Consultation de la collectivité sur la gestion des données de recherche (GDR) réalisée par CANARIE en 2018 : survol

Juillet 2018

Contexte

En janvier 2018, CANARIE a entrepris de consulter les membres de la profession afin de connaître leur point de vue sur les principales lacunes que présente l'infrastructure numérique de recherche (INR) au niveau de la gestion des données scientifiques. Le Comité consultatif sur la gestion des données de recherche (CCGDR) de CANARIE a analysé les commentaires obtenus, puis formulé des recommandations à CANARIE en prévision de son premier appel à projets en gestion des données de recherche (GDR), lancé en mai 2018.

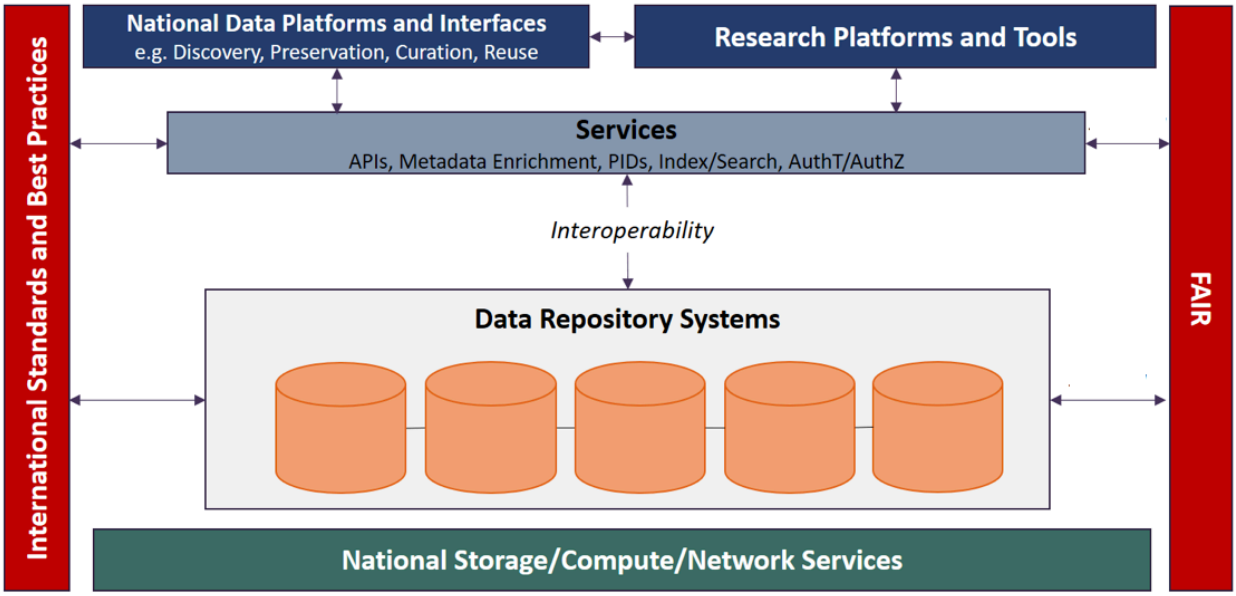
L'appel de 2018 est le premier que lance CANARIE pour encourager le développement d'innovations susceptibles d'améliorer la gestion des données dans l'INR. Ses principaux objectifs sont de faciliter la découverte, la consultation et la réutilisation des données produites par d'autres, de concourir à l'adoption des règles de l'art et au respect des politiques de financement, et d'engendrer l'homogénéité avec les normes nationales et internationales afin de favoriser un vaste partage des données entre institutions de recherche.

La consultation a aussi jeté les bases du nouvel appel à projets en priant les répondants d'expliquer en quoi la lacune ou la possibilité qu'ils avaient identifiée présente des liens avec les cadres de services de données nationaux (CSDN) en train de voir le jour au pays et avec les principes FAIR (***F****indable, **A**ccessible, **I**nteroperable and **U**sable*) voulant que les données soient repérables, accessibles, interopérables et utilisables.

