évoluera. Votre travail dans ce domaine doit être dirigé par les pratiques exemplaires actuelles et une nécessité de s'adapter au changement et de rester au fait des progrès réalisés ailleurs.

## Les réalités en matière de gestion des données de recherche

Même lorsque des processus bien systématisés sont en place, la gestion des données ne deviendra jamais un

exercice qui consiste à cocher des cases d'une liste. Il faudra toujours prendre des décisions et il y aura toujours des données qui ne correspondent pas tout à fait à la pratique existante. Dans cette dernière section, nous étudions la réalité de la GDR sur la ligne de front de la recherche. Comment les chercheuses et chercheurs (et les personnes qui les appuient) recueillent les données de recherche, y accèdent, les traitent, les organisent, les analysent et les archives de façon à satisfaire aux exigences des organismes subventionnaires et des établissements hôtes tout en s'inscrivant de manière pragmatique dans le travail de l'équipe de recherche?

Selon la complexité d'un projet, la gestion méthodique des données de recherche peut s'avérer accablante, désorganisée ou simplement négligée à une étape ou l'autre du processus. Les choses peuvent mal tourner qu'il s'agisse de planifier la GDR au début de l'étude, lors de la résolution de problèmes au milieu d'une étude, lorsque vous tentez de trouver un sens à des données existantes ou pendant que vous aidez d'autres personnes dans l'une ou l'autre de ces situations. En dépit des embûches, les données de recherche peuvent être gérées efficacement, même dans les situations les plus complexes, si vous faites preuve d'un esprit critique, que vous êtes conséquent, que vous documentez ce que vous faites et que vous recherchez les pratiques exemplaires ainsi que du soutien, le cas échéant.

Vous trouverez ci-après les grandes lignes de ce que vous devriez prendre en considération lorsque vous gérez activement des données de recherche sur le terrain. Elles ressemblent vaguement à un PGD, bien qu'elles aient été conçues du point de vue d'effectuer la GDR au lieu de la planifier. Si les PGD constituent des documents évolutifs, rien n'est plus immédiat pour les chercheuses et chercheurs que de devoir gérer des données en même temps qu'un certain nombre d'autres tâches urgentes. Le fait est que la GDR est casée dans un monde où le temps est déjà une denrée rare. Les idées formulées ici sont issues de ce que j'ai appris moi-même et de ce que des collègues ont appris au cours de leur carrière de chercheuse et chercheur et de gestionnaire de données pendant plus d'une décennie. Les sept éléments qui suivent mettent l'accent sur les chercheuses et chercheurs; toutefois, ils devraient être aussi utiles aux bibliothécaires de données et à d'autres spécialistes pour leur donner un meilleur aperçu de la GDR.

1. **La GDR repose autant sur la réflexion et la résolution de problèmes que sur l'action.** La gestion des données est une activité à grande échelle. Elle ne porte pas uniquement sur les données et leur gestion. Il s'agit d'un élément (relativement nouveau) d'un processus de recherche plus large. Inversement, au moment d'entreprendre une recherche précise, il peut y avoir très peu d'informations sur la manière de gérer les données. Pour utiliser une analogie fondée sur la recherche, les logiciels d'analyse comme SPSS et NVivo simplifient le travail avec des données. En revanche, de tels programmes ne les analysent pas. C'est le travail de la chercheuse ou du chercheur. Si les approches en GDR peuvent guider la gestion des données, les chercheuses, les chercheurs et les personnes qui les soutiennent doivent réfléchir de manière critique aux données en question et à la manière de mettre en application les principes et pratiques de façon pragmatique (et souvent novatrice) qui améliorent les processus et les résultats de recherche. En d'autres termes, en situation de GDR, en plus d'agir, il faut s'accorder

suffisamment d'espace pour réfléchir de manière critique.

2. **Rédigez un plan de gestion des données.** Les PGD sont utiles, car ils forcent les chercheuses et chercheurs à réfléchir aux aspects essentiels de la manière dont les données seront gérées au cours de l'étude. Même si une demande de subvention n'exige pas la présence d'un PGD, pensez à en créer un. Lorsque vous avez terminé, demandez-vous ce qu'il comprend et ce qu'il ne comprend pas. Souvenez-vous que les PGD sont à la fois ambitieux et orientés vers un objectif : ils vous indiquent votre destination et la manière dont vous souhaitez y arriver. Ils ne prennent pas en considération les réalités de la gestion quotidienne des données de recherche, comme composer avec une personne dans l'équipe de recherche qui ne nomme pas les fichiers correctement ou éprouver de la difficulté à trouver un dépôt convenable. C'est alors que le point 1 ci-dessus entre en jeu.
3. **Réfléchissez à ce qui stimule vos efforts en GDR.** De nos jours, la plupart d'entre nous qui participons à la recherche améliorons nos compétences en GDR, car nous croyons que nous devons le faire, surtout depuis que les organismes subventionnaires exigent de plus en plus une bonne gestion des données. Quels autres facteurs entrent en jeu dans votre projet? En tant que chercheuse ou chercheur en milieu de carrière, vous pourriez constater qu'une meilleure organisation de vos données peut avoir une incidence positive sur vos résultats de recherche ou sur votre capacité à collaborer avec d'autres personnes. En tant que membre de la communauté étudiante des cycles supérieurs, vous pourriez remarquer que la personne responsable de votre étude est nouvelle en GDR, elle aussi, ce qui vous permet d'étoffer vos compétences afin de jouer un rôle de premier plan dans ce domaine. Si vous entreprenez une analyse secondaire, vous pourriez devoir retourner vos données à la personne responsable de l'étude originale et devrez connaître quel niveau de gestion des données est attendu. Peu importe la situation, il est avantageux pour vous de définir pourquoi la GDR compte pour vous.
4. **À quoi ressemblerait le résultat idéal?** Cette étape est importante pour toute personne qui travaille dans une discipline où il y a peu de lignes directrices ou de pratiques exemplaires bien définies en GDR. Prenez le temps de réfléchir à l'approche idéale relativement à la gestion et à l'archivage de vos données. Si vous êtes une chercheuse ou un chercheur en quête de données pour réaliser une analyse secondaire, à quoi ressemblerait le jeu de données idéal au chapitre de l'organisation, de la documentation, des métadonnées, des ententes d'accès, et ainsi de suite? Si une solution parfaite n'existe pas, il existe sûrement des exemples qui s'en approchent quelque part dans le monde. Réfléchissez de manière critique à l'endroit où vous pourriez les trouver et commencez à les chercher. Posez des questions jusqu'à ce que vous obteniez des réponses avec lesquelles vous pouvez travailler.
5. **Préparez-vous à aborder vos données et leur gestion de manière itérative.** Les données de recherche ne sont presque jamais recueillies dans leur état définitif. Elles ont besoin d'être nettoyées, reformatées, anonymisées, agrégées, entre autres, avant d'atteindre un état convenable pour l'analyse et l'archivage. En tant que chercheuse ou chercheur, vous devez décider si toutes vos données ont une importance égale (pour le projet, pour la communauté de recherche). Il est essentiel que vous documentiez vos données et leur **provenance**, car ces détails offrent à d'autres personnes (les membres

de l'équipe, les utilisatrices et utilisateurs secondaires) des renseignements importants, notamment sur l'analyse que les données peuvent ou ne peuvent pas supporter. De tels efforts évoluent au fil du temps : ce que vous croyez et faites au début de l'étude peut changer en cours de projet. La GDR est rarement une entreprise unique. Le simple protocole pour nommer les fichiers, à lui seul, peut ne plus fonctionner adéquatement à l'étape de l'analyse.

6. **Qui fait quoi?** Une gestion des données efficaces, surtout lorsqu'il est question d'équipes de recherche, exige que les rôles et responsabilités soient définis et qu'ils soient continuellement examinés pour vérifier que ce qui est prévu correspond à ce qui se produit. Pour certains membres de l'équipe, voire l'ensemble des membres de l'équipe, il peut s'avérer nécessaire d'améliorer des compétences; par conséquent, évaluez la situation et déterminez les ressources externes en début de processus. Les réunions peuvent gruger un temps précieux, mais le fait de vous rencontrer régulièrement pour échanger de l'information au sujet de la GDR sur un projet vous permettra d'aborder les défis qui surviendront, à coup sûr, comme une postdoctorante ou un postdoctorant qui quitte le projet pour occuper un poste menant à la permanence. Comme toujours, faites en sorte que vous et votre équipe documentiez systématiquement vos efforts de GDR avec des ressources comme les **pistes de vérification** et les procédures opérationnelles standards.
7. **Convenez que tout ne se déroulera peut-être pas sans heurts; néanmoins, vous pourriez obtenir un résultat raisonnable.** La GDR ressemble aux processus de recherche qu'elle soutient : en constante évolution et jamais parfaite. Faites de votre mieux et mettez vos apprentissages en application au fur et à mesure que vous allez les acquérir.

## Conclusion

Ce manuel constitue un excellent point de départ quant aux enjeux principaux en gestion des données de recherche au Canada. Les divers chapitres présentent un large éventail de principes, de politiques, de stratégies et de pratiques très utiles que vous devriez connaître en tant que chercheuse ou chercheur, bibliothécaire universitaire ou spécialiste des données. Le principal point à retenir de ce chapitre est simple : la gestion des données requiert toujours de la réflexion et une ouverture pour les nouvelles idées et pratiques. Essentiellement, la GDR demeure la responsabilité des chercheuses et chercheurs qui travaillent en première ligne, dont la plupart sont encore novices, non pas tant en ce qui concerne la gestion des données de recherche que la gestion de ces données conformément aux nouvelles exigences externes. Malheureusement, de telles exigences ne se transposent pas facilement en pratique de recherche, ce qui crée plusieurs défis continus. Les bibliothécaires et d'autres spécialistes des données offrent un soutien précieux à ce travail, bien que leurs efforts doivent être évalués d'un point de vue critique alors que divers modèles de service émergent. La GDR n'est pas un élément unique ni une entreprise statique. L'apprentissage à partir de ce manuel est essentiel, mais

il faut aussi garder une perspective critique et faire preuve de curiosité quant à la manière dont les choses peuvent être faites ailleurs.

**Éléments clés à retenir**

- En plus de soutenir le partage et la réutilisation, la gestion efficace des données fait partie intégrante du processus de recherche; l'essentiel du travail de GDR se produit pendant le projet, plutôt qu'à la fin de celui-ci. La gestion cohérente des données est également importante entre des études interreliées, au fil du temps.
- La responsabilité de la GDR risque d'incomber à plus d'une personne. Les membres de l'équipe de recherche assument divers domaines de responsabilité et peuvent avoir des points de vue divergents ainsi que des niveaux de compétences variées. Les tâches quotidiennes en matière de GDR sont souvent déléguées à des chercheuses ou chercheurs en début de carrière qui n'auront pas de liens avec les données à long terme.
- Les approches actuelles et les pratiques exemplaires en GDR sont dynamiques. Préparez-vous à devoir vous adapter, à changer, à rester à l'affût de tendances émergentes et des autres façons de faire au chapitre de la résolution de problèmes.
- Ne vous attendez pas à tout réussir, car la « bonne » façon de faire n'existe peut-être pas encore!

## Lectures et ressources supplémentaires

Cheah, P. Y. et Piasecki, J. (2020). Data access committees. *BMC Medical Ethics, 21*(12), 1-8. https://doi.org/10.1186/s12910-020-0453-z (https://doi.org/10.1186/s12910-020-0453-z)

Kruse, F. et Thestrup, J. B. (2017). *Research data management – A European perspective*. De Gruyter Saur. https://www.degruyter.com/document/doi/10.1515/9783110365634 (https://www.degruyter.com/document/doi/10.1515/9783110365634/html?lang=en)

Pasek, J. E. (2017). Historical development and key issues of data management plan requirements for National Science Foundation grants: A review. *Issues in Science and Technology Librarianship, 87*. https://doi.org/10.5062/F4QC01RP (https://doi.org/10.5062/F4QC01RP)

Rice, R. et Southall, J. (2016). *The data librarian's handbook*. Facet Publishing.

Thompson, K. et Kellam, L. M. (dir.). (2016). Introduction to databrarianship: The academic data librarian in theory and practice. Dans L. M. Kellam et K. Thompson (dir.), *Databrarianship: The academic data librarian in theory and practice*. Association of College and Research Libraries. https://scholar.uwindsor.ca/cgi/viewcontent.cgi?article=1047&context=leddylibrarypub (https://scholar.uwindsor.ca/cgi/viewcontent.cgi?article=1047&context=leddylibrarypub)

Whyte, A. et Tedds, J. (2011). *Making the case for research data management*. Digital Curation Centre.

