Découverte des données de recherche et écosystème du savoir au Canada

Livre blanc

Préparation du Groupe d'experts sur la découverte du réseau Portage au nom de l'Association des bibliothèques de recherche du Canada (ABRC) :

Eugene Barsky (University of British Columbia)
John Brosz (University of Calgary)
Amber Leahey (Scholars Portal, Conseil des bibliothèques universitaires de l'Ontario)

JUILLET 2016

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

www.carl-abrc.ca

portage
SERVICES PARTAGÉS POUR LES DONNÉES DE RECHERCHE
SHARED STEWARDSHIP OF RESEARCH DATA
CARL ABRC
CANADIAN ASSOCIATION OF RESEARCH LIBRARIES
ASSOCIATION DES BIBLIOTHÈQUES DE RECHERCHE DU CANADA

Table des matières

Sommaire

Il faut que les données de recherche soient repérables pour être réutilisées. La découverte des données correspond au traitement descriptif et technique des données et des métadonnées ainsi qu'aux outils et à l'infrastructure visant à améliorer l'accès aux données de recherche sur Internet et leur réutilisation. Un service canadien de découverte des données faciliterait la découverte et la réutilisation des données de recherche conservées dans des dépôts institutionnels et disciplinaires. Nous aimerions voir un service qui offrirait un point d'accès cohérent à des descriptions (métadonnées) concernant les ensembles de données qui soient fiables, interrogeables, explorables et exploitables par ordinateur, et qui offrirait des moyens clairs pour y accéder, ce qui accroîtrait les chances de découverte et de réutilisation des données de recherche au Canada.

Dans le document qui suit, nous faisons ressortir les possibilités et problématiques actuelles concernant l'élaboration d'un tel service au Canada. À l'aide d'une enquête concernant les dépôts de données de recherche et les services de découverte des données nationaux et internationaux, nous proposons un ensemble de principes directeurs, de pratiques exemplaires et de recommandations pour la découverte des données :

Métadonnées communes : Les renseignements descriptifs qui accompagnent les données de recherche doivent respecter des normes minimales pour favoriser la découverte et soutenir la réutilisation des données. Cette exigence requiert un engagement pour un ensemble fondamental de composants de métadonnées dans tous les domaines. Les outils de métadonnées devraient permettre de multiples espaces de nommage de métadonnées – termes descriptifs assignés, gérés et regroupés en collections de catégories et attributs. Nous recommandons également de créer de multiples outils de moissonnage de données souples pour l'indexation des dépôts spécialisés afin que l'on puisse conserver la granularité et les métadonnées propres à un domaine dans leur format original.

Identification permanente : L'utilisation d'identifiants universels pour les chercheurs et les données de recherche. Nous recommandons d'envisager une entente nationale ORCID afin que les universités et les organismes gouvernementaux du Canada puissent intégrer des identifiants de chercheur dans des logiciels de gestion et de publication de recherche institutionnels ou d'autre nature. Nous recommandons également d'enregistrer des identifiants numériques d'objet (DOI) se rattachant aux ensembles de données dans les dépôts participants auprès de DataCite Canada. Ces DOI amélioreront grandement la découverte des ensembles de données par l'entremise des partenaires de métadonnées de DataCite (p. ex., ORCID, VIVO, etc.).

Libre accès et API : Une interface de programmation d'application (API) qui permet à un composant logiciel de mettre la fonctionnalité ou les données à la disposition d'un autre par l'entremise d'un ensemble de routines, de protocoles et d'outils. Les dépôts et les services de découverte des données participants devraient offrir une interface de programmation permettant à des programmes d'accéder aux données et aux métadonnées à des fins de réutilisation et de développement.

Licences communes : Des politiques et des licences devraient régir l'accès aux données et aux métadonnées et, dans la mesure du possible, celles-ci devraient être le moins restrictives possible. Nous recommandons d'utiliser des licences Creative Commons pour les données de recherche puisqu'elles transmettent efficacement des renseignements sur les intentions des titulaires du droit d'auteur et qu'elles clarifient les usages autorisés. Les licences peuvent s'appliquer aux données et aux métadonnées, même si nous recommandons fortement d'offrir les métadonnées le plus librement possible, avec des restrictions minimales ou nulles sur leur réutilisation pour en faciliter la découverte.

Collaboration : Un engagement partagé pour la reconnaissance et la collaboration entre les acteurs, les organisations, les producteurs de données et les chercheurs, que l'on peut aussi qualifier de « coexistence dans l'écosystème du savoir ». Nous insistons sur le fait que la collaboration constituera le facteur déterminant de l'amélioration de la découverte des données au Canada. Un projet national bien coordonné permettra de s'assurer que toutes les initiatives pour améliorer la découverte et la consultation des données seront éclairées et facilitées par les attentes, la participation et la collaboration des intervenants. La mobilisation des intervenants et l'établissement de canaux de communication clairs sont essentiels à la réussite d'un service national de découverte des données.

Le document qui suit est présenté avec un objectif commun, soit celui de favoriser le plus possible la découverte et la consultation des données de recherche, et d'améliorer conséquemment les possibilités de reproductibilité et de réutilisation des données. L'amélioration de la découverte des données constitue une manière de faciliter l'interopérabilité et la découverte accrues des résultats savants. L'élaboration d'une infrastructure nationale qui rehausse la découverte des données de recherche améliorera les possibilités d'intégration additionnelle au cœur de l'écosystème du savoir, ce qui comprend le soutien pour les métadonnées, les identifiants universels, et les interfaces programmatiques à libre accès.

Introduction

Il faut que les données de recherche soient repérables pour être réutilisées. Les pressions et les intérêts croissants pour une plus grande disponibilité des données ont renforcé la nécessité d'offrir de moyens nouveaux et améliorés pour la découverte de données de recherche existantes. On a toujours considéré la capacité de vérification des conclusions de recherche comme un principe fondamental d'une pratique scientifique rigoureuse, mais on entend actuellement des demandes pressantes d'ouverture dans le domaine de la recherche afin d'améliorer la communication, le partage et la réutilisation des données par les chercheurs; trois organismes canadiens ont d'ailleurs fait une telle demande dernièrement[1].

Dans le présent document, la découverte des données correspond au traitement descriptif et technique des données et des métadonnées ainsi qu'aux outils et à l'infrastructure visant à améliorer l'accès aux données de recherche sur Internet et leur réutilisation. Le document n'a pas pour but de présenter un historique ou un examen exhaustif. Il vise plutôt à faire une synthèse des différents dépôts de données de recherche et services de découverte de données existants et à proposer un ensemble de principes directeurs et de pratiques exemplaires en matière de découverte des données. Il est à souhaiter que cet examen fasse ressortir les possibilités et problématiques actuelles concernant l'élaboration d'un tel service au Canada. Les dépôts et les services de données en développement peuvent tirer avantage de cet ensemble de principes et de recommandations pour améliorer la découverte des données.

Au Canada, un service national de découverte des données permettrait à d'autres d'explorer les données de recherche contenues dans des dépôts institutionnels et disciplinaires en regroupant des métadonnées sur les collections et les ensembles de données. Nous aimerions voir un service qui ne sera pas utilisé comme méga-dépôt pour les ensembles de données proprement dits. Il accroîtrait plutôt les chances de découverte et de réutilisation des données de recherche au Canada en offrant un point d'accès cohérent à des descriptions (métadonnées) concernant les ensembles de données qui soient fiables, interrogeables, explorables et exploitables par ordinateur, et il expliquerait comment y accéder.