Le CSDN encadrera les efforts de développement afin de rendre les services interopérables et qu'on puisse les intégrer à l'INR aux paliers national et international. Bien que l'élaboration d'un CSDN au Canada n'en soit qu'à ses débuts, le développement d'une INR qui respecte les normes et les protocoles existants en parachèvera la genèse. Dans cette optique, on pourrait notamment réaliser ce qui suit :

- créer une INR accessible aux Canadiens, pareille accessibilité étant déterminée grâce aux politiques d'encadrement, au fil de leur évolution;
- mettre en place des services nationaux de logiciels de données accessibles par domaine ainsi que des plateformes logicielles plus générales au moyen des interfaces appropriées;
- développer ou améliorer les cadriciels spécialisés ou génériques afin d'en rehausser les capacités, pour qu'on puisse s'en servir comme plateforme ou service dans un CSDN;
- monter des équipes qui collaboreront pour intégrer les nombreux logiciels et plateformes de l'INR en existence et en faire un CSDN cohérent.

Le diagramme ci-dessous illustre le principe du CSDN, notamment les principaux thèmes qui sont ressortis de la consultation.



L'application des principes FAIR lors du développement d'un logiciel peut se faire de nombreuses manières. En effet, ces principes ont été établis pour promouvoir les pratiques exemplaires pouvant se refléter dans les services offerts par des humains ou des machines. Ils cadrent donc à merveille avec l'appel à projets de CANARIE qui vise l'élaboration de logiciels.

Thèmes

Après avoir pris connaissance des réponses au questionnaire employé pour la consultation, le CCGDR a cerné huit thèmes à l'intérieur desquels pourrait être proposé un projet. Ces thèmes s'inspirent des commentaires formulés par les membres du milieu canadien de la GDR et contribueront à ce que les efforts se concentrent sur les principales lacunes de l'INR.

La liste qui suit présentent ces thèmes dans un ordre correspondant au nombre de soumissions recueillies lors de l'exercice de consultation. Le premier est donc celui qui revenait le plus dans les présentations et le huitième, le thème mentionné le moins souvent. Dans les descriptions ci-dessous, le terme (méta)données désigne à la fois les métadonnées (les données sur les données) et les données proprement dites. Ces descriptions ont été tirées de la soumission originale pour ne pas divulguer des informations que l'auteur de la soumission pourrait juger confidentielles ainsi que pour donner à la collectivité un juste aperçu des difficultés et des possibilités qui ont été identifiées. Certaines soumissions ont toutefois été rendues publiques dans leur totalité par l'organisation qui en est l'auteur.

1) Enrichissement et découverte des (méta)données

Les résultats des recherches se faisant de plus en plus variés, trouver des ressources et des chercheurs précis devient de plus en plus difficile, surtout à l'échelon national. On pourrait regrouper les (méta)données existantes, venant des plateformes scientifiques et des étapes de la recherche, ou trouver de nouveaux moyens pour enrichir les métadonnées (par ex., prospection des données, associations, extraction d'entités) afin qu'on puisse découvrir plus facilement les résultats des

recherches canadiennes à travers un CSDN. Voici quelques-unes des lacunes ou possibilités spécifiques mises en relief dans les soumissions.

1. Une plateforme nationale qui intègrerait les métadonnées de tout l'écosystème de la recherche et dont un des principaux composants exploiterait les normes comme ORCID et DOI.
2. Des outils qui utiliseraient des API et des ontologies communes pour faciliter la collecte des métadonnées dans les différents systèmes qui émaillent l'écosystème de la recherche
3. Une plateforme nationale qui récolterait les (méta)données des plateformes scientifiques de domaines spécifiques ou les relierait entre elles pour que les données puissent être découvertes plus facilement et être mieux partagées entre les disciplines
4. Des outils qui faciliteraient l'identification des jeux de données à partager et permettraient l'addition d'autres métadonnées avec lesquelles découvrir les données sera plus aisé
5. Des outils qui faciliteraient la collecte et l'enrichissement des métadonnées durant la recherche
6. Des outils qui faciliteraient la liaison ou la découverte de données renfermant des informations susceptibles d'identifier quelqu'un, mais qui préserveraient les renseignements personnels et la vie privée
7. Des outils qui faciliteraient une extraction bonifiée des métadonnées enrichies à partir des résultats textuels de la recherche, donc concourraient à la découverte et à l'analyse