# Bibliographie

Cox, A. M. et Tam, W. W. T. (2018). A critical analysis of lifecycle models of the research process and research data management. *Aslib Journal of Information Management, 70*(2), 142-157. https://doi.org/10.1108/AJIM-11-2017-0251 (https://doi.org/10.1108/AJIM-11-2017-0251)

Murtagh, M. J., Blell, M. T., Butters, O. W., Cowley, L., Dove, E. S., Goodman, A., Griggs, R. L., Hall, A., Hallowell, N., Kumari, M., Mangino, M., Maughan, B., Mills, M. C., Minion, J. T., Murphy, T., Prior, G., Suderman, M., Ring, S. M., Rogers, N. T., ... Burton, P. R. (2018). Better governance, better access: practising responsible data sharing in the METADAC governance infrastructure. *Human Genomics, 12*(1), 1-12. https://doi.org/10.1186/s40246-018-0154-6 (https://doi.org/10.1186/s40246-018-0154-6)

Pinfield, S., Cox, A. M. et Smith, J. (2014). Research data management and libraries: Relationships, activities, drivers and influences. *PLoS ONE, 9*(12), e114734. https://doi.org/10.1371/journal.pone.0114734 (https://doi.org/10.1371/journal.pone.0114734)

Plomp, E., Dintzner, N. J. R.,. Teperek, M. et Dunning, A. (2019). Cultural obstacles to research data management and sharing at TU Delft. *Insights, 32*(1). https://doi.org/10.1629/uksg.484 (https://doi.org/10.1629/uksg.484)

## À propos de l'auteur

### Dr. Joel T. Minion

Joel T. Minion, PhD MLIS MA BA (spécialisé) est un chercheur qualitatif en santé, un bibliothécaire, un gestionnaire de données et un éducateur ayant de l'expérience en gestion de données de recherche (GDR) au

Canada et en Europe. Il est actuellement chercheur scientifique au sein du programme Translating Research in Elder Care (TREC) de la Faculty of Nursing de l'Université de l'Alberta où il est responsable de la planification de l'héritage des données longitudinales du TREC. Joel était auparavant responsable de la recherche qualitative pour la Health Technology Assessment Unit de l'Université de Calgary au sein du O'Brien Institute for Public Health et avant cela, il était associé principal de recherche au centre de recherche PEALS (Policy, Ethics and Life Sciences) de l'Université Newcastle au Royaume-Uni. Il est titulaire d'un doctorat en informatique de la santé de l'Université de Sheffield et d'un diplôme MLIS de l'Université Western. Depuis 2010, Joel est activement impliqué dans la gestion des données de recherche qualitative et dans les efforts en cours pour les intégrer dans des cadres de GDR plus larges.

# GLOSSAIRE

## Analyse exploratoire des données

Processus utilisé pour explorer, analyser et synthétiser des jeux de données au moyen de méthodes quantitatives et graphiques. L'analyse exploratoire des données aide à faire ressortir des patrons et facilite la découverte d'irrégularités et d'incohérences dans un jeu de données.

## Analyse secondaire

Dans le cadre d'une recherche, utilisation de données déjà existantes. L'analyse est généralement menée par des chercheuses et chercheurs qui n'ont aucun lien avec la recherche originale.

## Approbation éthique

Autorisation qui donne le feu vert à la tenue d'une étude. Elle est obtenue par le biais d'un comité dont les appellations varient : comité d'éthique en recherche, comité d'éthique indépendant ou comité de révision déontologique.

## Argument [logiciel tableur]

Valeur ou variable utilisée par une fonction d'un logiciel tableur pour effectuer un calcul. Par exemple, dans Excel, les fonctions sont des formules intégrées au logiciel.

## ASCII

American Standard Code for Information Interchange (Code américain normalisé pour l'échange d'information). Norme informatique de codage de caractères. Elle définit 128 codes qui représentent les chiffres arabes de 0 à 9, les 26 lettres de l'alphabet latin en minuscule et en capitales ainsi que des symboles mathématiques et de ponctuation.

## Assistant PGD

Outil en ligne qui pose, aux personnes qui l'utilisent, une série de questions sur leurs données et leurs plans de recherche. De l'aide et des conseils contextuels sont disponibles pour aider à répondre aux questions.

## Attaque par homogénéité

Moyen de porter atteinte à la confidentialité d'un groupe de participantes et de participants à une recherche quand toutes les personnes ayant le même ensemble d'attributs particuliers possèdent aussi un même attribut sensible.

## Authentification multifactorielle

Type d'authentification qui implique un mot de passe et un appareil. L'utilisation d'un mot de passe pour ouvrir une session dans un service entraîne une demande d'entrer un code à usage unique généré par un appareil tel qu'un cellulaire ou un ordinateur. Les codes à usage unique peuvent être transmis par message texte, par courriel ou ils peuvent être générés sur un appareil par le biais d'une application d'authentification telle que Google Authenticator. Plusieurs institutions bancaires et gouvernementales, telles que l'Agence du revenu du Canada, exigent maintenant l'activation de l'authentification multifactorielle.

## Autodétermination

Droit des peuples autochtones à déterminer ce qu'il y a de mieux pour leur développement social, culturel et économique afin d'assurer le bien-être de leurs membres. Cette définition s'inspire de la Déclaration des Nations Unies sur les droits des peuples autochtones (DNUDPA).

## Biobanque

Dépôt qui stocke des échantillons biologiques, physiques et des données biologiques.

## Boîte à moustache

Représentation graphique d'un jeu de données qui affiche la distribution des données et de toute valeur aberrante potentielle. Aussi appelé diagramme en boîte.

## Cahier de laboratoire électronique

Type d'outil en ligne basé sur la conception et l'utilisation des cahiers de laboratoire papier.

## Carte de base

Carte de référence sous-jacente qui sert d'assise aux données pour les mettre en contexte. Par exemple, sur une carte montrant des informations démographiques dans plusieurs zones de recensement (https://www.arcgis.com/apps/View/index.html?appid=00ec54e38f7b43f081c60956234bc8cb&extent=-80.0230,42.7964,-78.7047,43.3142), la carte sera plus difficile à lire sans l'ajout des limites de chacune des zones de

recensement. Une carte peut être considérée comme une représentation abstraite, mais sa lecture est enseignée et elle aide les individus à se situer. La carte de base permet donc de situer des données au-dessus d'une carte grâce aux informations sur la position.

## Chaînage [logiciel de sondage]

Texte inséré de façon automatique par un logiciel de sondage selon les réponses précédemment données par les personnes qui remplissent un questionnaire.

## Classe d'équivalence

Ensemble d'enregistrements qui comporte les mêmes valeurs d'identifiants indirects à l'intérieur d'un jeu de données.

## Clé d'anonymisation

Document utilisé par les chercheuses et chercheurs en recherche qualitative pour dépersonnaliser leurs données de manière systématique. Le document relie les informations qui sont supprimées des données originales (par exemple, le nom d'une personne dans la transcription d'une entrevue) et qui sont remplacées par un texte plus générique (p. ex., Personne 6). La chercheuse ou le chercheur travaille alors avec la transcription anonymisée, mais peut utiliser la clé pour réidentifier des personnes, des lieux, des organisations, etc., si ces informations redeviennent importantes au cours de l'analyse. Une clé d'anonymisation doit être protégée par un mot de passe, stockée en toute sécurité et ne jamais être conservée avec les données en question. Elle est souvent détruite à la fin de l'étude.

## Comité d'accès aux données

Organe décisionnel indépendant dont l'objectif est de superviser l'accès aux jeux de données à des fins de recherche.

## Communauté d'utilisateurs cible

Entité conceptuelle introduite par la norme OAIS qui se rapporte aux personnes utilisatrices potentielles d'un objet numérique préservé dans une archive. La Communauté d'utilisateurs cible est un concept essentiel pour la planification de la préservation à long terme parce qu'il nécessite une compréhension des besoins et des capacités de la Communauté d'utilisateurs cible, permettant ainsi de faire des choix éclairés en matière, notamment, de formats de fichiers et de rétention des données.

## Compression sans perte

Mécanisme de réduction de la taille des fichiers qui permet de conserver toutes les données originales.

(Manuel de préservation numérique (https://www.dpconline.org/docs/digital-preservation/handbook/translations-3/2519-handbook-2021-fr/file), s.d.).

## Conception descriptive

Type de conception d'étude qui se préoccupe des questions exploratoires (p. ex., quoi? quand? comment? où?). L'étude vise à explorer un phénomène ou à effectuer une observation pour décrire un effet.

## Conception explicative

Type de conception d'étude qui se préoccupe des liens de causalité (p. ex., les causes et leurs effets ou des questions liées au "pourquoi" d'un effet). L'étude vise à expliquer un phénomène ou une observation pour comprendre un effet.

## Conteneur informatique

Ordinateur autonome virtuel à l'intérieur d'un ordinateur. Il comprend tout ce qui est nécessaire pour faire fonctionner un logiciel (y compris le système d'exploitation), sans avoir à télécharger et installer des programmes ou des données.

## CONTENTdm

Outil d'OCLC pour la gestion et la présentation de contenu numérique. Consultez https://www.oclc.org/fr/contentdm.html (https://www.oclc.org/fr/contentdm.html) pour plus d'informations.

## Contrôle d'intégrité

Méthode permettant de garantir l'intégrité d'un fichier et de vérifier qu'il n'a pas été altéré ou corrompu. Pendant les transferts de fichiers, une archive peut effectuer un contrôle d'intégrité pour s'assurer qu'un fichier transmis n'a pas été altéré en cours de route. Au sein de l'archive, le contrôle d'intégrité est utilisé pour s'assurer que les fichiers numériques n'ont pas été altérés ou corrompus. Il est le plus souvent réalisé en calculant des sommes de contrôles telles que MD5, SHA1 ou SHA256 pour un fichier et en les comparant à une valeur stockée. (Manuel de préservation numérique (https://www.dpconline.org/docs/digital-preservation/handbook/translations-3/2519-handbook-2021-fr/file), s.d.).

## Couche [système d'information géographique]

Représentation visuelle d'un jeu de données géographiques dans un environnement cartographique numérique. De manière conceptuelle, une couche est une tranche ou une strate de la réalité

géographique dans une zone donnée. Elle équivaut plus ou moins à un élément de légende sur une carte papier. Sur une carte routière, par exemple, les routes, les parcs nationaux, les limites politiques et les fleuves sont autant de couches différentes. (ESRI (https://support.esri.com/fr-fr/gis-dictionary/layer), s.d.)

## Création de paquets de données

Traitement qui consiste à regrouper des données et des informations sur les données dans un ensemble logique qui sera utilisé dans un processus de préservation numérique.

## Cycle de vie des données

Cycle au cours duquel les données sont recueillies, traitées, analysées, préservées et ensuite partagées avec d'autres chercheuses et chercheurs qui pourront recommencer le cycle.

## Data Documentation Initiative (DDI)

Schéma de métadonnées basé sur des normes et développé pour les données en sciences sociales.

## Dégradation du média

Menace à la longévité des objets numériques basée sur la détérioration du support sur lequel ils sont stockés. Parfois appelé « pourriture de l'octet » (bit rot). Les menaces de dégradation du média sont souvent traitées par le biais d'actions de préservation qui assurent l'intégrité des bits, y compris la vérification active des objets numériques pour y déceler des altérations/pertes, en plus de mesures qui visent à créer de multiples copies d'un objet sur différents types de médias.

## Délimiteur

Caractère qui sépare les données.

## Dépendance [informatique]

Bibliothèque de logiciels supplémentaires qui peut être téléchargée à partir d'Internet et utilisée pour certaines tâches précises de programmation.

## Dépersonnalisation

Procédé par lequel tout renseignement qui pourrait compromettre la vie privée des participantes et participants à une recherche dans un jeu de données est retiré.

### Détail identifiant

Toute information dans un jeu de données qui, combinée, pourrait conduire à la divulgation de l'identité d'une personne.

### Dictionnaire de données

Fichier qui documente et décrit les différents éléments d'un jeu de données. Par exemple, il peut définir les variables, les unités de mesure utilisées, les valeurs acceptées pour les variables, etc. Le document est lisible et souvent exploitable par une machine, comme le guide de codification, et peut également contenir des informations détaillées sur la structure technique d'un jeu de données.

### Données administratives

Données recueillies dans le cadre d'un travail de gestion administrative. Les données administratives peuvent être utilisées pour faire le suivi de personnes, d'achats, d'inscriptions, de prix, etc.

### Données de recherche

Sources d'informations ou de preuves qui ont été compilées pour servir de base à la recherche.

### Données matricielles

Données qui représentent des espaces sous la forme d'une grille ou d'une série de cellules, chacune avec une valeur particulière – souvent considérées comme les pixels d'une image. Par exemple, un document numérisé comme une carte historique ou une photo aérienne.

### Données ouvertes

Données en ligne, gratuites et accessibles qui peuvent être utilisées, réutilisées et distribuées.

### Données probantes

Données qui se présentent sous diverses formes et qui sont issues d'une activité de recherche : analyse de données, modélisation, synthèse de la littérature, évaluation permettant de produire des lignes directrices, évaluation de mise en œuvre d'un procédé ou d'une technologie et de son coût-efficacité.

### Données qualitatives

Données générées par des recherches qui examinent les aspects sociaux de la condition humaine en utilisant des méthodes descriptives plutôt que des mesures.

### Données sensibles

Données qui ne peuvent être partagées sans risque de trahir la confiance ou de nuire à une personne, une entité ou une communauté.

### Données tabulaires

Données disposées sous la forme de tables ou de tableaux, c'est-à-dire en lignes et en colonnes.

### Données vectorielles

Données qui comprennent des points individuels qui se rapportent à des endroits particuliers. Ces points peuvent être reliés pour former des lignes ou des formes (polygones). Ces points, lignes et polygones peuvent être traités comme des unités individuelles avec des données associées.

### Droit à l'oubli

Droit qui permet à la personne concernée d'obtenir du responsable du traitement l'effacement, dans les meilleurs délais, de données à caractère personnel la concernant et le responsable du traitement a l'obligation d'effacer ces données à caractère personnel dans les meilleurs délais [traduction]. (GDPR.EU (https://gdpr.eu/), 2018).

### Dublin Core

Schéma de métadonnées simple et générique qui utilise 15 propriétés de base facultatives et répétables comme le titre, le créateur, le format et la date. Créé en 1995, Dublin Core est également une norme internationale (ISO 15836).