[1] Gouvernement du Canada. *Déclaration de principes des trois organismes sur la gestion des données numériques*, 2016. Consultation le 15 juin 2016 au http://www.science.gc.ca/default.asp?lang=Fr&n=547652FB-1

Le respect des principes de la découverte des données suppose une approche holistique visant à rassembler des sources de données disparates qui, en raison de l'effet des silos de données, sont souvent éparpillées dans une multitude de dépôts institutionnels et disciplinaires, d'organisation et de collections de données. On considère, par exemple, que les mauvaises pratiques en matière de citation de données dans les résultats de recherche publiés (p. ex., articles de revue, rapports) et l'accès aux données de recherche sous-jacentes freinent considérablement la vérification scientifique et la reproduction dans diverses disciplines[2] [3]. Les appels à l'amélioration de l'identification des données, au partage des données et à l'établissement de liens entre les données de recherche et les publications mettent l'accent sur une amélioration de la gestion des données de recherche dans l'écosystème de la recherche savante.

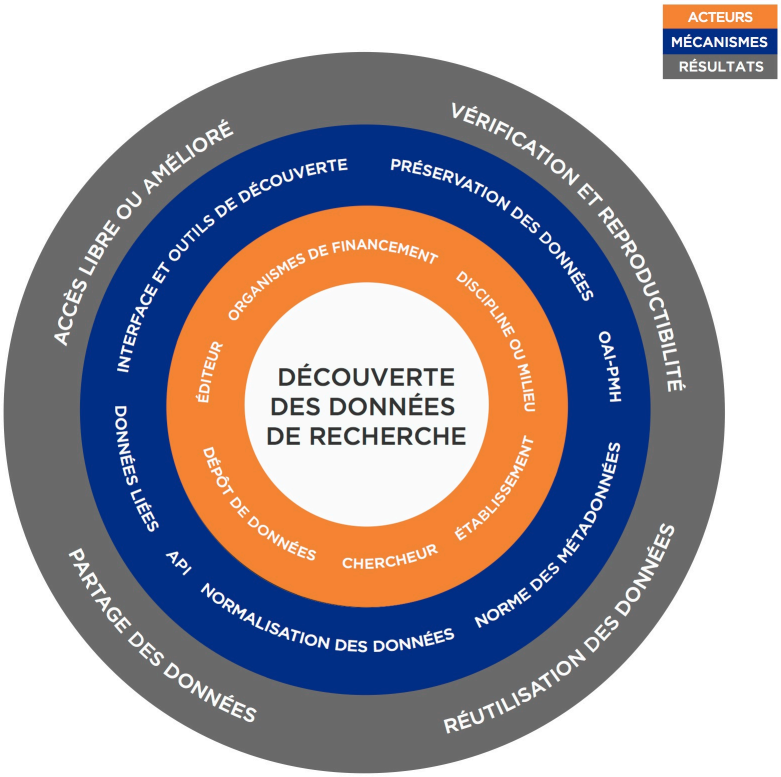

Figure 1 - Données de recherche et écosystème de la recherche savante

[2] G. King, « Replication, replication », *PS: Political Science & Politics, vol. 28*, n°3, 1995, p. 444 à 452.

[3] E. Yong, « Replication studies: Bad copy », *Nature, vol. 485*, n° 7398, 2012, p. 298 à 300. DOI : 10.1038/485298a

Principes de la découverte des données

Une évaluation des plates-formes et des normes utilisées par les dépôts de données et les services de découverte des données permet de dégager plusieurs thèmes. La découverte des données est le plus souvent liée à l'ouverture des données, y compris à un engagement pour l'ouverture des métadonnées, dans le but de favoriser la découverte et la réutilisation des données sur Internet. Ceci est possible seulement s'il y a interaction entre les acteurs et par le déploiement de différents mécanismes (voir la Figure 1). Voici un ensemble de principes qui, à notre avis, englobe des pratiques exemplaires en matière de découverte des données.

Ensemble de Principes:

1. Métadonnées communes
2. Identification permanente
3. Accès libre
4. Licences communes
5. Collaboration (coexistence dans l'écosystème du savoir)

Les métadonnées communes font référence à l'ensemble des renseignements descriptifs qui accompagnent des données de recherche. Les métadonnées doivent respecter des normes minimales pour permettre le repérage efficace et soutenir la réutilisation des données de recherche dans diverses disciplines. Cela n'exige pas nécessairement l'usage des mêmes normes dans tous les domaines, mais plutôt un engagement à utiliser un petit ensemble de base commun de métadonnées favorisant la découverte et la réutilisation.

L'**identification permanente** des données englobe l'usage d'identifiants universels pour les chercheurs et les données de recherche dans le but de permettre la publication des données dans l'écosystème de la recherche savante (p. ex., DOI, ISNI, ORCID).

Le **libre accès** aux données et aux métadonnées s'inspire d'autres principes d'ouverture (P. ex., OCDE [4]), comme un accès au coût le plus bas possible et un accès simple, opportun et convivial par l'entremise d'Internet de préférence. Partant de ce principe, le libre accès dans les outils de découverte devrait être conçu pour permettre l'exploitation tant par des personnes que par des machines (p. ex., accès au moyen d'API).

[4] Recommandation du Conseil de l'OCDE sur l'accès aux données de la recherche financée par des fonds publics (2006), consultation le 9 juin 2016 au https://www.oecd.org/fr/sti/sci-tech/38500813.pdf

Les licences communes englobent différentes conditions relatives au partage, à la consultation et à l'utilisation des données et des métadonnées. Dans la mesure du possible, il faudrait établir des politiques d'octroi de licences pour les données et les métadonnées qui soient le moins restrictives possible. L'élaboration de politiques nationales et internationales sur le partage et la réutilisation des données appuiera grandement les efforts déployés en vue de définir des licences communes et partagées pour les données provenant de diverses disciplines. Par exemple, toute personne qui souhaite fournir des données librement et sans restriction pourrait utiliser la licence CC0 de Creative Commons pour les données et les métadonnées.

La collaboration, ou plutôt la coexistence dans l'écosystème du savoir, demande un engagement pour la reconnaissance mutuelle et la collaboration entre les acteurs, les organisations, les producteurs de données, les chercheurs et autres intervenants concernés.

Dans l'ensemble, le présent document et la présentation des principes de découverte des données ont pour but de réunir autour d'un objectif commun les grandes organisations et les principaux acteurs du milieu savant. Leurs fondements sont de favoriser le plus possible la découverte et la consultation des données de recherche, d'améliorer les possibilités de reproductibilité et de réutilisation des données et de favoriser la découverte des nouvelles connaissances au Canada. L'amélioration de la découverte des données constitue une manière de faciliter l'interopérabilité et la découverte accrues des résultats savants, ce qui comprend les données de recherche, dans l'ensemble de l'écosystème du savoir.

Création d'un service national de découverte des données au Canada

La recherche multidisciplinaire a besoin d'un accès à des données et renseignements numériques provenant d'une multitude de sources, de communautés et de dépôts de données. La demande pour un service regroupé de données et de métadonnées exige une infrastructure et des outils de métadonnées efficaces pour permettre des recherches interdisciplinaires. Parmi les exemples de services qui récupèrent des données de diverses sources et de catalogues collectifs, on trouve le service OAIster de l'OCLC[5], le portail Europeana[6], la Digital Public Library of America (DPLA)[7], et

[5] OAIster de l'OCLC, consultation le 9 juin 2016 au http://www.oclc.org/fr-CA/oaister.html

[6] Portail Europeana, consultation le 9 juin 2016 au http://www.europeana.eu/portal/

[7] DPLA, consultation le 9 juin 2016 au https://dp.la/

Blacklight,[8] une application de catalogue public en ligne, pour n'en nommer que quelques-uns.