2) Dépôts fédérés / Interopérabilité

On recense plus de 200 dépôts de données au Canada et un nombre plus considérable de bases de données devraient contenir des (méta)données scientifiques utiles. Or peu d'entre elles sont accessibles en raison d'une interface de programmation d'applications (API) mal conçue. On pourrait s'inspirer des normes internationales et des pratiques exemplaires pour mieux documenter les dépôts de (méta)données canadiens et en accroître l'accessibilité, ainsi qu'élaborer des approches mieux pensées pour exploiter et relier ces sources à l'échelon national, ce qui en rendrait la découverte et la réutilisation plus faciles entre les différentes disciplines. Voici quelques-unes des lacunes ou possibilités spécifiques mises en relief dans les soumissions.

1. Un dépôt national assorti d'une plateforme où tous les chercheurs, peu importe leur champ d'activité, pourraient verser leurs données
2. Des outils qui rehausseraient l'interopérabilité des plateformes de recherche de domaines différents pour faciliter la connexion entre chercheurs et jeux de données
3. L'adoption de normes communes pour les métadonnées, l'interopérabilité et les API de manière à faciliter la fédération des dépôts de données disparates
4. Des outils et des services horizontaux auxquels on aurait accès à partir de n'importe quel dépôt de données ou plateforme de recherche, et qui offriraient des fonctionnalités communes facilitant la GDR
5. Une source nationale d'informations administratives et techniques sur les jeux et les dépôts de données canadiens
6. Des outils et des ontologies qui faciliteraient l'harmonisation des données entre les jeux ou dépôts de données du même domaine ou de domaines différents

3) Dépôts particuliers à un domaine

aux chercheurs des sources où déposer et trouver des (méta)données, de manière à mieux répondre aux exigences des organismes subventionnaires ou des diffuseurs, et soutiennent les aspects propres à

un domaine, ce qui accroît l'utilité des (méta)données. On pourrait bonifier les dépôts de données du domaine Canada pour qu'ils soient interopérables à l'échelon national et que ces systèmes s'appuient sur des règles de l'art internationalement reconnues. Voici quelques-unes des lacunes ou possibilités spécifiques mises en relief dans les soumissions.

1. Des dépôts dans les disciplines privées d'un dépôt aux nombreuses fonctionnalités
2. Des outils qui garantiraient un meilleur accès aux données présentes dans les dépôts auxquels il est très difficile d'accéder ou qui exigent passablement de connaissances pour qu'on en exploite les données

4) Dépôt et préservation des données

Les diffuseurs et les bailleurs de fonds adoptent des politiques afin que les chercheurs augmentent l'accessibilité et la réutilisabilité des données à l'origine des résultats qu'ils publient, mais les moyens qui le permettraient sont complexes et peuvent être difficiles à maîtriser pour un chercheur. On pourrait créer des logiciels pour qu'un chercheur puisse identifier et documenter l'approche idéale en gestion des données avec un minimum d'efforts (avec l'aide de services automatisés, par exemple), dans un contexte articulé sur des normes et des pratiques qui fera en sorte que les données adhèrent aux principes FAIR. Voici quelques-unes des lacunes ou possibilités spécifiques mises en relief dans les soumissions.

1. Des outils qui recourraient à des services automatisés pour faciliter les étapes de la préservation des données et appliqueraient des normes à la création de métadonnées bonifiées
2. Des outils qui favoriseraient un versement plus dynamique et précis des données dans les dépôts qui leur sont destinés
3. Des outils qui aideraient les personnes qui ne connaissent pas nécessairement les particularités des données dans une discipline quelconque ou qui n'ont pas coutume de fournir des données à en verser plus facilement dans un dépôt spécifique
4. Des outils et des services qui concourraient à mettre en place une approche nationale à la préservation des données scientifiques bien coordonnée

5) Conservation

Déterminer quelles (méta)données de recherche doivent être conservées et pendant combien de temps est parfois un défi. Cependant, il n'en reste pas moins qu'elles ne sont pas conservées, ces données ne pourront être consultées ni réutilisées à long terme. On pourrait créer des outils TI qui faciliteront la prise de telles décisions, d'une manière qui prendrait en compte les pratiques exemplaires et ferait en sorte que les résultats des recherches demeurent accessibles et exploitables durant une période raisonnable. Voici quelques-unes des lacunes ou possibilités spécifiques mises en relief dans les soumissions.