### Échelle d'intervalles

Échelle qui utilise des chiffres dont la distance entre eux est équivalente, soit en ordre croissant ou décroissant, et où zéro pourrait représenter un point sur l'échelle (c'est-à-dire que zéro n'implique pas une absence de valeur). La température et l'heure en sont de bons exemples. Dans le cas de l'échelle de température en degré Celsius, le zéro se rapporte au point où l'eau gèle, non pas à une absence de température.

### Échelle de Likert

Outil élaboré en additionnant ou en faisant la moyenne d'un certain nombre d'items de Likert liés entre eux. Un item de Likert est une question ou un énoncé dans un sondage où la personne interrogée doit exprimer son degré d'accord ou de désaccord.

## Échelle de rapport

Échelle qui peut augmenter ou baisser en fonction d'un dénominateur plutôt que de distances équivalentes. Sur une échelle de mesure de rapports, le zéro n'est pas un point sur l'échelle, mais plutôt une absence de valeur. La densité de population est un exemple de mesure de rapports. Dans le cas de densité de population, zéro se rapporte à un endroit sans résidants.

## Émulation

Moyen de surmonter l'obsolescence technologique du matériel et des logiciels en développant des techniques permettant d'imiter des systèmes obsolètes sur les futures générations d'ordinateurs. (Manuel de préservation numérique (https://www.dpconline.org/docs/digital-preservation/handbook/translations-3/2519-handbook-2021-fr/file), s.d.).

## Énoncé de politique des trois conseils sur l'éthique de la recherche avec des êtres humains (EPTC 2)

Cadre principal harmonisé qui guide l'établissement des lois canadiennes et des paradigmes éthiques plus larges en lien avec le droit des êtres humains en recherche.

## Environnement de développement intégré

Application logicielle qui fournit un environnement complet de développement de logiciels. RStudio est un environnement de développement intégré qui permet aux personnes qui l'utilisent d'écrire, de déboguer, d'exécuter du code R et d'afficher les sorties correspondantes.

## Étude longitudinale

Type d'étude qui s'intéresse aux effets du temps sur un résultat quelconque. Autrement dit, une étude qui mesure un résultat à plusieurs moments dans le temps. Par exemple, une enquête longitudinale implique une même enquête sur les mêmes individus répétée à plusieurs moments pour comprendre les changements d'attitude ou de comportement au fil du temps.

## Évaluation de la maturité de la gestion des données de recherche

Évaluation de l'état actuel des services et du soutien en gestion des données de recherche, généralement pour un établissement particulier.

### Extension de fichier

Suffixe attribué à un fichier afin de l'identifier. Par exemple, un fichier créé avec le logiciel Word portera l'extension DOCX.

### Fichier CSV

Fichier texte délimité qui utilise la virgule pour séparer les valeurs d'un enregistrement de données. Chaque ligne du fichier correspond à un enregistrement de données.

### Fichier TSV

Fichier texte délimité qui utilise une tabulation pour séparer les valeurs. Chaque ligne du fichier correspond à un enregistrement de données.

### Format de fichier

Méthode normalisée qui répartit des uns et des zéros pour qu'ils puissent être utilisés pour codifier certains types particuliers d'informations.

### Format non propriétaire

Format qui n'appartient pas à une entreprise.

### Format ouvert

Format dont les spécifications techniques sont publiques. Les renseignements qui permettent de comprendre le fonctionnement et la structure du format sont accessibles.

### Format tabulaire

Informations intégrées à des tableaux avec des rangées et des colonnes.

### Fourche [Github]

Dans GitHub, copie d'un jeu de données qui conserve son lien vers la création originale.

### Frais de traitement d'article

Frais de publication facturés aux autrices, auteurs ou à leurs établissements pour rendre une œuvre disponible en libre accès.

### Gestion des données de recherche

Terme qui décrit toutes les activités que les chercheuses et chercheurs effectuent pour structurer, organiser et préserver les données de recherche avant, pendant et après le processus de recherche.

### Gestion des versions

Système qui fait automatiquement le suivi de chaque modification à un document ou fichier, permettant aux personnes qui l'utilisent de revenir à des versions sauvegardées antérieures sans avoir à continuellement enregistrer des copies sous différents noms.

### Gestionnaire de mots de passe

Logiciel qui stocke les mots de passe. Certains gestionnaires de mots de passe peuvent aussi créer et suggérer des mots de passe plus complexes à utiliser.

### Guide de codification

Fichier surtout utilisé par des sondeurs qui fournit des informations détaillées sur l'outil de sondage. Par exemple, on y retrouve les questions du sondage, les noms et définitions des variables utilisés pour coder les réponses du sondage, les valeurs acceptées pour chacune des variables, des statistiques sommaires pour chacune des questions, etc.

### Histogramme

Représentation graphique de la distribution d'un jeu de données continues ou de valeurs énumérables et identifiables séparément.

### Humanités numériques

Domaine de recherche qui s'intéresse à l'application d'outils et de méthodes informatiques aux disciplines traditionnelles des sciences humaines telles que la littérature, l'histoire et la philosophie.

### Identifiant direct

Renseignement recueilli par la chercheuse ou le chercheur qui permet d'identifier des participantes ou des participants à une recherche. Les noms, numéros de téléphone, numéros d'assurance sociale et numéros d'étudiant sont des exemples d'identifiants directs.

### Identifiant indirect ou quasi-identifiant

Attribut d'un individu qui n'est pas identifiant en soi mais qui, en combinaison avec d'autres

renseignements, peut permettre d'identifier une personne. Un attribut ne peut être quasi-identifiant que si des pirates informatiques peuvent raisonnablement jumeler cet attribut à des informations de source externe.

## Identifiant numérique d'objet (DOI)

Nom pour une entité dans un réseau numérique; il ne s'agit pas d'une localisation. Le nom fournit un système pour l'identification pérenne et exploitable ainsi que pour l'échange interopérable d'informations gérées sur des réseaux numériques. Un DOI est un type d'identifiant pérenne émis par la Fondation internationale DOI. Cet identifiant permanent est associé à un objet numérique, ce qui permet à l'objet d'être fidèlement cité en référence, et ce, même si sa localisation et ses métadonnées sont modifiées au fil du temps [traduction]. (CODATA Research Data Management Terminology (https://codata.org/rdm-terminology/digital-object-identifier/), s.d.).

## Identifiant unique pérenne

Référence durable à un objet numérique qui fournit des informations sur cet objet indépendamment de ce qui lui arrive. Développé pour lutter contre des liens qui deviennent obsolètes (*link rot*), un identifiant pérenne peut être résolu pour fournir une représentation appropriée d'un objet, que celui-ci change d'emplacement en ligne ou qu'il soit mis hors ligne [traduction]. (CODATA Research Data Management Terminology (https://codata.org/rdm-terminology/persistent-identifier/), s.d.).

## Infonuagique

Système informatique qui est réparti sur plus de deux serveurs dans plus de deux emplacements, permettant ainsi un accès à distance par l'entremise de navigateurs Web ou d'interfaces de programmation (API) pour la puissance de calcul et/ou le stockage des données.

## Intégration [informatique]

Processus consistant à relier des systèmes ou des outils différents, souvent disparates, en une infrastructure cohérente.

## Intégrité

Concept lié à la permanence des objets numériques. L'uniformité des objets numériques est complexe à établir; la façon dont ils sont stockés implique que les objets sont souvent copiés ou transférés et il faut s'assurer qu'ils restent identiques aux objets avant la copie ou le transfert. Dans la pratique courante, l'intégrité est intimement liée à la génération et la vérification des sommes de contrôle, ce qui peut aider à assurer qu'une série ordonnée de bits est restée inchangée.

## Interface de programmation d'application (API)

Pour une application donnée, ensemble de fonctions et de procédures fournies par une bibliothèque de logiciels ou un service Web avec lequel une autre application peut communiquer.

## Interopérabilité

Capacité des données ou des outils provenant de ressources non coopératives à travailler ou à communiquer entre eux avec un minimum d'effort et en utilisant un langage commun. L'interopérabilité exige que les données et les métadonnées utilisent des formats normalisés, accessibles et largement utilisés. Par exemple, lors de la sauvegarde de données tabulaires, il est recommandé d'utiliser un fichier CSV plutôt qu'un fichier propriétaire tel que XLSX (Excel). Un fichier CSV peut être ouvert et lu par davantage de logiciels qu'un fichier XLSX.

## Jumeau de données

Dans un jeu de données ayant des identifiants indirects, enregistrement qui a la même valeur ou les mêmes attributs qu'un autre enregistrement. Par exemple, dans un jeu de données, deux hommes blancs entre 25-30 ans sont des jumeaux de données.

## K-anonymisation

Approche permettant de démontrer mathématiquement qu'un jeu de données a été anonymisé. L'approche part du principe que ce ne devrait pas être possible d'isoler moins de « k » cas individuels dans un jeu de données et ce, pour toutes les combinaisons possibles de variables identificatoires – « k » correspond au numéro établi par la chercheuse ou le chercheur.

## K-anonymisation p-sensible

Évaluation des risques à la vie privée fondée sur la k-anonymisation, mais plus contraignante.

## L-diversité

Évaluation des risques à la vie privée fondée sur la k-anonymisation, mais plus contraignante. La l-diversité est appliquée à un jeu de données quand chaque groupe d'enregistrements qui partage une même série d'attributs démographiques comporte au moins « l » valeurs différentes pour chacune des variables confidentielles.

## Les organismes subventionnaires

Le Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG), le Conseil de

recherches en sciences humaines du Canada (CRSH) et les Instituts de recherche en santé du Canada (IRSC) (les organismes subventionnaires) représentent les trois agences fédérales de financement de la recherche au Canada. Ils sont à la source d'une importante proportion des fonds de recherche au Canada.

### Libre accès

Disponibilité libre et immédiate d'informations sans limites d'utilisation dans l'environnement numérique.

### LISEZ-MOI

Document qui fournit des renseignements à propos d'un fichier ou d'un jeu de données. Il permet d'assurer la pérennité de l'interprétation correcte des données par toutes les personnes qui les consulteront.

### Littératie en matière de codage

Capacité de bien comprendre le code informatique, au même titre que la littératie en mathématiques (ou numératie) est la capacité de bien comprendre les mathématiques. L'apprentissage du code en informatique a souvent été comparé à l'apprentissage d'une nouvelle langue.

### Métadonnées

Éléments d'information utilisés pour décrire le contenu ou le contenant d'une ressource. Elles peuvent être structurées ou non.

### Métadonnées lisibles par machine

Métadonnées qui sont dans un format qu'un ordinateur peut utiliser et comprendre.

### Modèle d'évaluation de la maturité

Outil utilisé pour évaluer le niveau de sophistication d'un service ou d'un produit. Les différents modèles mesurent le niveau d'aboutissement de capacités dans des domaines clés en utilisant une échelle de valeurs numériques, permettant ainsi aux personnes qui les utilisent de quantifier ces capacités et de favoriser l'amélioration continue des processus.

### Modèle pour l'évaluation de la maturité au Canada (MEMAC)

Version française du *Maturity Assessment Model in Canada* (MAMIC). Outil d'évaluation de la gestion

des données de recherche (GDR) proprement canadien conçu pour aider les établissements dans l'évaluation de l'état actuel de leurs services et soutien à la GDR, ce qui constitue un des éléments dans le processus d'élaboration de stratégies institutionnelles de GDR. Le MEMAC s'articule autour de quatre champs de services et de soutien : les politiques et processus de l'établissement, l'infrastructure informatique, les services de soutien et le support financier. Il permet donc aux personnes qui l'utilisent d'évaluer la maturité et l'ampleur de ces services.

## Nettoyage des données

Processus qui vise à identifier et corriger les données altérées, inexactes ou non pertinentes. Cette étape fondamentale du traitement des données améliore la cohérence, la fiabilité et la valeur des données. (Talend (https://www.talend.com/fr/resources/what-is-data-cleansing/#:~:text=D%C3%A9finition%20du%20nettoyage%20des%20donn%C3%A9es,fiabilit%C3%A9%20et%20valeur%20des%20donn%C3%A9es.), s.d.).

## Niveau de maturité [MEMAC]

Mesure du degré d'aboutissement d'un élément particulier en lien avec la gestion des données de recherche. Plus la note est faible, moins l'élément est développé (mature).

## Normalisation

Lors de l'ingestion des fichiers dans un système de préservation, processus qui consiste à convertir une copie des fichiers originaux dans un format non propriétaire, largement utilisé et respectueux de la préservation. La normalisation standardise les formats des objets numériques ingérés et permet aux archives d'éviter de gérer un grand nombre de formats. Cependant, la normalisation peut également modifier la taille et les propriétés des fichiers. [traduction]. (Scholars Portal (https://learn.scholarsportal.info/all-guides/handling-digital-archives/concepts/#Normalization), s.d.).

## Notation chameau

Écriture sans espace ni ponctuation qui utilise des lettres majuscules afin de distinguer les mots.

## OAIS

Modèle conceptuel publié en 2002, révisé en 2012 (et traduit en 2017), le modèle du système ouvert d'archivage d'information (*Open Archival Information System*, OAIS) établit une série de recommandations pour un système d'information dont le but est de maintenir la capacité d'utilisation des objets numériques au fil du temps. Devenu une norme ISO (ISO 14721) en 2003.

## Objet numérique

Tout morceau d'information, soit unique, soit groupé, qui est stocké par un ordinateur. L'utilisation du terme numérique s'explique parce que toutes les versions des données lisibles par un ordinateur sont codées sous la forme d'une série de uns et de zéros qui sont les seules entrées que les systèmes informatiques peuvent comprendre.

## Objet R

Structure de données qui contient un ensemble de valeurs de type particulier. Les objets R peuvent être créés, modifiés et utilisés pour effectuer des calculs et des analyses.