Voici une liste de services de découverte de données nationaux qui regroupent des données de diverses sources en vue d'accroître la repérabilité et la consultation des données de recherche. Il n'existe actuellement aucun point d'accès central pour les données de recherche et les ressources connexes au Canada. Il s'avère donc utile d'évaluer des méthodes adoptées ailleurs avant d'élaborer un tel service ici.

Tableau 1 - Comparaison des services de découverte nationaux regroupant des données [9]

Pays ou région	Nom du fournisseur	Nom du service	Modèle de service	Fournisseurs ou sources de données	Métadonnées ou normes	Modèle de coûts ou maintenance
États-Unis	Association of Research Libraries (ARL) et Center for Open Science (COS)	SHARE[10], SHARE Notify	Service de notification des activités de publication de recherche (publications, plans de gestion des données, présentations, données de recherche) avec des options offertes aux développeurs pour l'accès aux bases de données. Service de notification : Fil Atom-XML, outil de recherche et de navigation et interface de programmation JSON.	Plus de 100 fournisseurs (arXiv, CrossRef, PubMed Central et autres dépôts).	VIVO, ORCID, DataCite, OAI-PMH, Dublin Core	Programme des adjoints en organisation 2016-2017, financé à l'aide de subventions (Institute of Museum and Library Services [IMLS], et Alfred B. Sloan Foundation) et offert aux bibliothécaires professionnels en tant que formation et modèle de maintenance.

[8] Projet Blacklight, consultation le 9 juin 2016 au http://projectblacklight.org/

[9] Le tableau propose un ensemble structuré de renseignements pour comprendre les initiatives et les services variés qui sont en cours d'élaboration dans le contexte de cinq services nationaux ou régionaux différents. Il s'appuie sur diverses sources, principalement électroniques et imprimées, dont celle-ci : Data Service Infrastructure for the Social Sciences and Humanities (DASISH), 2012 *Report, Appendix: Data Archive Description Sheets (DADS)*, p. 169 à 179.

[10] SHARE, consultation le 9 juin 2016 au http://www.share-research.org/

Australie	Partenariat dirigé par la Monash University, de concert avec l'Australian National University et la Common-wealth Scientific and Industrial Research Organisation.	ANDS[11], Research Data Australia	ANDS possède un service centralisé, Research Data Australia. Il s'agit d'un service de découverte Web qui s'appuie sur des enregistrements de données provenant de plus de 90 établissements dans le but de rassembler et de présenter des données australiennes à l'échelle nationale et internationale. Research Data Australia couvre une vaste gamme de domaines de recherche, comme les sciences, les sciences sociales, les arts ainsi que les lettres et sciences humaines.	Appui sur des enregistrements de données provenant de plus de 90 établissements	Norme de métadonnées RIF-CS, moissonnage du Service de catalogue Web (CSW), OAI-PMH	Financement par le gouvernement australien
Royaume-Uni	Joint Information Systems Committee (JISC), Digital Curation Centre (DCC)	UK Research Data Discovery Service [12]	DCC et projet-pilote UK Data Service en vue de créer un service national de registres pour regrouper les métadonnées des données de recherche conservées dans des universités du Royaume-Uni et des centres de données disciplinaires nationaux.	Neuf établissements d'enseignement supérieur, sept centres de données, dont l'UK Data Archive, l'Archaeology Data Centre et les centres de données du NERC	Plate-forme CKAN, OAI-PMH, Core Metadata Schema Version 1.0[13] -(mappage de Dublin Core, MODS, DDI-Codebook, DataCite, UK Gemini, EPrints / ReCollect)	Financement en grande partie par l'ESRC du Royaume-Uni, le JISC et la University of Essex.

[11] ANDS Research Data, consultation le 9 juin 2016 au http://www.ands.org.au/

[12] Research Data Discovery Service du R.-U., consultation le 9 juin 2016 au http://ckan.data.alpha.jisc.ac.uk/fr/dataset

[13] Core Metadata Schema du Research Data Discovery Service du R.-U., consultation le 9 juin 2016 au https://docs.google.com/document/d/1pdSPfOTDPL8n6MiHDuqRF_zqESIQnb0gKOtVrkIpvBs/edit

Europe	OpenAIRE 2020 Project, Research Data Alliance Europe/WDS Publishing Data Interest Group et ICSU World Data System	Data Literature Inter-linking (DLI) Service[14]	Le DLI Service propose un service ouvert pour la collecte, le partage et la réutilisation de couplages de données entre des ressources publiées et des données de recherche sous-jacentes. Les liens et les métadonnées sont disponibles aux utilisateurs (repérage), aux développeurs et aux fournisseurs de contenu (amélioration des fonds).	Plus de 20 fournisseurs, dont CrossRef, PubMed, l'IEEE, l'ICPSR, l'ANDS, PANGAEA et DataCite	DOI de DataCite, OAI-PMH, infrastructure D-NET	Projet OpenAIRE financé par l'UE
Pays-Bas	DANS	DANS Search[15], NARCIS et EASY[16]	Les DANS proposent divers services, dont le National Academic Research and Collaborations Information System (NARCIS), le portail national principal de consultation des résultats de recherches savantes menées partout aux Pays-Bas.	DANS Search et NARCIS permettent d'accéder aux résultats savants de nombreux établissements. Téléversement des ensembles de données vers EASY par les chercheurs directement. Organisation par les archivistes des DANS.	Dublin Core, Dublin Core qualifié, attributs DDI avec ajouts optionnels de métadonnées du FGDC ou non normalisées.	KNAW et la Netherlands Organization for Scientific Research (NWO)

[14] DLI Service, consultation le 9 juin 2016 au http://dliservice.research-infrastructures.eu/#/

[15] Service DANS Search, consultation le 9 juin 2016 au http://www.dans.knaw.nl/en/search

[16] DANS. Service EASY, consultation le 9 juin 2016 au https://easy.dans.knaw.nl/ui/home

Les modèles de financement des projets nationaux de découverte européens et du service SHARE des États-Unis s'appuient tous sur des subventions nationales. Cela ne signifie pas qu'il est impossible d'établir un modèle différent au Canada, mais plutôt qu'il s'agit d'un élément de réflexion, surtout si l'on considère sa durabilité.

Au Canada, divers dépôts et centres de données institutionnels et disciplinaires hébergent des données de recherche à réutiliser. Le Tableau 2 présente une liste de dépôts de données institutionnels et disciplinaires qui, selon le Groupe de travail, s'inscriraient dans l'élaboration d'un service national de découverte des données au Canada. Pour connaître la liste complète des dépôts de données et des centres de données au Canada, veuillez consulter la Passerelle de données de recherche du Conseil national de recherches du Canada[17].

Tableau 2 - Liste de dépôts de données et de centres de données institutionnels canadiens pertinents [18]

Fournisseur ou institution	Nom du dépôt	Disciplines couvertes	Modèle du dépôt	Métadonnées ou normes	Taille[19]	Exploration des données
Bibliothèque de la University of British Columbia	Dépôt Abacus et Dataverse	Multidisciplinaire	Intendance des données, accès contrôlé au dépôt, multi-institutionnel	DDI-Codebook (mappage Dublin Core, DataCite, ISO 19115)	1 875 études, 30 555 fichiers	Accès libre, OAI-PMH, API publique, métadonnées communes (champs obligatoires), exploration au niveau des variables, descripteurs, DOI

[17] Passerelle de données de recherche du CNRC, consultation le 9 juin 2016 au https://dr-dn.cisti-icist.nrc-cnrc.gc.ca/eng/home/collection/Gateway%20to%20Research%20Data/

[18] La liste n'est pas exhaustive. On a jugé qu'il serait pertinent de consulter plus avant les dépôts figurant sur la liste en vue de l'élaboration d'une politique sur les collections dans le cadre d'un service national de découverte des données. Ce tableau s'appuie sur le document *Research Data Repositories: Review of current features, gap analysis, and recommendations for minimum requirements*, publié en 2015 par le Comité des normes et de l'interopérabilité (SINC) de Données de recherche Canada (DRC). Consultation le 9 juin 2016 au http://www.rdc-drc.ca/download/review-of-research-data-repositories-2015/?wpdmdl=669

[19] Articles dans le dépôt, s'il est connu, au moment de la rédaction (juin 2016).