1. Des outils qui aideraient à déterminer l'utilité à long terme des données ou la période durant laquelle il faudrait les conserver une fois la période de conservation obligatoire initiale terminée
2. Un service ou un registre de logiciels qui serait accessible de n'importe quelle plateforme et qui fournirait des informations canoniques sur le format des fichiers et les normes s'y appliquant, ainsi que sur les mesures connexes qui faciliteraient l'exploitation et la conservation des données

3. L'intégration des outils de préservation à un réseau national fédéré de dépôts de données et de plateformes

6) Persistance des identifiants / Possibilité de citation

Une contrainte à laquelle on ne peut échapper pour rendre les résultats scientifiques accessibles concerne l'adoption des meilleures pratiques reconnues pour rendre les identifiants uniques, autant pour le chercheur que pour son équipe, en raison de la diversité des résultats ou de l'équipement scientifique et des services employés. On pourrait intégrer les plateformes de recherche canadiennes aux meilleures normes opérationnelles pour créer des identifiants persistants qui relieraient plus facilement les différents produits de la recherche et les attributs générés par ceux qui participent à cette activité. Voici quelques-unes des lacunes ou possibilités spécifiques mises en relief dans les soumissions.

1. Le développement de moyens qui faciliteraient une citation dynamique des données dans les plateformes de recherche ou de données
2. L'élaboration de moyens qui permettraient d'identifier plus facilement et de manière unique les échantillons dans les plateformes de recherche ou de données
3. Des outils qui faciliteraient l'intégration des identifiants persistants (DOI ou ORCID, par exemple) aux plateformes et aux dépôts de données existants ou en train d'être mis en place

7) Consultation et analyse des données

Les chercheurs recourent à des approches de plus en plus variées pour analyser et transformer leurs données, que ce soit avec des systèmes locaux ou au moyen de vastes systèmes de calcul de pointe nationaux ou internationaux. On pourrait bâtir des services qui réuniront ces sources de données disparates afin que le chercheur puisse analyser plus aisément les données à toutes les étapes du cycle de recherche. Voici quelques-unes des lacunes ou possibilités spécifiques mises en relief dans les soumissions.

1. Des outils qui créeraient des produits de données évolués susceptibles de faciliter l'accès à des jeux de données volumineux ou complexes et leur analyse
2. Des API qui permettraient des interactions plus complexes avec des jeux de données scientifiques brutes

8) Protection et sécurité des données

Dans certaines disciplines (par ex., santé humaine, biodiversité, sciences sociales), il est difficile de partager les (méta)données sans recourir à un complexe et laborieux processus qui les rendra anonymes ou en assurera la protection. On pourrait créer des logiciels qui faciliteront ce travail grâce à des méthodes semi-automatiques ou à des approches qui relieront les jeux de données sur les individus sans pour autant divulguer les informations de nature personnelle, sauf si on l'autorise. Voici quelques-unes des lacunes ou possibilités spécifiques mises en relief dans les soumissions.

1. Des outils qui faciliteraient le partage des données sensibles, y compris l'intégration des ententes concernant la consultation des données ou des modèles de partage de données à d'autres systèmes

Le but de CANARIE est d'encourager le développement d'une INR à partir des lacunes et des possibilités signalées par les membres de la profession dans le cadre de cette consultation. À mesure que les projets du premier appel avanceront, CANARIE et Données de recherche Canada (organisme que finance

CANARIE) collaboreront avec la collectivité pour faire en sorte que les Canadiens aient tous accès à un réseau d'INR cohérent et interopérable. CANARIE souhaite aussi activement recueillir les commentaires de la collectivité sur la reconduction de son mandat (2020-2025) dans la proposition duquel des fonds sont prévus pour poursuivre le financement des travaux du programme GDR.