## Obsolescence des formats

Menace à la longévité des objets numériques basée sur l'incapacité de décoder la séquence de bits qui constitue l'objet numérique. Les menaces d'obsolescence des formats sont souvent traitées par le biais d'un programme d'identification et de validation des formats de fichiers et – au besoin – de la normalisation ou migration des formats obsolètes vers des formats courants.

## Obsolescence des médias

Menace à la longévité des objets numériques basée sur la notion que le média sur lequel ils sont stockés pourrait devenir inutilisable parce que la personne qui veut les utiliser ne détient pas le matériel informatique (ou le logiciel, comme les pilotes de périphérique) nécessaire pour accéder aux données sur le média. Au moment de rédiger ce manuel, l'obsolescence des médias est généralement associée aux disquettes et à une variété de formats de cartouches de données qui, au fil du temps, ne font plus partie de l'usage courant. Les menaces d'obsolescence des médias sont traitées par le biais de méthodes qui assurent l'intégrité au niveau des bits, dont la migration régulière des objets numériques vers des supports modernes plus récents.

## Ontologie

Représentation théorique d'un domaine de connaissances dont les concepts sont liés par des relations sémantiques et logiques.

## OpenRefine

Outil de manipulation de données à code source libre qui nettoie, remodèle et édite par lots les données désordonnées et non structurées.

## Opérationnaliser des variables

Action qui implique l'établissement de définitions mesurables et quantifiables pour des concepts ou des constructions abstraites qui ne peuvent être directement mesurés.

## ORCiD

Identifiant unique pour les membres de la communauté de la recherche. Il est défini par un code numérique permanent ayant deux fonctions principales : lier la personne à ses activités de recherche, dont ses publications, et la distinguer de ses homonymes.

## Outil en ligne de commande

Programme informatique qui peut fonctionner à partir d'une interface en ligne de commande (ILC) d'un système d'exploitation. L'ILC est une interface à base de texte qui permet à une personne d'interagir avec un ordinateur en utilisant des commandes écrites plutôt que d'utiliser une interface graphique avec des menus et des icônes.

## Paquet d'information archivé

Ensemble d'informations, comprenant les informations de contenu et les informations de description de la préservation associée, qui sont préservées dans un système OAIS. (Manuel de préservation numérique (https://www.dpconline.org/docs/digital-preservation/handbook/translations-3/2519-handbook-2021-fr/file), s.d.).

## PCAP®

Acronyme qui signifie propriété, contrôle, accès et possession. Ces quatre principes gouvernent la manière dont les données et l'information relatives aux Premières Nations devraient être collectées, protégées, utilisées et partagées. Les principes PCAP® ont été créés pour combler une lacune dans les lois occidentales qui ne reconnaissent pas les droits des communautés et des peuples autochtones à contrôler leur information.

## Personne responsable de l'intendance des données

Bien que son rôle puisse varier, la personne responsable de l'intendance des données dans un contexte de recherche est chargée de veiller à ce que les données soient traitées de manière systématique et uniforme.

### Personne unique à l'échantillon

Personne dont les renseignements en matière de quasi-identifiants ne correspondent à ceux d'aucune autre personne dans le jeu de données.

### Personne unique à la population

Personne dans une population qui peut être identifiée en raison d'une combinaison unique d'attributs démographiques.

### Perte de la provenance

Menace à la longévité des objets numériques basée sur l'incapacité des membres de la communauté des utilisatrices et utilisateurs à identifier des informations importantes sur l'objet numérique, notamment sa source, l'historique des modifications et ultimement, son authenticité. Les menaces à la provenance d'un objet numérique sont souvent traitées par le biais de la création et de la mise à jour des métadonnées de préservation.

### Photographie oblique

Photographie aérienne prise avec l'axe de la caméra tenu à un angle entre le plan horizontal du sol et le plan vertical perpendiculaire au sol. Une image oblique basse affiche uniquement la surface de la Terre; une image oblique élevée inclut l'horizon. (ESRI (https://support.esri.com/fr-fr/gis-dictionary/oblique-photograph), s.d.).

### Piste de vérification

Documentation qui retrace l'activité et la prise de décision tout au long de la vie d'un projet en détaillant ce qui s'est passé, quand et pourquoi.

### Plan de gestion des données

Description formelle de tout le processus de la chercheuse ou du chercheur, de la collecte des données à leur analyse puis comment elles seront traitées à la fin du projet.

### Politique des trois organismes sur la gestion des données de recherche

Politique qui s'applique aux données générées grâce au financement de la recherche par l'une des trois agences fédérales de financement du Canada. Cette politique vise à encourager l'amélioration de la recherche en obligeant les chercheuses et chercheurs à créer des plans de gestion de données et à préserver leurs données.

## PREMIS

Norme pour les métadonnées ainsi qu'un dictionnaire de données développés pour uniformiser la façon dont les systèmes de préservation enregistrent et comprennent les concepts importants liés à la préservation à long terme d'objets numériques. Les fichiers PREMIS peuvent comprendre des informations techniques (p. ex., l'information sur le format de fichier, les sommes de contrôle) ainsi que des informations sur la provenance (p. ex., les journaux des changements (*changelogs*), les informations sur les acquisitions).

## Prépublication

Version préliminaire d'un article qui n'a pas encore passé le processus d'examen par les pairs, mais qui peut être partagé à des fins de rétroaction. Les prépublications (ou préimpressions) peuvent être considérées comme de la littérature grise.

## Préservation au niveau des bits

Niveau de préservation qui préserve la séquence de uns et de zéros qui compose un objet numérique, mais qui ne traite pas nécessairement de la compréhension des données codées.

## Préservation numérique

Série d'activités gérées nécessaires pour garantir un accès continu aux objets numériques aussi longtemps que nécessaire.

## Principes FAIR

FAIR est un acronyme qui signifie facile à trouver, accessible, interopérable et réutilisable. Les principes directeurs FAIR ont été élaborés en 2014 et visent à améliorer la réutilisation des données, tant par les machines que par les personnes.

## Processus réflexif

Processus par lequel la chercheuse ou le chercheur en recherche qualitative reconnaît, examine et tient compte de l'impact de ses propres jugements, pratiques et croyances sur la collecte et l'analyse des données.

## Programmation lettrée

Affichage de façon linéaire de code, commentaires et sorties, un peu comme une œuvre de littérature.

### Provenance

Documentation faisant référence à la source, l'historique et la propriété d'un artéfact, que celui-ci soit analogique ou numérique.

### Quartile

Valeur qui divise une liste de numéros en quartier.

### Récapitulation entre collègues

Sessions où les membres d'une équipe de recherche se questionnent sur ce qu'elles ou ils ont vu et entendu. Ces discussions peuvent parfois faire partie de l'ensemble final des données de l'étude.

### Recherche computationnelle

Recherche qui dépend des ordinateurs pour la création ou l'analyse des données.

### Recherche itérative

Approche où des révisions ou des modifications font partie intégrante du processus de recherche. Conséquemment, le plan d'étude peut être adapté selon les constats identifiés au fil de la collecte et de l'analyse des données.

### Réduction globale des données

Modification de certaines variables dans l'ensemble d'un jeu de données, par exemple regrouper des réponses en catégories.

### Renseignement identificatoire

Tout renseignement dans un jeu de données qui, seul ou en combinaison avec d'autres renseignements, risque de permettre d'identifier une personne.

### Réplicabilité de la recherche

Caractère d'une recherche qui peut être reproduite par d'autres chercheuses ou chercheurs qui, avec des données différentes ou nouvelles, arriveront à des résultats semblables ou identiques à ceux de la recherche originale.

## Reproductibilité de la recherche

Caractère d'une recherche qui peut être reprise par des chercheuses ou chercheurs qui ne faisaient pas partie de l'équipe de recherche originale, mais qui utilisent les mêmes données pour arriver aux mêmes résultats.

## Rétrocompatibilité

Caractéristique d'un logiciel, d'un programme ou d'un appareil qui fonctionne avec un système avancé, mais qui peut également fonctionner avec les versions antérieures de ce système. (OQLF (https://vitrinel inguistique.oqlf.gouv.qc.ca/fiche-gdt/fiche/8400245/retrocompatibilite), 2008).

## Savoir traditionnel

Connaissances collectives des traditions et des pratiques développées au fil du temps et dont se servent les groupes autochtones pour subvenir à leurs besoins et s'adapter à leur environnement. Le savoir traditionnel est transmis de génération en génération au sein des communautés autochtones. Le savoir autochtone prend diverses formes notamment récits, cérémonies, danses, art, artisanat, chasse, trappage, cueillette, préparation de la nourriture, stockage des aliments, spiritualité, croyances, conceptions du monde et plantes médicinales.

## Schéma de métadonnées

Regroupement d'éléments destinés à décrire une ressource. Pour chaque élément, le nom et la sémantique (la signification de l'élément) sont spécifiés. Les règles de contenu (comment celui-ci doit être formulé), les règles de représentation (par exemple, les règles de capitalisation) et les valeurs d'élément autorisées (par exemple, à partir d'un vocabulaire contrôlé) peuvent être spécifiées en option, mais ce n'est pas toujours le cas.

## Science ouverte

Mouvement visant à rendre la recherche, les données et la diffusion scientifiques transparentes et largement accessibles, sans barrières financières ou autres.

## Sciences sociales

Catégorie métadisciplinaire englobant les disciplines académiques qui utilisent des méthodologies et approches scientifiques pour étudier des phénomènes sociaux, culturels, affectifs et de comportements humains. Des exemples de disciplines de sciences sociales comprennent la sociologie, la science politique, l'économie, la psychologie, les études sur l'information et plus.

### Scripts

Fichier texte qui contient des séquences de commandes dans un langage de programmation particulier (par exemple, R) pouvant être exécutées de façon consécutive.

### Séparateur [informatique]

Caractère spécial réservé par les systèmes ou langages informatiques pour désigner des objets ou éléments indépendants.

### Séquence de bits

Enchaînement précis de bits (0 ou 1) qui ensemble ont une signification (p. ex., un caractère, une opération à effectuer (instruction-machine), une sélection de couleur, un objet numérique).

### Signature [format]

Série de bits qui s'enchaînent de façon prévisible au début, à la fin ou aux deux extrémités d'un fichier.

### Signature numérique

Équivalent d'une signature manuscrite sur papier qui offre des garanties sur l'authenticité de l'identité de la personne signataire.

### Somme de contrôle

Chaînes numériques ou alphanumériques uniques de longueurs potentielles variées produites par un algorithme cryptographique tel que CRC, MD5, SHA1 et SHA256. Aussi appelée empreinte numérique, même la plus petite modification apportée au fichier entraînera une modification complète de la somme de contrôle.

### Source libre

Lorsque du code ou un logiciel est ouvert ou en source libre, les personnes qui l'utilisent sont autorisées à inspecter, utiliser, modifier, améliorer et redistribuer le code sous-jacent. Plusieurs programmeuses et programmeurs utilisent la licence MIT lors de la publication de leur code, ce qui implique que toutes les itérations ultérieures du logiciel incluent également la licence MIT.

### Souveraineté des données autochtones

Droit des peuples autochtones de collecter, d'analyser, d'interpréter, de gérer, de distribuer et de

réutiliser les données auxquelles ils ont accès qui sont dérivées de leurs communautés ou en lien avec elles.

### Stockage actif

Niveau de stockage qui prend en charge les données à l'étape active du projet de recherche, pendant que les données sont créées, modifiées et consultées fréquemment.

### Stockage de type archive

Niveau de stockage qui prend en charge une série d'activités gérées nécessaires pour soutenir la préservation à long terme des documents numériques.

### Stockage de type dépôt

Niveau de stockage qui prend en charge le versement, le stockage, la découverte et l'accès approprié de copies sûres de documents numériques dans divers formats.

### Suppression locale [anonymisation]

Processus utilisé lors de l'anonymisation d'un jeu de données. Le processus implique la suppression de réponses ou de cas individuels.

### Traçabilité de la recherche

Caractère d'une recherche où des chercheuses ou chercheurs externes peuvent comprendre et répéter chacune des modifications apportées aux données brutes pour les préparer à l'analyse.

### Unicode

Standard pour le codage des caractères qui n'est pas lié aux formats ni aux codages des alphabets. Il permet l'échange de textes dans différentes langues.

### Valeur aberrante

Point de données qui diffère de façon importante des autres points d'un jeu de données; elle peut entraîner des problèmes avec certains types de modèles ou d'analyses de données.

### Variable catégorique

Type de données qui représente des catégories discrètes. Les données catégoriques ordinales peuvent être mises dans un ordre ou classées en séquence. Des exemples comprennent les notes de cours qui utilisent

des lettres (p. ex., A, B, C, D, F) et l'échelle de Likert (une échelle avec 5 choix de réponses qui mesurent des constructions latentes ou des phénomènes qui ne peuvent être observés de façon directe). Il existe également des variables catégoriques nominales qui ne peuvent être mises en ordre sur une échelle ou en séquence. Celles-ci peuvent être codées avec des variables factices et incluses dans des analyses quantitatives. Des exemples de variables catégoriques non scalaires comprennent le genre, la race, l'ethnicité, les villes, etc.

## Variable factice

Variable textuelle ou non quantitative à laquelle un chiffre a été attribué à des fins d'analyses quantitatives. Par exemple, un jeu de données qui comprend une variable pour le genre pourrait être codé avec 1 pour l'option féminin, 2 pour masculin, 3 pour non-binaire et 4 pour « préfère ne pas répondre. »

## Versionnage

Permet de garder une trace des modifications apportées à un fichier, aussi petites soient-elles. Également connue sous le nom de contrôle de version, cette opération s'effectue généralement à l'aide d'un système de contrôle de version automatisé tel que GitHub. De nombreux services de stockage de fichiers tels que Dropbox, OneDrive et Google Drive, conservent des versions historiques d'un fichier chaque fois qu'il est enregistré. Il est possible d'accéder à ces versions en consultant l'historique du fichier.