Conseil des bibliothèques universitaires de l'Ontario (CBUO)	<odesi>	Sciences sociales, multidisciplinaire	Intendance des données, accès contrôlé au dépôt	DDI-Codebook (mappage Dublin Core, MARC)	3 535 ensembles de données	Certaines restrictions pour les données, accès libre aux métadonnées, OAI-PMH, API publique, métadonnées communes (pratique exemplaire), exploration au niveau des variables
Conseil des bibliothèques universitaires de l'Ontario (CBUO)	Dépôt Scholars Portal Dataverse	Multidisciplinaire	Dépôt autonome, multi-institutionnel	DDI-Codebook (mappage Dublin Core, DataCite, ISO 19115) Autres normes disciplinaires (astronomie, biomédecine, sciences de la Terre, revue)	473 études, 6 342 fichiers	Accès libre, OAI-PMH API publique, métadonnées communes (champs obligatoires), exploration au niveau des variables, descripteurs, DOI
Bibliothèques de la University of Alberta	Dépôt Dataverse de la University of Alberta	Multidisciplinaire	Intendance des données, accès contrôlé au dépôt	DDI-Codebook (mappage Dublin Core, DataCite, ISO 19115)	247 études, 2 407 fichiers	Accès libre, OAI-PMH API publique, métadonnées communes (champs obligatoires), exploration au niveau des variables, descripteurs, DOI
Initiative de démocratisation des données (IDD) de Statistique Canada	IDD	Multidisciplinaire	Intendance des données, accès contrôlé au dépôt	DDI-Codebook	Inconnue	Certaines restrictions pour les données, accès libre aux métadonnées, OAI-PMH, API publique, métadonnées communes (pratiques

						exemplaires), exploration au niveau des variables
Réseau canadien des Centres de données de recherche (RCCDR)	Dépôt des métadonnées des fichiers principaux des centres de données de recherche (réseau multi-institutionnel pour accéder aux données des CDR)	Sciences sociales, multidisciplinaire	Intendance des données, accès gouvernemental contrôlé seulement	Base de mégadonnées intégrée (BMDI) de Statistique Canada, DDI-Codebook, DDI-Lifecycle	114 enquêtes	Accès contrôlé aux données, accès libre aux métadonnées, OAI-PMH API publique, métadonnées communes (pratiques exemplaires), exploration au niveau des variables
Polar Data Catalogue (PDC)	Canadian Cryospheric Information Network, University of Waterloo	Sciences de la Terre, multidisciplinaire	Intendance des données, accès contrôlé au dépôt	Métadonnées du PDC, FGDC, ISO 19115	2 442 ensembles de données	OAI-PMH DOI
Nordicana D	Centre d'études nordiques, Université Laval	Multidisciplinaire	Intendance des données, accès contrôlé au dépôt	Inconnue	Inconnue	Ne s'applique pas en bonne partie, DOI pour les données et les publications
Simon Fraser University	RADAR[20]	Multidisciplinaire	Intendance des données, accès contrôlé au dépôt	DDI, Dublin Core	269 articles	Accès libre Formats libres
Oceans Network Canada	Oceans 2.0 Data Search[21]	Multidisciplinaire	Intendance des données, accès contrôlé au dépôt	Inconnue	Inconnue	Quality Assurance of Real Time Oceanographic Data (QARTOD)

[20] RADAR, consultation le 9 juin 2016 au http://researchdata.sfu.ca/

[21] Recherche dans Oceans 2.0, consultation le 9 juin 2016 au http://dmas.uvic.ca/DataSearch

Centre canadien de données astronomiques[22]	CCDA	Astronomie, physique	Intendance des données, accès contrôlé au dépôt	Inconnue	115 instruments	TAP[23], téléchargement direct
University of Calgary	**University of Calgary**	Biologie, sciences de la vie	Intendance des données, accès contrôlé au dépôt	Inconnue	Inconnue	Accès libre

Les dépôts de données de recherche mentionnés ci-dessus constituent de vastes collections de données de recherche qui sont utilisées pour soutenir les recherches dans un domaine précis, qui servent à une région ou à une communauté en particulier (dans le cas d'établissements ou de consortiums) ou qui font partie d'administrations gouvernementales ou d'autre nature, empêchant ainsi le développement de collections ou de dépôts externes ou interdisciplinaires. Signalons que, si les dépôts sont souvent responsables de l'intendance des données de recherche et qu'ils fournissent des outils pour leur partage et leur consultation, le besoin existe d'instaurer un service commun de découverte des données de recherche au Canada qui englobe les données dans toutes leurs formes, provenant des diverses disciplines, pour soutenir la consultation et la découverte des données à partir d'un point central.

Enjeux et considérations

Il faut prendre en considération plusieurs éléments si l'on veut regrouper des métadonnées provenant de différents dépôts et centres de données au Canada. Il faut notamment définir l'objectif poursuivi, la portée du projet et le public-cible. Les données canadiennes ne se trouvent pas toutes dans des dépôts canadiens. Par exemple, on trouve des données recueillies par des chercheurs canadiens ou au sujet du Canada dans une multitude d'endroits, y compris des dépôts internationaux comme Dryad[24], FigShare[25] et PANGAEA[26]. Se concentrer uniquement sur les dépôts et les centres de données canadiens serait limitatif; d'autres considérations et enjeux sont à examiner.

---

[22] CCDA, consultation le 9 juin 2016 au http://www.cadc-ccda.hia-iha.nrc-cnrc.gc.ca/fr/

[23] TAP, consultation le 9 juin 2016 au http://www.ivoa.net/documents/TAP/20100327/REC-TAP-1.0.html

[24] Dryad, consultation le 9 juin 2016 au http://datadryad.org/

[25] FigShare, consultation le 9 juin 2016 au https://figshare.com/

[26] Pangaea, consultation le 9 juin 2016 au https://www.pangaea.de/

Il faut collaborer davantage avec les partenaires qui gèrent les dépôts institutionnels et disciplinaires au Canada pour refléter la gamme complète des cas d'utilisation et des enjeux pour bâtir un tel service national. Il convient de noter que, parmi tous les cas de services nationaux de découverte des données présentés dans le Tableau 1, la consultation avec la communauté et les intervenants des dépôts est primordiale.

Voici une liste des enjeux, présentée sans ordre d'importance, qui nécessiteront la tenue de consultations auprès des communautés :

- le dédoublement des données;
- la granularité des ensembles de données, surtout pour les grandes quantités de données (p. ex., données génomiques);
- les plates-formes et les outils;
- la volonté de participer aux métadonnées, de les partager et d'y contribuer[27];
- les normes de métadonnées;
- le niveau et les méthodes de regroupement;
- le modèle de maintenance et de durabilité.

La portée multidisciplinaire d'un service national de découverte pose des défis uniques pour l'infrastructure, car la gestion de la granularité des ensembles de données peut s'avérer difficile pour les groupes d'utilisateurs, les disciplines et les intervenants. Il est à noter qu'il peut s'avérer irréaliste de trouver une seule et même approche pour toutes les données au Canada. Nous pouvons néanmoins commencer à examiner ces enjeux et élaborer des solutions pour répondre aux exigences disciplinaires à mesure qu'elles se manifestent.

Il est aussi important de discuter des plates-formes et des outils, puisqu'ils ont un effet direct sur la capacité d'adhérer à des principes en matière de découverte des données. Par exemple, si un service de découverte des données utilise des plates-formes et des outils qui ne permettent pas le libre accès, il sera difficile de respecter nos principes et nos normes. Le Groupe de travail n'évalue pas les plates-formes ou les outils de dépôts, mais de telles évaluations pourraient être incluses dans le mandat de comités consultatifs ayant pour but d'aider à l'élaboration de tels services.

[27] Le Groupe de travail n'a pas mobilisé activement les dépôts ou établissements décrits dans le tableau ci-dessus (au-delà de nos propres affiliations). Il faut donc les consulter pour sonder leur volonté à participer à un service national et à y contribuer.

Recommandations

Nous recommandons de lancer un projet visant à définir la portée d'un éventuel service national de découverte des données, et de le faire de préférence en mettant sur pied des groupes consultatifs pour les dépôts et les centres de données que nous souhaitons consulter. Ces groupes, ou un plus grand groupe consultatif combiné, prendraient part à des consultations tout au long des étapes du projet, ce qui comprend la définition de la portée des collections, les questions liées aux métadonnées et aux outils, la mobilisation des intervenants et des dépôts, le développement technique et la mise à l'essai. Plutôt que de réinventer la roue, les groupes consultatifs évalueraient les services nationaux de découverte des données existants (R.-U., ANDS, DANS, etc.) et s'appuieraient sur ces efforts pour dégager une certaine compréhension commune.

Représentation recommandée au sein des groupes consultatifs :

- les utilisateurs (chercheurs, corps professoral, bibliothécaires, etc.);
- les dépôts ou les centres de données (un groupe choisi pour la mise à l'essai d'un service);
- les établissements (établissements d'enseignement de toutes les régions du Canada).

Outils de découverte et de visualisation des données améliorés

Un outil de découverte à la fine pointe de la technologie n'offre pas seulement une liste de ressources avec une boîte de recherche textuelle et des mots-clés. Les dernières technologies de recherche aident également les utilisateurs à formuler et à préciser leurs requêtes.

Voici un ensemble de fonctions jugées particulièrement utiles pour la découverte des données :

- des options de recherche avancée (p. ex., plage de dates, recherche au niveau des variables ou d'autres éléments des données, couverture géographique, etc.);
- la visualisation des métadonnées normalisées (affichage des champs, des valeurs, des liens ou d'autres éléments pour la compréhension);
- le facettage (par type, format, collection, dépôt, couverture géographique, sujet, date, etc.);
- des vues d'ensemble (niveau des études scientifiques) avec la possibilité de forer jusqu'aux éléments de données;

- des représentations liées qui relient les données et les ressources connexes, dans le but de contextualiser les données et les recherches s'y rapportant;
- des API ouvertes qui permettent de réutiliser des métadonnées et des résultats de recherche pour développer des applications.

Le Tableau 3 (ci-dessous) comprend des systèmes et techniques offrant des possibilités de découverte améliorée des données. La plupart de ces systèmes n'ont pas pour objet principal la découverte des données de recherche, mais le traitement qu'ils font des métadonnées bibliographiques pour les documents de recherche savante plus traditionnels est directement applicable à des collections d'ensembles de données.

Tableau 3 – Outils de découverte améliorés

Projet	Description	Données[28]	Béta[29]	API	Lien
Bohemian Bookshelf	Découverte au hasard par l'entremise de visualisations liées ou modulaires	N	O	N	http://www.alicethudt.de/BohemianBookshelf
Collection Diver	Interface de recherche qui met en évidence le processus de recherche et qui présente des commentaires visuels clairs sur les filtres et les facettes.	N	O	N	http://hci.uni-konstanz.de/downloads/CollectionDiver.mp4
PivotPaths	Exploration par facettes de réseaux de citations bibliographiques	N	O	N	http://mariandoerk.de/pivotpaths
PivotSlice	Construction visuelle de demandes de données dynamiques intégrant des sources de données en ligne, une recherche en direct et des historiques d'interactions graphiques.	N	O	N	http://vialab.science.uoit.ca/portfolio/pivotslice

[28] Cette marque indique que le projet a été conçu spécialement pour l'exploration des ensembles de données.

[29] Cette marque indique si le projet est un projet de recherche ou un prototype (O) ou un système de production (N).

VisGets	Éléments visuels interactifs coordonnés servant à créer des requêtes qui offrent des résumés graphiques et des formulations de requête.	N	O	N	http://innovis.cpsc.ucalgary.ca/Research/VisGets
Aperçu visuel de données gouverne-mentales	Utilisation de tableaux de bord visuels pour prévisualiser des ensembles de données afin d'évaluer rapidement la qualité et l'applicabilité de ces ensembles de données à des fins précises.	O	O	N[30]	http://dl.acm.org/citation.cfm?id=2757407 https://github.com/niclabs/visual-overview
Ariadne's Thread	Graphiques en réseau d'entités bibliographiques pour favoriser l'exploration du contexte Interopérabilité des prototypes avec ArticleFirst, WorldCat et Astrophisics.	N	O	N	http://www.oclc.org/research/themes/data-science/ariadne.html
Scopus	Ce système propose des options d'analyse des résultats de recherche de la base de données sur les résumés et les citations qui permettent d'interagir au moyen de graphiques pour des sélections fondées sur l'année, la source, l'auteur, l'affiliation, le pays, le type de document et le domaine.	N	N	O	https://www.elsevier.com/solutions/scopus
Open Collections de la University of British Columbia	Open Collections rassemble du contenu créé et géré localement dans les quatre dépôts à accès libre de la bibliothèque de la University of British Columbia (DSpace, CONTENTdm, Dataverse et AtoM).	O	N	O	https://open.library.ubc.ca

[30] Le code source de démonstration est disponible sur GitHub et il est soumis à une licence Apache 2.0.

Outre les outils conçus pour la découverte des ensembles de données, plusieurs dépôts de données fournissent leurs propres outils ou utilisent des outils externes pour intégrer des fonctions d'analyse et de visualisation des données, ce qui permet aux utilisateurs d'explorer un ensemble de données sur place pendant que l'analyse se déroule sur le serveur du dépôt. Cette méthode améliore la découverte et la consultation des données de recherche en permettant au chercheur d'explorer des éléments de données sur Internet sans logiciel spécialisé pour la lecture des données. En plus de ces types de systèmes de visualisation, le téléchargement des données offre aux utilisateurs la souplesse nécessaire pour explorer et évaluer l'applicabilité des données en vue d'une réutilisation.