## Vocabulaire contrôlé

Liste de terminologies, mots et expressions utilisés pour indexer ou analyser du contenu et pour retrouver de l'information, généralement dans un domaine spécifique d'information [traduction]. (CODATA Research Data Management Terminology (https://codata.org/rdm-terminology/controlled-vocabulary/), s.d.).

## Vol et exploitation du savoir

Collecte de connaissances autochtones sans demander la permission de partenaires au sein de la communauté ou sans consulter les communautés.

# ANNEXE 1: MODÈLE D'UN PLAN DE GESTION DE DONNÉES

## Collecte de données

- Quels types de données recueillerez-vous, créerez-vous, couplerez-vous, acquérez-vous ou consignerez-vous?
- Dans quels formats de fichiers vos données seront-elles collectées? Ces formats permettront-ils la réutilisation, le partage et l'accès à long terme aux données?
- Quelles conventions et procédures utiliserez-vous pour structurer vos fichiers, les nommer et en contrôler les versions afin de vous aider, vous et toute autre personne, à mieux comprendre la façon dont vos données sont organisées?

## Documentation et métadonnées

- Quelle documentation sera requise pour que les données soient lues et interprétées correctement dans le futur?
- Comment ferez-vous en sorte que la documentation soit créée ou saisie de manière cohérente tout au long de votre projet?
- Si vous utilisez une norme de métadonnées ou des outils pour documenter et décrire vos données, veuillez les énumérer ici.

## Stockage et sauvegarde

- Quels sont les besoins prévus en matière de stockage pour votre projet (en mégaoctets, gigaoctets, téraoctets, etc.) et quelle sera la durée de sauvegarde?
- Comment et où vos données seront-elles stockées et sauvegardées pendant votre projet de recherche?
- De quelle manière l'équipe de recherche et les autres collaboratrices et collaborateurs vont-ils accéder aux données, les modifier et y contribuer tout au long du projet?

## Préservation

- Où déposerez-vous vos données pour les conserver à long terme et y accéder à la fin de votre projet de recherche?
- Indiquez comment vous vous assurerez que vos données soient prêtes pour la conservation. À prendre en considération: formats de fichier appropriés pour la conservation mais qui préservent l'intégrité des données; anonymisation et dépersonnalisation des fichiers, y compris les fichiers de documentation.

## Partage et réutilisation

- Quelles données partagerez-vous et sous quelle forme (p. ex., brutes, traitées, analysées, finales)?
- Avez-vous songé au type de licence d'utilisateur final à inclure avec vos données?
- Quelles mesures prendrez-vous pour faire savoir à la communauté de recherche que vos données existent?

## Responsabilités et ressources

- Désignez qui sera responsable de la gestion des données de ce projet pendant et après le projet et les principales tâches de gestion des données dont cette personne sera responsable.
- De quelle façon allez-vous administrer les responsabilités quant aux activités de gestion des données s'il y a des changements importants de personnel chargé de superviser les données du projet, notamment un changement de la chercheuse ou du chercheur principal?
- De quelles ressources aurez-vous besoin pour mettre en œuvre votre plan de gestion des données? À combien estimez-vous le coût global de la gestion des données?

## Conformité éthique et juridique

- Si votre projet comprend des données sensibles, comment vous assurerez-vous qu'il soit géré de manière sécuritaire et que les données soient accessibles uniquement aux membres approuvés du projet?
- Quelles stratégies allez-vous mettre en œuvre pour gérer les utilisations secondaires des données sensibles, le cas échéant?
- Comment allez-vous gérer les questions juridiques, éthiques et de propriété intellectuelle?

*– Adapté du modèle Portage (https://assistant.portagenetwork.ca/public_templates), sous licence Creative Commons Attribution 4.0 (https://creativecommons.org/licenses/by/4.0/deed.fr).*

# ANNEXE 2: UN EXEMPLE D'UNE SECTION COMPLÉTÉE DU MEMAC

À noter: les informations ci-dessous ont été anonymisées.

## Catégorie – Politiques et procédures de l'établissement

Ce domaine d'activité couvre l'élaboration et la mise à jour des politiques en matière de gestion des données de recherche (GDR), ainsi que les procédures qui s'appliquent au soutien des services de GDR.

Quelques pistes de réflexion qui auront un impact sur votre évaluation :

- Portée (p. ex., l'intendance des données, la destruction des dossiers, la sécurité et la protection, etc.);
- Lignes directrices du comité d'éthique de la recherche (CER);
- Plan de sensibilisation;
- Autres documents de l'établissement qui contiennent des éléments pertinents.

**Niveaux de maturité:**

Sans objet
○ Passer cet élément
0 – N'existe pas OU inconnu
1 – L'élément est informel ou ponctuel.
  ◦ Les politiques et procédures peuvent être peu élaborées, dépassées et/ou incohérentes.
  ◦ Certaines politiques connexes peuvent exister, mais elles sont insuffisantes.
2 – L'élément est en cours d'élaboration.
  ◦ Les politiques et procédures sont en cours de conceptualisation et de formulation.
3 – L'élément est fonctionnel et lancé.
  ◦ Les politiques et procédures sont définies et normalisées.
4 – L'élément est robuste et se concentre sur l'évaluation continue.
  ◦ Les politiques et procédures font l'objet d'une révision et d'une amélioration.

**Échelle:**

Sans objet – si 0 ou SO sont choisis pour le niveau de maturité

1 – Offert uniquement à certains utilisateurs sur demande.

2 – Disponible dans certaines unités ou cohortes.

3 – Disponible pour tous.

| **Catégorie : Politiques et procédures de l'établissement** | | | | |
|---|---|---|---|---|
| **Élément** | **Définitions** | **Niveau de maturité** | **Échelle** | **Vos commentaires** |
| Stratégie de GDR de l'établissement | Selon la définition des trois organismes subventionnaires. Comprend notamment toute feuille de route institutionnelle de GDR précisant la manière dont la stratégie sera mise en œuvre. | 1/2 | 3 | Groupe de travail a été mis en place.<br>Modalités et conditions du groupe consultatif ont été rédigées. |
| Politiques de l'établissement en matière de GDR | Comprend toute politique pertinente de l'établissement qui peut traiter de la GDR ou des éléments liés à la GDR | 1 | 3 | Peut-être STI / Cybersécurité<br>« La conduite responsable de la recherche » (lien retiré)<br>Lignes directrices du CER |
| Procédures et directives relatives à la planification de la gestion des données | Toute procédure ou ligne directrice de l'établissement pour les chercheuses et chercheurs sur la manière d'aborder les plans de gestion de données (p. ex., les attentes concernant la création, la soumission et/ou l'examen des PGD). | 3 | 3 | Bibliothécaire de GDR<br>Ressources actuelles du Réseau<br>Portage<br>Comité institutionnel de calcul informatique<br>STI et cybersécurité institutionnelle |
| Politiques et procédures de sécurité et d'évaluation des risques | Toute procédure ou politique de l'établissement qui traite de la sécurité et de l'évaluation des risques liés aux données de recherche (p. ex., les questions juridiques ou de confidentialité, les évaluations de vulnérabilités, etc.) | 2.5 | 2 | La sécurité du service de STI effectue l'évaluation – Évaluation complète de la sécurité pour Sharefile<br>Plan d'évaluation des risques<br>Responsabilités de l'utilisateur<br>Politique pour les utilisateurs des TI |
| Plan de communication et de sensibilisation | Tout plan de promotion de la GDR, notamment la sensibilisation aux politiques et lignes directrices nationales touchant la GDR (les politiques des trois organismes, des bailleurs de fonds, des publications savantes, etc.) et la prestation de liens et de ressources pour les bonnes pratiques et les outils. | 1 | 3 | Bibliothèque – GDR<br>Webinaires<br>Site Web – Stratégie institutionnelle<br>Réflexions sur un plan à long terme (encourage la participation de tout l'établissement) |

*Nom et rôle de la(des) personne(s) ayant rempli ce tableau:*

STI, Bibliothèque, Bureau de la recherche

# ANNEXE 3: EXERCICES DU CHAPITRE 10

## Introduction

L'objectif de cet exercice est de démontrer la relation entre les données ouvertes, les **cahiers de laboratoire électroniques** et les conteneurs informatiques dans la recherche reproductible. Vous allez interagir avec le code d'un cahier de laboratoire hébergé sur GitHub et rendu interopérable à l'aide de myBinder. La plupart des principes fondamentaux décrits au chapitre 10 seront illustrés ici.

Cet exercice inclut une activité d'introduction et une activité avancée. Dans l'activité d'introduction, vous allez explorer le code sur GitHub et examiner une version statique d'un cahier de laboratoire. Dans l'activité avancée, vous lancerez un conteneur informatique dans une interface appelée Binder. Le conteneur héberge un cahier de laboratoire électronique qui interroge un jeu de données ouvert. Vous pouvez interagir avec le jeu de données en ligne sans modifier la copie originale. Le conteneur en ligne vous permet d'exécuter le code sans installer de logiciels sur votre ordinateur. L'activité avancée nécessite des connaissances plus poussées en matière de codage ou simplement de la persévérance. Le conteneur informatique ne se charge pas toujours du premier coup et le code ne fonctionnera pas s'il n'est pas parfaitement saisi. Cet exercice a pour but de montrer les avantages et la complexité de la recherche reproductible. N'hésitez pas à chercher sur Google les termes que vous ne comprenez pas. De plus, ChatGPT est un bon outil pour expliquer le code et son fonctionnement.

À la toute fin de l'exercice, il y a une question de réflexion. Vous pouvez répondre à cette question même si vous n'avez pas fait l'activité avancée.

## Partie 1 (introduction) : explorer les données et le dépôt de codes

Le Programme international pour le suivi des acquis des élèves (PISA) (https://www.oecd.org/pisa/) est une initiative internationale qui mesure les réalisations en éducation des élèves de 15 ans. Ce jeu de données (https://www.oecd.org/pisa/data/) (en anglais uniquement) libre d'accès est disponible aux chercheuses et chercheurs qui mènent leurs propres analyses. Cette activité utilise une analyse du jeu de données du PISA menée par Klajnerok (2021) et publiée dans GitHub à l'aide d'un cahier Jupyter.

Pour les fins de cette activité, nous avons créé une fourche à partir du dépôt de ce projet dans notre propre

dépôt Git : https://github.com/mediagestalt/PISA (https://github.com/mediagestalt/PISA) (en anglais uniquement). Dans GitHub, une **fourche** (ou *fork*) représente une copie d'un jeu de données qui conserve son lien aux créateurs originaux (Documentation GitHub, s.d.). Dans la capture d'écran ci-dessous, vous pouvez voir le symbole de fourche et un lien vers le jeu de données qui le précède. Ces liens sont importants puisqu'ils démontrent la **provenance** du jeu de données.

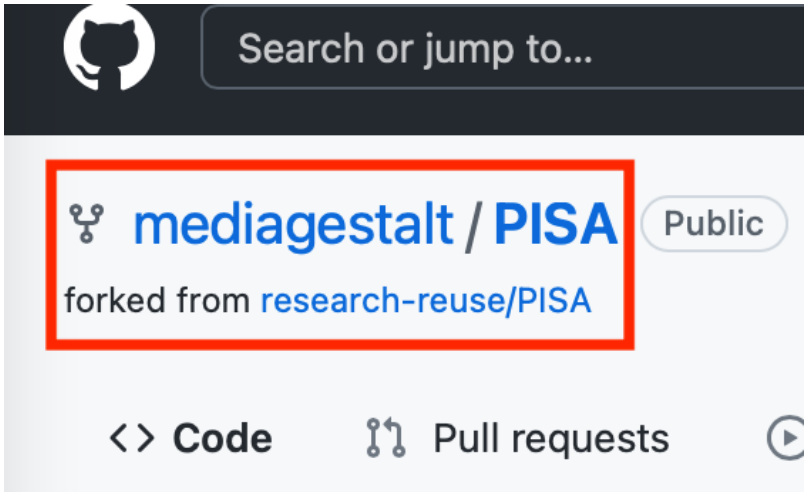



Figure 1. Fourche d'un projet dans un dépôt Git.

**QUESTION 1:** Quel est le nom du dépôt où ce code a été déposé à l'origine?

**Réponse:** Le créateur d'origine de ce code est : <https://github.com/mklajnerok/PISA (https://github.com/mklajnerok/PISA)>.Pour ce projet, le code et les données ont été réutilisés par <https://github.com/research-reuse/PISA (https://github.com/research-reuse/PISA)> et placées dans un conteneur informatique appelé Binder. Cette activité est une fourche de <https://github.com/research-reuse/PISA (https://github.com/research-reuse/PISA)> et a été adaptée pour ce manuel. Le jeu de données original a été publié par PISA.

Vous pouvez naviguer sur GitHub comme vous le feriez avec tout autre répertoire de fichiers imbriqués. La figure 2 est une capture d'écran de GitHub. Les noms de fichiers figurent dans la colonne de gauche; la colonne du milieu montre le commentaire qui décrit les dernières modifications apportées au fichier et la colonne de droite indique la dernière fois que le fichier a été modifié. Vous pouvez également voir la dernière

personne qui a contribué au dépôt de code en haut à gauche du tableau et les informations sur la version en haut à droite du tableau. Celles-ci sont illustrées dans la figure 2 avec le terme « *83 commits.* »

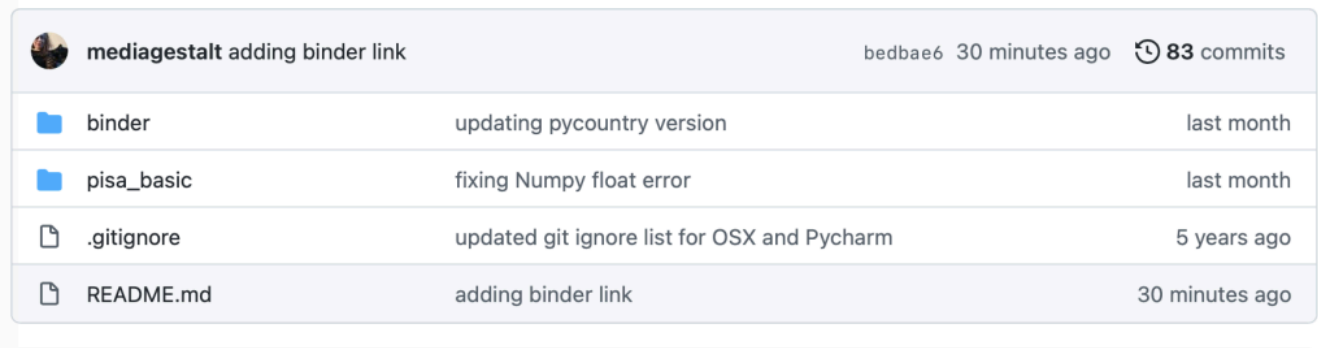



Figure 2. Dossier GitHub.

## Dossiers GitHub

Pour la prochaine question, trouvez les fichiers suivants dans le dépôt du PISA. Vous trouverez les fichiers dans différents dossiers, donc n'hésitez pas à fouiller.

- `requirements.txt`
- `pisa_project_part1.ipynb`

Cliquez sur le titre du fichier pour l'afficher. Puis, défilez vers le bas pour regarder le contenu de chacun des fichiers. Vous cherchez une liste de dépendances, c'est-à-dire les progiciels nécessaires a l'exécution du code dans le cahier. Dans le fichier *pisa_project_part1.ipynb*, vous trouverez la liste sous le titre « *Extracting PISA dataset* » comme le montre l'image ci-dessous.

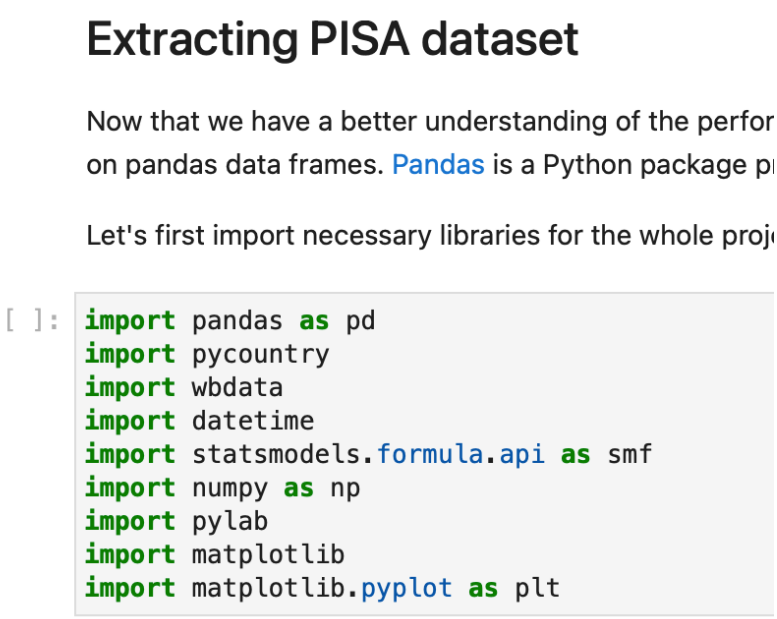



Figure 3. Dépendances du cahier électronique.

**QUESTION 2:** Comparez les dépendances énumérées dans le fichier *requirements.txt* avec celles du cahier *pisa_project_part1.ipynb*. De quelles façons sont-elles différentes?

**Réponse:** Le fichier *requirements.txt* comprend des numéros de version des dépendances tandis que le fichier du carnet ne fait que lister leur nom. L'information sur le versionnage est très importante pour les dépendances, car des changements inconnus effectués sur le code peuvent l'empêcher de fonctionner correctement. Il s'agit ici d'un scénario où mettre à jour la plus récente version d'un programme n'est pas la meilleure option. Effectuer la curation de code à des fins de réutilisation équivaut essentiellement à « geler » le code dans le temps afin qu'il roule exactement de la même manière que lorsqu'il a été créé.

Les noms de fichiers et leurs répertoires démontrent l'importance des chemins d'accès relatifs. Dans le répertoire Git, trouvez l'emplacement des fichiers CSV suivants. Faites-les correspondre à l'endroit où ils sont nommés dans le fichier du cahier.

- *pisa_math_2003_2015.csv;*

- *pisa_read_2000_2015.csv;*
- *pisa_science_2006_2015.csv*. Indice : les fichiers se trouvent dans la deuxième cellule de code sous les dépendances.

## Partie 2 (avancée): Exécuter et modifier le code

C'est le temps d'explorer le conteneur informatique. Puisque le chercheur original a écrit le code dans un cahier Jupyter (une sorte de cahier de laboratoire électronique couramment utilisé), il est possible de mettre en conteneur le code et les données pour qu'ils puissent être utilisés par d'autres.

Revenez à la page principale du dépôt GitHub aussi connu sous le nom de « *README* » (ou **LISEZ-MOI**). Cliquez sur le bouton « *launch binder* », tel qu'illustré dans la figure 4.

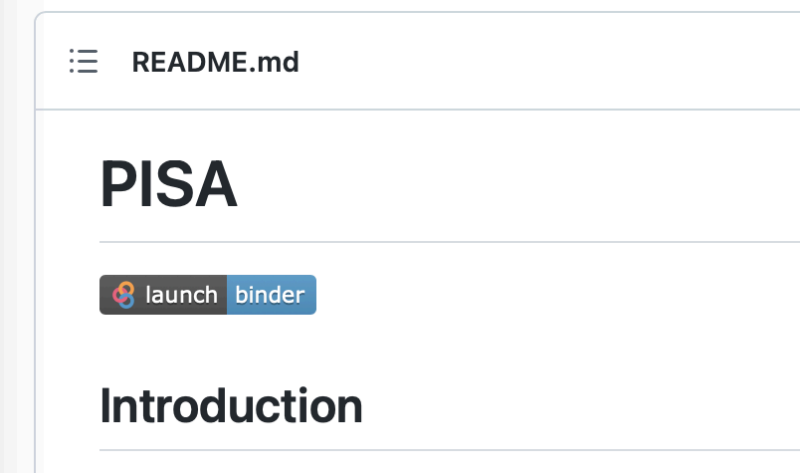



Figure 4. Bouton « *launch binder.* »

Selon votre ordinateur ou votre vitesse Internet, le chargement du conteneur peut prendre quelques minutes. Si cela prend trop de temps, fermez la page et essayez de lancer à nouveau le chargement à partir du lien GitHub Binder. Vous pouvez voir l'écran de chargement du Binder à la figure 5.

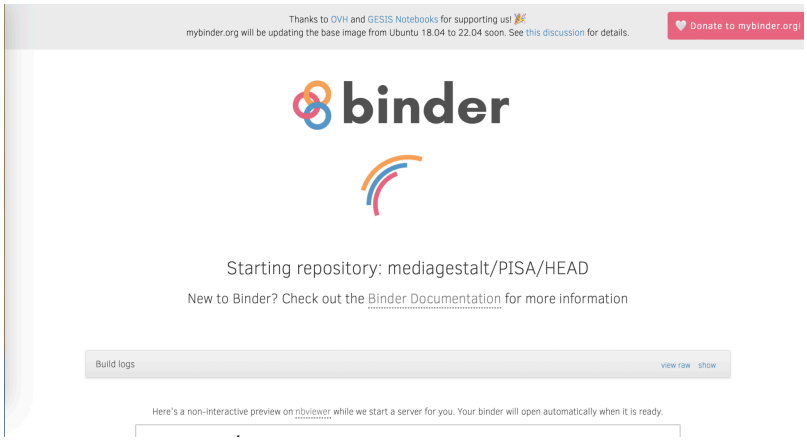



Figure 5. Lancement de binder.

Lorsque le chargement du cahier est terminé, défilez vers le bas et explorez la page. Le cahier chargé ressemble exactement au cahier consulté dans le dépôt Github.

Alors que vous examinerez le cahier, vous verrez du texte narratif parsemé de cellules de blocs de code. Des commentaires supplémentaires se retrouvent à l'intérieur des cellules de code. Il s'agit là d'un exemple de **programmation lettrée**.

Afin de faciliter la suite de l'activité, activez la fonction de numérotation des lignes dans le fichier. Les numéros de chacune des lignes de blocs de code seront affichés, permettant d'identifier plus facilement des lignes de code spécifiques. L'emplacement de cette commande est indiqué à la figure 6. Vous ne verrez pas de changement immédiat sur la page, car il s'agit simplement d'un changement de paramètre.

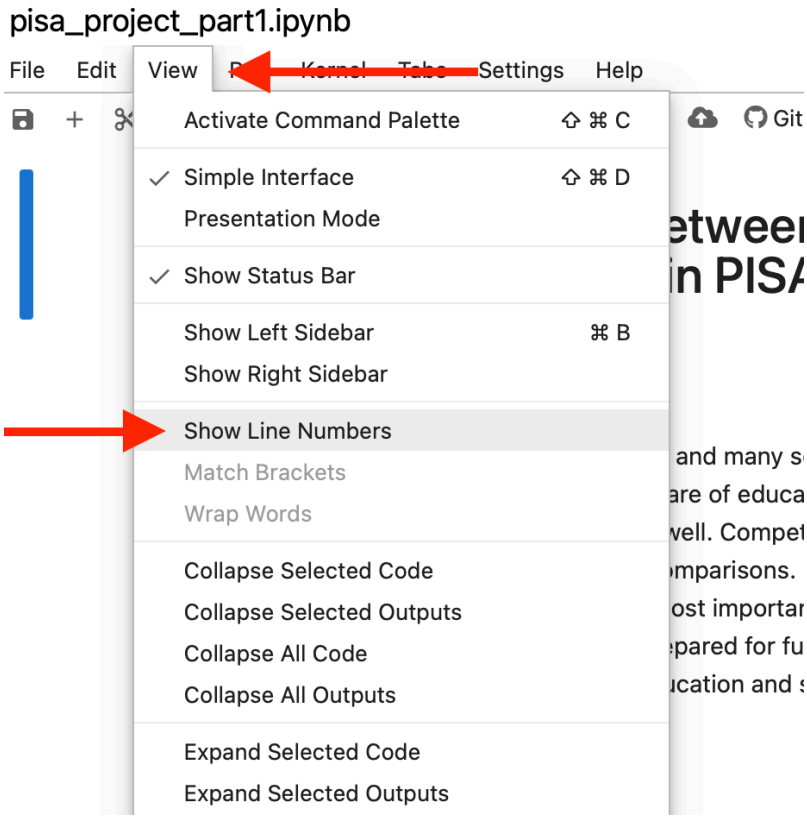



Figure 6. Activation de la fonction de numérotation des lignes.

C'est maintenant le temps d'exécuter le code. Pour débuter, exécutez toutes les cellules de code. L'emplacement de cette commande est indiqué à la figure 7. En faisant défiler la page, vous trouverez du nouveau contenu sous quelques-uns des blocs de code. Il s'agit des résultats de l'analyse pour laquelle le code a été écrit. Il peut s'agir de texte, de tableaux ou de visualisations.

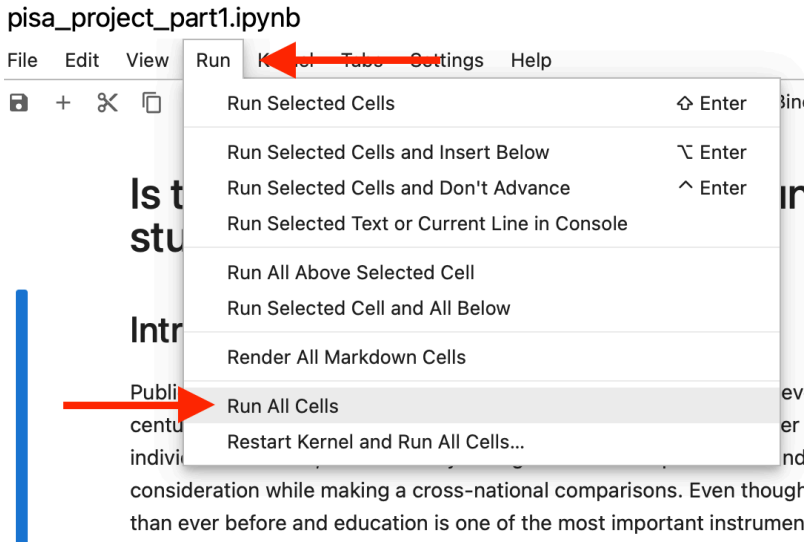



Figure 7. Exécution de toutes les cellules.

Vous verrez également un numéro entre crochets dans la marge gauche à côté de chaque bloc de code. Commencez la lecture au début de la page et rendez-vous à la cellule numéro 6. **Si vous n'arrivez pas à comprendre le code, ne vous en faites pas**. Portez plutôt votre attention sur les descriptions textuelles et sur les commentaires à l'intérieur des cellules. Un commentaire est facile à identifier parce qu'il est précédé du symbole [#] ou [“””]. Lisez les descriptions narratives jusqu'à la cellule n° 6. Voir la figure 8.