Tableau 4 - Dépôts et systèmes de visualisation des données

Outil	Dépôt	Description	Lien
CKAN	CKAN	Plusieurs outils de navigateur Web pour afficher des tableaux, créer des graphiques et faire du mappage.	http://ckan.org/
R	Dataverse	Intégration avec Dataverse pour offrir des fonctions d'analyse statistique avancées.	https://www.r-project.org/
Two Ravens	Dataverse	Fonctions d'analyse statistique avancées pour les données quantitatives et utilisation de graphiques interactifs pour filtrer et afficher les résultats.	http://datascience.iq.harvard.edu/about-tworavens
Chemistry Solution Pack	Islandora	Ajout de fonctions propres à la chimie, comme l'affichage tridimensionnel de molécules et l'analyse checkmol (analyse de fichiers de structure moléculaire pour déceler la présence de groupes fonctionnels et d'éléments structuraux).	https://github.com/discoverygarden/islandora_solution_pack_chemistry
Islandora Data Solution Pack	Islandora	Affichage de données sous forme de tableau (feuilles de calcul) dans un navigateur.	https://github.com/axfelix/islandora_solution_pack_data

Geoserver	Nesstar	Fonctions de cartographie géospatiale.	http://geoserver.org/
Nesstar	Nesstar	Création de graphiques et de sous-tableaux Les fonctions statistiques comprennent des tableaux croisés, des corrélations, des régressions et l'application de pondérations variables.	http://www.nesstar.com/

Enjeux et considérations

Dans le contexte de la découverte des données, le volume de document dans de tels systèmes pose des défis. Par exemple, les recherches par liste ou sur l'ensemble du texte peuvent être difficiles à utiliser. Les systèmes qui sont multidisciplinaires de façon inhérente exigent diverses techniques pour aider les utilisateurs à composer avec les différences présentes dans les vocabulaires, l'usage des termes et les représentations d'un ensemble de données. Par conséquent, le processus de visualisation des systèmes de découverte ne se limite pas à la création d'un simple graphique ou plan avec un aperçu des ensembles de données. Il comprend plutôt des éléments visuels et interactifs pouvant servir à accéder à des ressources spécifiques, à exclure des éléments, à parcourir des collections et à aider les utilisateurs à formuler des requêtes de recherche dans la collection.

En ce qui concerne les systèmes améliorés de découverte et de visualisation des données, il existe deux volets importants :

- il faut démonter les silos de données ainsi que favoriser le couplage et la réutilisation des données et des collections connexes, particulièrement pour la recherche interdisciplinaire;
- il faut faciliter le couplage des données avec d'autres résultats de recherche, permettre la citation de données (à l'aide d'identifiants permanents), simplifier l'utilisation des références et, conséquemment, intégrer des données dans l'évaluation des résultats de la recherche et de l'impact en général.

En outre, pour qu'un tel service s'avère réellement utile pour le milieu canadien de l'éducation et de la recherche, il est crucial de confier aux utilisateurs du milieu de la recherche un rôle central dans la définition des exigences et la formulation de commentaires sur tous les aspects du développement de ce service.

Principes de la découverte des données et autres recommandations

Métadonnées communes

Il faut des normes minimales reconnues sur les métadonnées de gestion des données de recherche pour permettre une exploration adéquate et soutenir l'administration et la gestion de la recherche tout au long du cycle de vie des données de recherche. Le projet UK National Data Discovery[31] a récemment comparé différents schémas de métadonnées provenant de dépôts de données de recherche institutionnels pour en dégager les éléments communs. D'autres systèmes de découverte des données adoptent des approches différentes concernant la question des métadonnées communes ou obligatoires et l'accommodement de diverses normes de métadonnées provenant de producteurs et de fournisseurs de métadonnées distincts.

Comme le montre le Tableau 1, ANDS utilise le schéma RIF-CS, lequel comprend de nombreux éléments obligatoires. Il en est ainsi parce qu'ANDS tente de trouver une solution nationale couvrant l'ensemble du cycle allant de la création à la préservation des données de recherche. Cela comprend des normes concernant la découverte, les valeurs, l'accès et la réutilisation qui exigent l'inclusion de métadonnées administratives et disciplinaires dans chaque enregistrement.

En revanche, DataCite ne compte que cinq champs de métadonnées obligatoires. DataCite[32] est une norme importante qui tient particulièrement compte de la découverte et du couplage sur Internet. Il convient de préciser que certaines disciplines possèdent leurs propres normes de métadonnées et ontologies. Il reste à voir si les solutions techniques locales peuvent accommoder ces éléments issus de diverses disciplines sans rendre trop complexes les flux des dépôts.

Remarque sur la granularité

En outre, de nombreux dépôts de données au Canada acceptent les descriptions de métadonnées au niveau granulaire des variables, comme la norme DDI (Initiative de documentation des données). Abacus de la University of British Colombia, <odesi> du CBUO et le dépôt de l'Initiative de démocratisation des données (IDD) de Statistique Canada offrent l'encodage DDI d'ensembles de données pour permettre la description au niveau des variables et la réutilisation. Il est donc important de signaler que la granularité doit être prise en compte et demeurer souple pour permettre ces descriptions favorisant une découverte riche des données.

[31] Blogue de JISC, consultation le 9 juin 2016 au https://rdds.jiscinvolve.org/wp/2016/03/18/how-much-metadata-is-enough/

[32] Norme de métadonnées DataCite, consultation le 9 juin 2016 au https://www.datacite.org/

Cela ne signifie toutefois pas que les métadonnées descriptives au niveau des variables soient superflues. Il faut intégrer la granularité selon le contexte, mais il ne faut pas le faire au détriment de la convivialité. Il en va de même pour d'autres disciplines pour lesquelles les différences entre les éléments distinctifs d'un ensemble de données peuvent devenir écrasantes, surtout avec de grands volumes de données (p. ex., données génomiques, cristaux, etc.).

Remarque sur les données et les systèmes multilingues
Dans le contexte canadien, il faut offrir aux utilisateurs du contenu et une interface de recherche dans les deux langues officielles, soit le français et l'anglais. Il y a également le cas des autres langues. Pour favoriser une exploration efficace, les systèmes sous-jacents de l'affichage et de la visualisation des métadonnées et des données devraient pouvoir traiter de nombreuses langues.

Par exemple, les plates-formes d'affichage des métadonnées devraient offrir aux utilisateurs une expérience identique ou très similaire dans les deux langues. Il faut donc examiner minutieusement la configuration du système, les règles de définition des éléments indexables et interrogeables ainsi que l'expérience globale relative à la recherche, aux interactions avec l'interface d'utilisation et aux affichages dynamiques.

Plusieurs projets d'internationalisation de dépôts de données sont en cours au Canada dans le but d'offrir des logiciels et des outils multilingues pour la gestion des données de recherche, dont un partenariat entre la University of Alberta, l'Université de Montréal et le Scholars Portal (CBUO).

Tous ces projets ont pour but d'aider les chercheurs à trouver et à utiliser des ensembles de données sans égard à la langue d'origine de la source de recherche. Ce travail nécessite donc de résoudre les problèmes relatifs aux normes de métadonnées et aux capacités langagières au niveau des champs, aux vocabulaires contrôlés, à l'encodage des caractères des ensembles de données, aux délimiteurs de données et à d'autres éléments. CERIF[33] d'euroCRIS[34], un modèle de données relationnelles international pour les renseignements administratifs sur les recherches, pourrait éclairer l'intégration d'ensembles de données provenant de dépôts et d'établissements francophones et bilingues.

[33] CERIF, consultation le 9 juin 2016 au http://www.eurocris.org/cerif-cornerstone-creation-research-information-infrastructures

[34] Association eutoCRIS, consultation le 9 juin 2016 au http://www.eurocris.org/what-eurocris

Recommandations

Les organisations qui mettent en œuvre une norme de métadonnées communes connaissent une multitude de complications, comme Open Collections de la University of British Colombia, l'UK National Research Data Discovery Service et d'autres. Nous recommandons l'analyse d'un ensemble d'outils de métadonnées et d'une plate-forme qui acceptent des espaces de nommage de métadonnées multiples et en chevauchement tout en prenant pour acquis qu'un schéma générique existant, comme Dublin Core[35], MODS[36] ou METS[37], puisse être utilisé pour tous les objets, à l'aide de correspondances entre les schémas s'il le faut. Nous recommandons également des outils de moissonnage de données souples indexer des dépôts spécialisés, au besoin, afin que l'on puisse conserver la granularité et les métadonnées propres à un domaine dans leur format original pour la découverte.