The latest PISA results were collected in year 2015, so we can use a filter to extract those rows. In the n
reduced significantly the amount of data. We can now merge all the data frames in one, so we have resu
single row.

[6]:

```
def filter_dict_by_year(df_dict, year):
    """Create a copy of df_dict and extract rows for a given year
    :param df_dict: dictionary with string keys and data frames values
    :param year: int
    :returns df_dict_year: dictionary with string keys and data frames values
    """
    df_dict_year = df_dict.copy()
    for k, v in df_dict_year.items():
        v = df_dict_year[k]
        v = v[v['Time'] == year]
        df_dict_year[k] = v
    return df_dict_year

#extract PISA results for 2015
pisa_2015 = filter_dict_by_year(pisa_data, 2015)

def merge_dict_by_year(df_dict_year):
    """Take df_dict_year and merge each data frame in one based on Code,
    then drop columns with year number
```

Figure 8. Cellule 6.

**QUESTION 3: Que dit le commentaire sur la ligne 14 de la cellule 6?**

**Réponse:** « *#extract PISA results for 2015.* » C'est-à-dire, extraire les résultats du PISA pour 2015. *Indice: si vous ne l'avez pas trouvé, utilisez la fonction « Rechercher » de votre navigateur pour rechercher la phrase. Vous verrez alors le numéro de la ligne et de la cellule.*

Le jeu de données du PISA pour ce projet comporte des données qui remontent à l'année 2000. Nous pouvons charger plus de données en modifiant le code. Pour la suite de cette activité, vous devez ajouter un nouveau code au cahier de laboratoire électronique et réexécuter le bloc de code. Pour obtenir les lignes de

code supplémentaires, consultez cet extrait de code (https://gist.github.com/mediagestalt/78de91092f21ad8b279f0f07f961a2f2) (appelé un Gist) sur GitHub. Il s'agit d'une version modifiée de la cellule 6 du cahier.

La ligne 14 du Gist et la ligne 14 du bloc de code 6 du cahier sont identiques. Le '#' devant le texte signifie que la ligne est un commentaire et non du code. C'est à la ligne 15 que le code commence. Dans ce Gist, il y a des lignes de code supplémentaires en dessous de la ligne 15 qui n'apparaissent pas dans le cahier. Copiez le code des lignes 16 et 17 et collez-les dans le cahier. Assurez-vous que le cahier correspond aux lignes 14-17 du Gist.

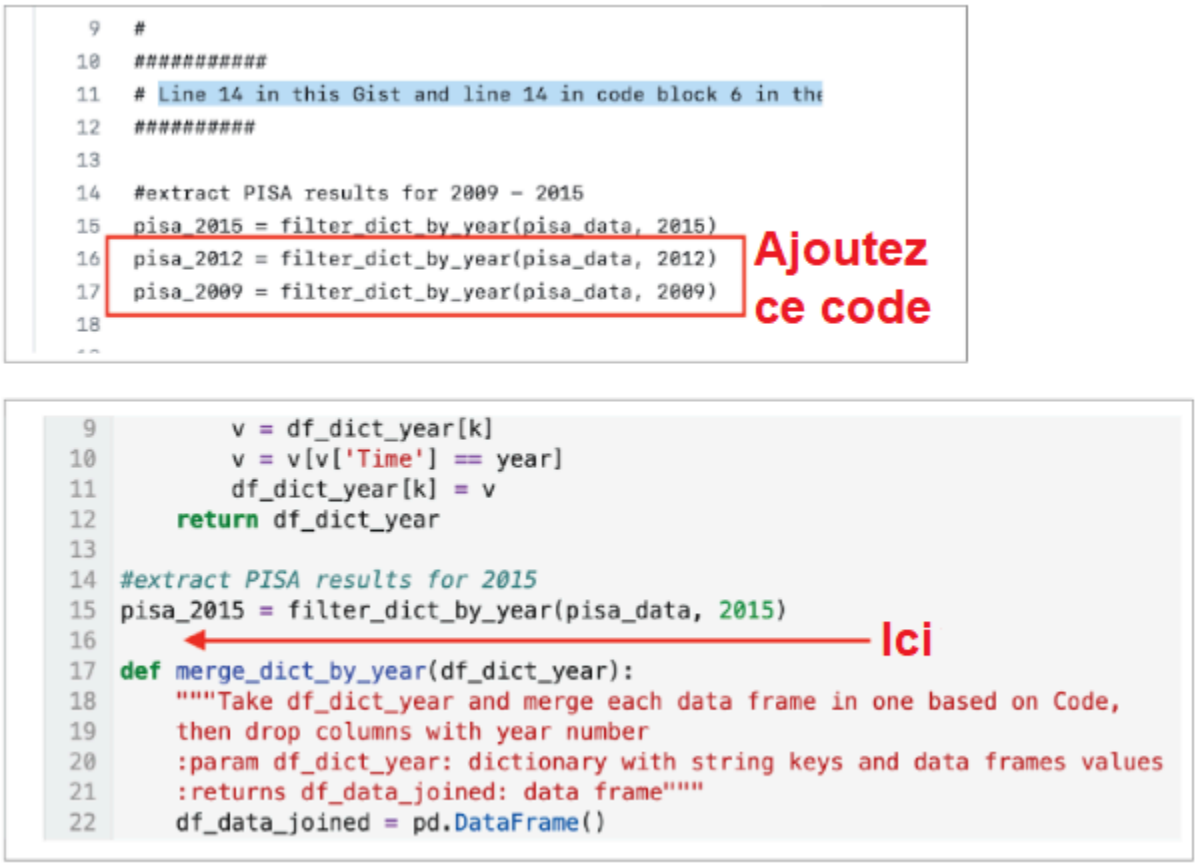



Figure 9. Code du Gist.

Ce code fait appel au jeu de données PISA. Avant l'ajout des lignes supplémentaires, les données de PISA ne concernaient que l'année 2015. L'ajout des deux lignes de code supplémentaires importe des années supplémentaires de données PISA, soit 2012 et 2009. Si vous souhaitez expérimenter davantage, vous pouvez ajouter des lignes supplémentaires avec des années différentes. Veillez simplement à suivre le format à la lettre.

Il ne suffit pas d'ajouter ces lignes. Vous allez devoir suivre le même processus pour les lignes 31 et 40. Ce code et d'autres instructions peuvent également être trouvés dans le Gist. **Notez que les numéros de ligne dans le cahier changeront lorsque vous ajouterez du code supplémentaire.**

Une fois que les paramètres supplémentaires ont été ajoutés au cahier, exécutez une nouvelle fois la cellule 6 du cahier en cliquant sur la cellule et en appuyant sur la touche Majuscule + Retour. En cas d'erreurs, vérifiez que votre code ne contient pas de fautes de frappe et essayez à nouveau. Vous pouvez également utiliser la commande de menu « *Run* > *Run Selected Cells* ».

**À partir de maintenant, les numéros de cellules du cahier vont changer selon le nombre de fois que vous exécutez le code de la cellule.**

Ensuite, laissez votre curseur sur la cellule que vous venez de modifier et insérez une nouvelle cellule pour chacune des années supplémentaires que vous avez ajoutées.

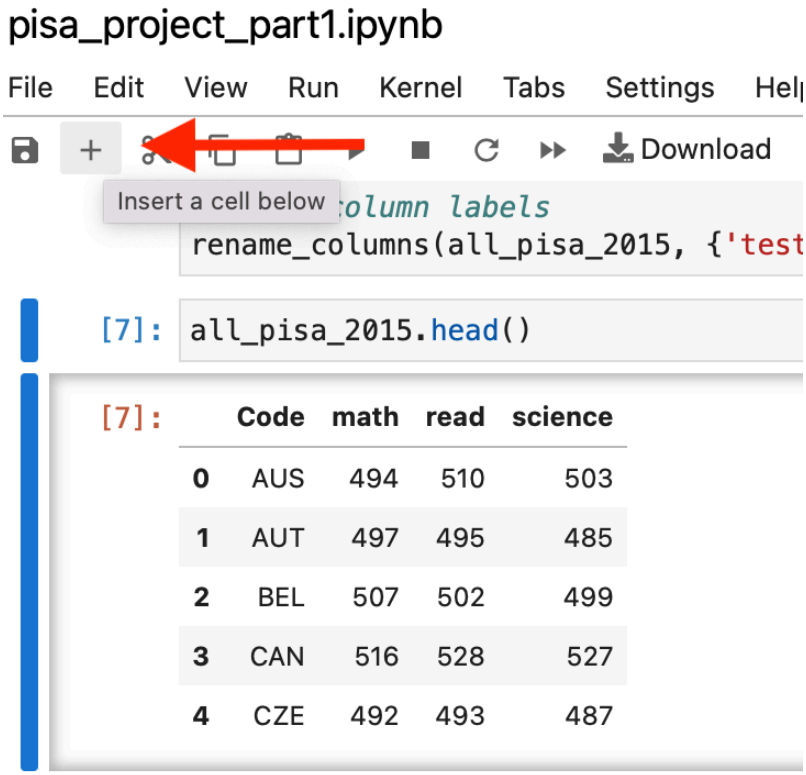



Figure 10. Ajout de nouvelles cellules.

Dans les nouvelles cellules, inscrivez le nom des variables supplémentaires pour chacune des années que vous avez ajoutées et appuyez sur la touche Majuscule + Retour pour exécuter chacune d'elles.

Par exemple:
all_pisa_2012.head()
all_pisa_2009.head()

En cas d'erreurs, cherchez les fautes de frappe et essayez à nouveau.

Essayez de voir combien de cellules de plus vous pouvez faire fonctionner! Tant avec les variables existantes qu'avec les nouvelles que vous avez créées.

Si vous faites une erreur qui entraîne une défaillance irréparable dans le code, vous pouvez consulter le fichier source pour recopier et coller le code original. Vous pouvez aussi recharger le fichier au complet en sélectionnant « *File* > *Reload Notebook from Disk* » dans le menu principal du carnet.

**Questions de réflexion**

Selon ce que vous avez retenu du chapitre 10 et par votre exploration du conteneur informatique, quelles modifications apporteriez-vous à la structure du répertoire de fichiers pour en améliorer l'organisation ? Les données et logiciels sont-ils suffisamment documentés ? Passez à travers ce cadre pour la reproductibilité: *Reproducibility Framework (https://docs.google.com/document/d/1E0c5-DDVo2MMoF2rPOiH2brIZyC_3YZZrcgp0x6VCPs/edit)* (Khair *et al.*, 2019) pour vous aider dans votre évaluation.

1. La provenance de ces données est-elle claire pour vous? Expliquez.
2. Quelles caractéristiques de ce jeu de données ont facilité sa reproductibilité? Qu'est-ce que vous voudriez améliorer?

# Bibliographie

Documentation GitHub. (s.d.). *Fork a repo*. https://docs.github.com/fr/get-started/quickstart/fork-a-repo (https://docs.github.com/fr/get-started/quickstart/fork-a-repo)

Klajnerok, M. (2021, 23 novembre). *Is there a relationship between countries' wealth or spending on schooling and its students'*. Towards Data Science. https://towardsdatascience.com/is-there-a-relationship-between-countries-wealth-or-spending-on-schooling-and-its-students-a9feb669be8c (https://towardsdatascience.com/is-there-a-relationship-between-countries-wealth-or-spending-on-schooling-and-its-students-a9feb669be8c)

Khair, S., Sawchuk, S. et Zhang, Q. (2019) *Reproducibiity Framework*. https://docs.google.com/document/d/1E0c5-DDVo2MMoF2rPOiH2brIZyC_3YZZrcgp0x6VCPs/edit (https://docs.google.com/document/d/1E0c5-DDVo2MMoF2rPOiH2brIZyC_3YZZrcgp0x6VCPs/edit)

# SOLUTIONNAIRE

## Chapitre 7, Le nettoyage de données dans le processus de gestion des données de recherche

| ID | AGE | ETABLISSEMENT | NOTE |
|---|---|---|---|
| 1 | 17 | Université de Guelph | 88 |
| 2 | 21 | Université de Guelph | 60 |
| 3 | 18 | Université de Guelph | 80 |
| 4 | 19 | Université de Guelph | 75 |
| 8 | 18 | Université de Guelph | 72 |
| 12 | 21 | Université de Guelph | 60 |
| 13 | 18 | Université de Guelph | 80 |
| 14 | 19 | Université de Guelph | 77 |
| 15 | 18 | Université de Guelph | 49 |
| 16 | 21 | Université de Guelph | 60 |
| 17 | 18 | Université de Guelph | 88 |
| 19 | 19 | Université de Guelph | 73 |
| 20 | 18 | Université de Guelph | 72 |

Solution à l'exercice du conseil no 1

| ID | AGE | ETABLISSEMENT | NOTE |
|---|---|---|---|
| 1 | 17 | Université de Guelph | 88 |
| 2 | 21 | UOG | 60 |
| 3 | 18 | Université de Guelph | 80 |
| 4 | 19 | Université de Guelph | 75 |
| 12 | 21 | Université de Guelph | 60 |
| 13 | 18 | Université de Guelph | 80 |
| 14 | 19 | Guelph University | 77 |
| 15 | 18 | Université de Guelph | 49 |
| 16 | 21 | U of G | 60 |
| 17 | 18 | Université de Guelph | 88 |
| 19 | 19 | Guelph University | 73 |
| 20 | 18 | Université de Guelph | 72 |