Il faut aussi examiner attentivement les problèmes relatifs aux données et aux contenus multilingues ainsi qu'à l'uniformité de l'expérience d'utilisation, peu importe la langue employée.

Identification permanente

Un identifiant universel unique, comme un DOI ou un ORCID, propose un mécanisme distinct et stable pour identifier des objets et des personnes sur Internet. Cela signifie qu'il ne change pas même si l'on déplace ou renomme l'élément ou l'objet en question.

DataCite, une organisation internationale formée et soutenue par divers acteurs du milieu savant, comme des organismes gouvernementaux, des éditeurs de revues, des producteurs de données, des établissements et des bibliothèques, fait la promotion de l'identification des données de recherche au moyen de la norme DOI (identifiant numérique d'objet), laquelle soutient l'identification des données et des publications sur Internet. Un certain nombre de dépôts de données, de bibliothèques et de centres de données au Canada souhaitent fournir des DOI pour les ensembles de données. Ils espèrent que cette approche mènera tôt ou tard à l'adoption de normes partagées pour l'identification d'ensembles de données sur Internet. Toutefois, l'identification ne constitue qu'un seul aspect du défi consistant à améliorer la découverte des données.

---

[35] Norme de métadonnées Dublin Core, consultation le 9 juin 2016 au http://dublincore.org/

[36] Norme de métadonnées MODS, consultation le 9 juin 2016 au http://www.loc.gov/standards/mods/

[37] Norme de métadonnées METS, consultation le 9 juin 2016 au http://www.loc.gov/standards/mets/

La norme DOI (ISO 26324:2012[38]) est le fondement du service de couplage DataCite, lequel permet la localisation et le suivi des références citées et comportant des citations dans la documentation savante. Le système DOI propose un cadre pour l'identification permanente, la gestion du contenu intellectuel et, plus important encore, la gestion des métadonnées. On utilise abondamment les DOI dans les publications savantes pour citer des articles de revue et des données de recherche.

ORCID[39] est une initiative sans but lucratif, semblable à DataCite, qui propose un registre des identifiants de chercheur uniques qui met en correspondance de façon transparente des activités de recherche et des résultats savants publiés. ORCID, tout comme les fichiers d'autorité qui existent dans les bibliothèques depuis de nombreuses années, peut couvrir une multitude de disciplines, de secteurs de recherche et de frontières nationales pour clarifier toute ambigüité quant au nom d'un chercheur et ainsi s'assurer que chaque auteur ou chercheur obtient pleinement le mérite auquel il a droit pour son travail. Les identifiants ORCID simplifient également le processus d'identification des publications d'un chercheur lors du suivi des citations, du calcul de l'indice h ou de la création d'un curriculum vitæ pour les organismes de financement. On envisage aussi de s'en servir pour établir des liens entre des sources de données disparates et des résultats savants quand il n'existe aucun lien technique, favorisant grandement l'interopérabilité des systèmes.

Enjeux et considérations

Il existe un certain nombre d'éléments à considérer pour les pratiques exemplaires relatives aux identifiants DOI et ORCID, notamment ceux concernant le contrôle d'autorité et le dédoublement des données. Par exemple, l'attribution de multiples DOI à un même ensemble de données n'est pas considérée comme une pratique exemplaire pour l'identification des données. Il faut des pratiques exemplaires et un certain niveau d'autorité pour l'attribution de DOI aux ensembles de données. Nous suggérons d'établir une version principale de cet ensemble et d'attribuer un DOI uniquement à cette version. Sinon, lorsqu'il faut absolument publier un ensemble de données à plusieurs endroits avec un DOI distinct, les métadonnées pour toutes les manifestations de cet ensemble devraient faire état de cela[40].

Nous avons également constaté qu'il y a beaucoup de données en double dans les dépôts de données au Canada. Il faudrait évaluer la situation activement et, dans la mesure du possible, conclure des ententes entre les parties pour éviter de transmettre des métadonnées en double dans un service de découverte des données.

---

[38] Norme ISO DOI, consultation le 9 juin 2016 au http://www.iso.org/iso/fr/home/store/catalogue_tc/catalogue_detail.htm?csnumber=43506

[39] ORCID, consultation le 9 juin 2016 au http://orcid.org/

[40] Institut canadien de l'information scientifique et technique (ICIST), « Ensembles publiés de nouveau ou en double », consultation le 9 juin 2016 au http://cisti-icist.nrc-cnrc.gc.ca/obj/cisti-icist/doc/datacite/datasets.pdf

Recommandations

Nous recommandons d'envisager une entente nationale pour ORCID[41] que les universités et les organismes gouvernementaux du Canada pourraient utiliser pour intégrer des identifiants de chercheur dans des logiciels de gestion et de publication de recherche institutionnels ou d'autre nature. L'attribution d'un identifiant semblable à ORCID à chaque chercheur permettrait de simplifier et de clarifier le processus de dépôt et de publication des données.

Nous recommandons aussi l'enregistrement des DOI auprès de Datacite Canada[42] pour les ensembles de données présents dans les dépôts participants du Tableau 2 ci-dessus, surtout si le dépôt ou le centre de données n'a aucun moyen de le faire. L'attribution de DOI aux ensembles de données de recherche améliorera grandement la découverte des ensembles de données par l'entremise des partenaires de métadonnées de DataCite (p. ex., DataOne, ORCID, VIVO, etc.). Il convient toutefois de noter que cette mesure doit être prise en collaboration avec les participants et pas au détriment des pratiques exemplaires en matière d'identification des données.

Enfin, nous recommandons la mise en œuvre d'une approche axée sur la communauté et la collaboration dans le but de trouver une solution technique nationale durable pour l'attribution de DOI et d'identifiants ORCID au Canada.

API ouvertes

Il s'agit d'une interface de programmation d'applications (API) qui permet à un composant logiciel d'utiliser la fonctionnalité ou les données mises à la disposition d'un autre par l'entremise d'un ensemble de routines, de protocoles et d'outils. Une API efficace propose une interface d'utilisation fixe qui facilite le développement de programmes informatiques et qui n'oblige pas les futures parties à comprendre et à utiliser l'ensemble du système. Dans un système de découverte, une interface de programmation peut offrir des mécanismes servant à enregistrer des ensembles de données, à consulter les métadonnées de la collection ou à extraire des renseignements (comme des identifiants, des fichiers connexes, des métadonnées, etc.) pour un ensemble de données précis. Il permet surtout aux tiers de créer des outils en s'appuyant sur les systèmes existants et à différents systèmes d'interagir en échangeant des renseignements de manière structurée et uniforme.

[41] Adhésion à l'ORCID, consultation le 9 juin 2016 au https://orcidpilot.jiscinvolve.org/wp/2015/02/03/next-steps-for-orcid-adoption-orcid-consortium-membership-for-the-uk/

[42] DataCite Canada, consultation le 9 juin au http://www.nrc-cnrc.gc.ca/fra/publications/library_services/datacite/index.html

Beaucoup de dépôts et de plates-formes, dont le projet Open Collections de la University of British Columbia[43], utilisent des API pour permettre aux utilisateurs et aux développeurs de lancer des requêtes efficaces, d'effectuer des analyses avancées et de créer des vues, des applications et des composantes personnalisées ayant un accès complet aux transcriptions et aux métadonnées d'Open Collections. Dans le cas de la University of British Colombia, une demande consiste en une adresse URL transmise au serveur Web par protocole HTTP dans le but de recevoir des ressources en format lisible par une machine et un humain. L'adresse URL fournit au serveur Web tout ce dont il a besoin pour créer et retourner une bonne réponse. Il s'agit de l'approche REST pour la conception d'API.

Les API personnalisées et propres à un système donné sont souvent nécessaires pour exposer une fonctionnalité, mais diverses API normalisées sont pertinentes pour les outils de découverte. Le protocole Open Archives Initiative (OAI) est une norme particulièrement pertinente. Il permet l'échange et le moissonnage des métadonnées, et on l'appelle « OAI-PMH »[44]. Cette norme propose des formats uniformes, structurés et interopérables pour l'échange de métadonnées. Elle est donc utilisée par de nombreux services regroupant des données pour le moissonnage des données (Tableau 1) et par des dépôts pour l'exposition des données (Tableau 2).

Enjeux et considérations

- Les données de recherche ne sont pas toutes comprises dans des dépôts ouverts, et les centres de données n'ont pas souvent d'API ou ne soutiennent pas de protocole comme l'OAI-PMH.
- Les API normalisées proposent des interfaces communes pour une multitude de systèmes, mais il faut du temps pour qu'elles deviennent des normes, et elles n'offrent pas nécessairement les fonctionnalités et éléments de découverte les plus récents.
- Le moissonnage des métadonnées ne répond pas aux questions ou aux préoccupations concernant la qualité et l'exhaustivité des métadonnées ou l'adoption de métadonnées communes dans tous les systèmes de dépôt.
- Il faut planifier la fréquence du moissonnage en fonction du cycle de mise à jour et de rafraîchissement, et le processus devrait être automatisé de préférence.

[43] Interface de programmation d'Open Collections de la University of British Colombia, consultation le 9 juin 2016 au https://open.library.ubc.ca/research

[44] OAI-PMH, consultation le 9 juin 2016 au http://www.openarchives.org/OAI/openarchivesprotocol.html

Recommandations

Nous recommandons que les dépôts participants et toute plate-forme nationale de découverte de données offrent une API permettant à des programmes d'accéder aux données et aux métadonnées à des fins de réutilisation et de développement. En plus d'offrir des API propres aux systèmes, il faudrait utiliser des normes d'API, comme l'OAI-PMH, s'il y a lieu.

Licences

Nous sommes d'avis que personne n'a encore résolu tous les problèmes complexes liés à la libre consultation et à la réutilisation des données. En fait, personne n'a encore formulé les questions auxquelles il faut trouver des réponses. Il existe un écart entre les données disponibles et les données réutilisables. En l'absence de métadonnées efficaces et fiables, y compris de métadonnées administratives sur les licences, il est souvent difficile de reproduire ou de réutiliser les données. Le modèle d'octroi de licences Creative Commons[45] est bien établi et il propose une gamme de licences standards, claires, compréhensibles et légalement contraignantes. Les licences Creative Commons (CC) accomplissent deux tâches : elles permettent aux créateurs de partager facilement leurs travaux et elles permettent à tous de trouver des travaux à utiliser gratuitement sans autorisation.

Le dépôt Dataverse, utilisé par Abacus de la University of British Colombia et d'autres au Canada, a recours à la norme CC0 de Creative Commons comme licence par défaut pour tous les ensembles de données de recherche ouverts. Il permet aussi toutefois aux chercheurs de choisir d'autres licences Creative Commons à partir d'un menu déroulant lors du téléversement des ensembles de données. La promotion de l'octroi de licences ouvertes pour les ensembles de données favorisera grandement une réutilisation simple et opportune.

Recommandations

Nous recommandons d'utiliser des licences Creative Commons pour les données de recherche puisqu'elles transmettent efficacement des renseignements sur les intentions des titulaires du droit d'auteur et qu'elles précisent les données à utiliser et les données pour lesquelles il faut obtenir une autorisation. Les licences Creative Commons aident les auteurs et les créateurs à gérer leurs droits d'auteur et à partager leur travail de création sans en perdre le contrôle. En outre, les licences Creative Commons identifient une personne avec qui communiquer pour obtenir des autorisations au besoin. Les licences peuvent s'appliquer aux données et aux métadonnées, même si nous recommandons fortement d'offrir les métadonnées le plus librement possible, avec des restrictions minimales ou nulles sur leur réutilisation pour faciliter la découverte.

---

[45] Licences Creative Commons, consultation le 9 juin 2016 au https://creativecommons.org/licenses/

Prochaines étapes

Une plate-forme nationale de découverte au Canada permettra aux chercheurs de consulter des métadonnées décrivant des ensembles de données provenant d'une multitude de disciplines. Un tel outil pourrait grandement étendre et accélérer la production de connaissances à l'échelle nationale et internationale. Les établissements et les centres de données de recherche participants verraient probablement une hausse marquée de la consultation de leurs ensembles de données en raison de l'exposition accrue. Les administrateurs de recherche pourraient tirer avantage de meilleures statistiques sur la réutilisation et l'impact des données de recherche produites par leur établissement, justifiant ainsi d'une autre manière la reconnaissance des ensembles de données de recherche un des principaux produits de la recherche et atouts institutionnels à part entière.

Nous insistons sur le fait que la collaboration constituera le facteur déterminant de l'amélioration de la découverte des données au Canada. Un projet national bien coordonné permettra de s'assurer que toutes les initiatives pour améliorer la découverte et la consultation des données seront éclairées et facilitées par les attentes, la participation et la collaboration des intervenants. Ce projet garantira l'avancement de l'infrastructure de découverte des données au Canada avec un processus décisionnel éclairé et collaboratif. La mobilisation des intervenants et l'établissement de canaux de communication clairs sont essentiels à la réussite d'un service national de découverte des données.

Nous recommandons d'élargir le Groupe d'experts sur la découverte[46] pour inviter d'autres intervenants à y participer de façon proactive et jouer le rôle de groupe d'experts national sur la découverte des données. Ce groupe élargi pourrait à son tour créer des groupes de travail (utilisateurs, dépôts ou centres de données, établissements de recherche) et leur confier les tâches suivantes :

- choisir les dépôts ou les centres de données à inclure dans le projet-pilote du service;
- créer une politique de développement des collections pour déterminer les données à inclure ou à exclure;
- évaluer et choisir les éléments de métadonnées pour la découverte (p. ex., spatial, temporel, descriptif, technique) et collaborer avec des experts pour mettre au point un modèle de métadonnées;
- mobiliser les établissements d'enseignement supérieur du Canada et les

[46] Groupe d'experts sur l'exploration, consultation le 10 juin 2016 au https://portagenetwork.ca/fr/a-propos-de-portage/reseau-dexperts/membres-/des-groupes-dexperts

centres de données nationaux et propres à des domaines précis, et communiquer avec eux à propos de la collecte des exigences et des cas d'utilisation de recherche pour le projet-pilote d'un service;
- dans la mesure du possible, en tenant d'autres consultations, résoudre les problèmes concernant le dédoublement des données, les métadonnées communes, les licences et l'adoption de normes adéquates avec le milieu des données au Canada (Portage de l'ABRC, DRC, IDD, Calcul Canada, Environnement Canada et d'autres).

La pratique en évolution de la recherche exige de plus en plus que les données et autres sources sur lesquelles s'appuient les conclusions soient diffusées pour permettre la vérification et la réutilisation. Nous le réitérons : il faut que les données de recherche soient repérables pour être réutilisées. Nous sommes enchantés de voir le début de ces travaux au Canada et nous sommes impatients d'entamer d'autres collaborations.

Remerciements

Nous souhaitons remercier Diane Sauvé (Université de Montréal), Chuck Humphrey (Portage) et Martha Whitehead (Queen's University) qui ont fait l'examen par les pairs de notre document et qui ont formulé de précieux commentaires sur son contenu, ses orientations et son style.