Solution à l'exercice du conseil no 2

| ID | OISEAU | LIEU | TOTAL |
|---|---|---|---|
| 1 | 17 | chemin Québec | 6 |
| 2 | 21 | chemin Cork | 5 |
| 3 | 18 | chemin March | 8 |
| 4 | 19 | chemin Victoria | 5 |
| 5 | 18 | chemin Steffler | 8 |
| 6 | 21 | chemin Extra | 0 |
| 7 | 18 | chemin Doyle | 2 |
| 8 | 19 | chemin Oxford | 7 |
| 9 | 18 | chemin Dublin | 4 |
| 10 | 21 | chemin First | 6 |
| 11 | 18 | chemin Church | 1 |
| 12 | 19 | chemin North | 3 |
| 13 | 18 | chemin Dulac | 2 |

Solution à l'exercice du conseil no 3

| ID | AGE | NOM | COURRIEL |
|---|---|---|---|
| 1 | 17 | James Smith | jsmith@gmail.com |
| 2 | 21 | Michael Smith | msmith@gmail.com |
| 3 | 18 | Robert Smith | smithr@aol.com |
| 4 | 19 | Maria Garcia | mgarcia@hotmail.com |
| 8 | 18 | David Smith | davidsmith@gmail.com |
| 12 | 21 | Maria Rodriguez | mariar@gmail.com |
| 13 | 18 | Mary Smith | marysmith@gmail.com |
| 14 | 19 | Maria Hernandez | hernandez@outlook.com |
| 15 | 18 | Maria Martinez | mmartinez@mail.com |
| 16 | 21 | James Johnson | james@gmail.com |
| 17 | 18 | Lee Hartman | hartman@mail.com |
| 19 | 19 | Patricia Smith | smithp@mail.ca |
| 20 | 18 | Ben Smith | bensmith@mail.com |

Solution à l'exercice du conseil no 4

| ID | OISEAU | LIEU | JUVENILE | JUVENILE_NUM |
|---|---|---|---|---|
| 1 | rouge-gorge | ch Québec | non | 0 |
| 2 | hirondelle | rue Cork | oui | 1 |
| 3 | corbeau | ch March | non | 0 |
| 4 | pigeon | chemin Victoria | non | 0 |
| 5 | corbeau | ch Steffler | non | 0 |
| 6 | corbeau | ch Extra | oui | 1 |
| 7 | rouge-gorge | ch Doyle | oui | 1 |
| 8 | rouge-gorge | chemin Oxford | non | 0 |
| 9 | corbeau | ch Dublin | non | 0 |
| 10 | pigeon | chemin First | non | 0 |
| 11 | pigeon | chemin Sixth | oui | 1 |
| 12 | pigeon | Ch Church | non | 0 |
| 13 | hirondelle | ch Dulac | oui | 1 |

Solution à l'exercice du conseil no 6

| ID | AGE | PRENOM | NOM | COURRIEL |
|---|---|---|---|---|
| 1 | 17 | James | Smith | jsmith@gmail.com |
| 2 | 21 | Michael | Smith | msmith@gmail.com |
| 3 | 18 | Robert | Smith | smithr@aol.com |
| 4 | 19 | Maria | Garcia | mgarcia@hotmail.com |
| 8 | 18 | David | Smith | davidsmith@gmail.com |
| 12 | 21 | Maria | Rodriguez | mariar@gmail.com |
| 13 | 18 | Mary | Smith | marysmith@gmail.com |
| 14 | 19 | Maria | Hernandez | hernandez@outlook.com |
| 15 | 18 | Maria | Martinez | mmartinez@mail.com |
| 16 | 21 | James | Johnson | james@gmail.com |
| 17 | 18 | Lee | Hartman | hartman@mail.com |
| 19 | 19 | Patricia | Smith | smithp@mail.ca |
| 20 | 18 | Ben | Smith | bensmith@mail.com |

Solution à l'exercice du conseil no 8

| ID | OISEAU | LIEU |
|---|---|---|
| 1 | rouge-gorge | chemin Québec |
| 3 | corbeau | chemin March |
| 4 | pigeon | chemin Victoria |
| 5 | corbeau | chemin Steffler |
| 8 | rouge-gorge | chemin Oxford |
| 9 | corbeau | chemin Dublin |
| 10 | pigeon | chemin First |
| 12 | pigeon | chemin Chruch |

Solution à l'exercice du conseil no 9

# Chapitre 8, Nouvelles aventures en nettoyage des données

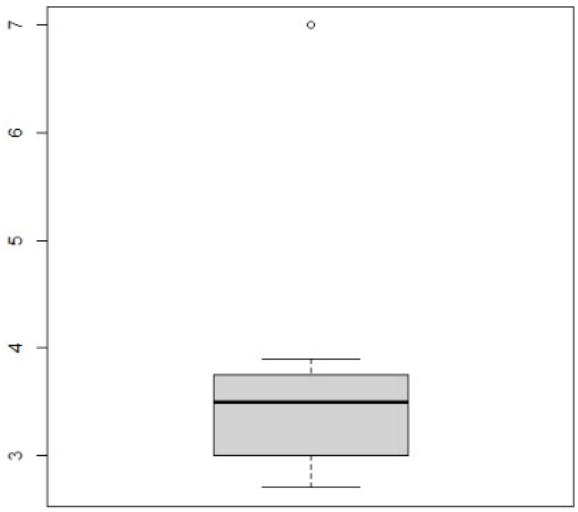
**Réponse à la question 1 :**

| | A | B | C |
|---|---|---|---|
| 1 | **Données importées** | **Formule** | **Résultats** |
| 2 | L' | =CONCATENER(SUPPRESPACE(A2)," ",SUPPRESPACE(A3)," ",SUPPRESPACE(A4)) | L'école de service social |
| 3 | école de | | |
| 4 | service social | | |
| 5 | | | |
| 6 | 3 ou plus | =VALEURNOMBRE(GAUCHE(A6,1)) | 3 |
| 7 | | | |
| 8 | à la | =NOMPROPRE(A8) | À La |
| 9 | | | |
| 10 | !!Rapport mensuel!! | =EPURAGE(A10) | Rapport mensuel |
| 11 | | | |
| 12 | Taxe | =MINUSCULE(A12) | taxe |
| 13 | | | |
| 14 | SACR-126 | =DROITE(A14,3) | 126 |
| 15 | | | |
| 16 | 789 | =TEXTE(A16,"00000") | 00789 |
| 17 | | | |

**Réponse à la question 2 :**

Selon la boîte à moustache, il existe une valeur aberrante. Nous remplaçons cette valeur par « NA ». Nous calculons ensuite la moyenne en éliminant tous les « NA ».

```
> boxplot(mydata_csv$Largeur)
```




Boîte à moustache de l'exercice 2

```
> summary(mydata_csv$Largeur)
```

```
Min.   1st Qu.  Median  Mean   3rd Qu.  Max.   NA's
2.700  3.000    3.500   3.814  3.750    7.000  1
```

```
> mydata_csv$Largeur[mydata_csv$Largeur==7] = NA
> mean(mydata_csv$Largeur, na.rm = T)
```

```
[1] 3.283333
```

# Chapitre 13, Les données sensibles: des considérations pratiques et théoriques

**Réponse à la question 1 :**

*Énoncé de politique des trois conseils : Éthique de la recherche avec des êtres humains* – EPTC 2

**Réponse à la question 2 :**

Les identifiants directs sont les suivants:

- nom entier ou partiel ou initiales;
- dates précises d'événements personnels tels que la naissance, la graduation, l'admission à l'hôpital (seul le

mois ou l'année peut être acceptable);
- adresse complète ou partielle (de grandes zones géographiques, telles que des villes, appartiennent à la catégorie d'identifiants indirects et doivent être révisées);
- code postal complet ou partiel (les trois premiers chiffres peuvent être acceptables);
- numéros de téléphone ou de télécopieur;
- adresse courriel;
- identifiants ou noms d'utilisateurs Web ou de médias sociaux tels que le pseudonyme Twitter;
- numéros de protocole Internet ou IP; renseignements précis relatifs au navigateur Web ou au système d'exploitation (ces informations peuvent être recueillies par certains types d'outils de sondage ou de formulaires Web);
- identifiants de véhicule tels que la plaque d'immatriculation;
- identifiants liés à des dispositifs médicaux ou autres;
- tout autre numéro d'identification unique lié directement ou indirectement à un individu tel que le numéro d'assurance sociale, numéro d'étudiant ou numéro d'identification d'un animal de compagnie;
- photos d'individus ou de leur domicile ou emplacement; des enregistrements vidéos les montrant; des images médicales;
- enregistrements audio de personnes (Han *et al*., 2020);
- données biométriques;
- tout attribut personnel unique ou reconnaissable (p. ex., maire de Kapuskasing ou gagnant du prix Nobel).

Les quasi-identifiants peuvent comprendre l'un des éléments suivants :

- âge (peut être un identifiant direct dans le cas de personnes très âgées);
- identité de genre;
- revenu;
- emploi ou secteur d'activité;
- variables géographiques;
- variables ethniques ou d'immigration;
- appartenances à des organismes ou utilisation de services particuliers.

Il existe de nombreux autres exemples !

**Réponse à la question 3 :**

La combinaison de la latitude et de la longitude ou les informations sur la proximité du site industriel le plus proche sont des variables de localisation qui peuvent indiquer avec précision où se trouve une espèce menacée et peuvent être considérées comme des données sensibles.

# Chapitre 14, La gestion des données de recherche qualitatives

**Réponse à la question 1 :**

- Peuvent mener à la réidentification des personnes participantes;
- Difficiles à dépersonnaliser; plus souvent sous forme textuelle ou audio;
- Recueillies par des êtres humains;
- Dépendent du contexte;
- Souvent recueillies auprès de communautés marginalisées ou d'individus vulnérables;
- Souvent issues de questions de recherche hautement sensibles;
- Sont moins souvent archivées, partagées ou réutilisées.

**Réponse à la question 2 :**

- Histoires orales;
- Journaux personnels de participantes ou participants;
- Photos;
- Vidéos;
- Documents;
- Artéfacts;
- Réponses ouvertes aux questions de sondage.

**Réponse à la question 3 :**

Faire le suivi des activités et de la prise de décisions pendant toute la durée du projet, précisant ce qui s'est passé, à quel moment et pourquoi.

**Réponse à la question 4 :**

- Capture;
- Traitement;
- Sécurité et sauvegarde des données;
- Transfert pour la transcription;
- Transfert vers d'autres membres de l'équipe;
- Traduction.

**Réponse à la question 5 :**

- Enregistrement original;
- Transcription originale;
- Transcription vérifiée;
- Transcription anonymisée;
- Transcription modifiée;
- Transcription codée.

**Réponse à la question 6 :**

La coproduction vise à réunir les compétences complémentaires des chercheuses et chercheurs en recherche qualitative et des bibliothécaires/archivistes/spécialistes des données afin d'établir et de mettre de l'avant des normes de gestion de données qualitatives.

## Chapitre 17, Gestion des données de recherche et mouvement de la science ouverte : positions et enjeux

**Réponse à la question 1 :**

Tout d'abord, les deux définitions caractérisent toutes deux la science ouverte par l'importance de la collaboration dans la conduite de la recherche. La libre mise à disposition des résultats de la recherche est aussi une caractéristique commune bien que la définition de Foster Open Science mette beaucoup plus l'accent sur le libre accès traditionnel et est en ce sens plus réductrice. La définition de Vicente-Saez et Martinez-Fuente parle davantage d'accès et de partage que de libre accès comme tel, le partage pouvant être soumis à des restrictions légales ou éthique. Cette définition est donc davantage formellement circonscrite par les principes FAIR que celle de Foster Open Science. Les principes de transparence sont également moins développés dans la définition de Foster Open Science. On y mentionne seulement des conditions favorisant la réutilisation sans être explicites sur ces conditions et on ne fait aucune allusion à l'assurance qualité et l'audit qui sont facilités par des principes de transparence.

**Réponse à la question 2 :**

La réponse peut varier selon les point de vue et les expertises qui ouvriraient sur de nombreux exemples non listés dans ce chapitre. Par contre, l'école pragmatique façonne plus d'un domaine de la science ouverte (tableau 1.2) : ouverture des protocoles de recherche, réseaux sociaux académiques et autres plateformes de

collaboration, comme les cahiers de laboratoires ouvert, et finalement ouverture du processus de révision par les pairs.

**Réponse à la question 3 :**

Faux. Les éditeurs commerciaux ont consolidé leur place en rendant dominant le modèle d'affaires du libre accès basé sur les frais de traitement d'articles et on ne peut pas exclure que l'acquisition d'infrastructure en lien avec les données de la recherche ne soit pas dans leur ligne de mire. Elsevier offre déjà son répertoire de données, Mendeley Data (https://www.elsevier.com/authors/tools-and-resources/research-data/mendeley-data-for-journals). Le blogue *Scholarly Kitchen* aborde assez régulièrement les acquisitions et fusions dans le domaine de l'édition dans le domaine de la santé. Consultez cet exemple : *Elsevier to Acquire Interfolio (https://scholarlykitchen.sspnet.org/2022/04/25/elsevier-acquire-interfolio/)*.

**Réponse à la question 4 :**

Parce que les données de recherche produites sont souvent dépendantes de leur contexte de production. Il devient alors problématique de penser la reproductibilité au regard de contextes qui sont souvent uniques. La reproductibilité de la recherche qualitative doit donc être envisagée à la lumière de diverses postures épistémologiques appelant elles-mêmes leurs propres méthodologies et balises d'analyse.

**Réponse à la question 5 :**

Le domaine des *Critical Data Studies.*