Documentation et bibliothèques

Les données et leurs impacts théoriques et pratiques sur les professionnels de l'information

Lyne Da Sylva

Les données et les sciences de l'information
Volume 63, numéro 4, octobre–décembre 2017

URI : id.erudit.org/iderudit/1042308ar
https://doi.org/10.7202/1042308ar

Aller au sommaire du numéro

Éditeur(s)

Association pour l'avancement des sciences et des techniques de la documentation (ASTED)

ISSN 0315-2340 (imprimé)
2291-8949 (numérique)

Découvrir la revue

Résumé de l'article

Les données sont présentes dans l'environnement informationnel actuel sous différentes formes : données confidentielles commerciales ou gouvernementales, mégadonnées, données ouvertes des gouvernements, données ouvertes liées (*Linked Open Data*) du Web sémantique. Comment les professionnels de l'information devraient-ils se préparer pour traiter ces divers types de données ? Nous proposons que cette préparation repose sur trois éléments : une connaissance éclairée des différents types de données en jeu, une initiation aux ressources nécessaires pour traiter chaque type et une compréhension de l'impact qu'aura chacun sur la discipline des sciences de l'information et sur la pratique des professionnels de l'information.

Citer cet article

Da Sylva, L. (2017). Les données et leurs impacts théoriques et pratiques sur les professionnels de l'information. *Documentation et bibliothèques*, 63(4), 5–34. https://doi.org/10.7202/1042308ar

LES DONNÉES ET LEURS IMPACTS THÉORIQUES ET PRATIQUES SUR LES PROFESSIONNELS DE L'INFORMATION

Lyne Da Sylva
Professeure agrégée
Université de Montréal
lyne.da.sylva@umontreal.ca

RÉSUMÉ | ABSTRACT

Les données sont présentes dans l'environnement informationnel actuel sous différentes formes : données confidentielles commerciales ou gouvernementales, mégadonnées, données ouvertes des gouvernements, données ouvertes liées (*Linked Open Data*) du Web sémantique. Comment les professionnels de l'information devraient-ils se préparer pour traiter ces divers types de données ? Nous proposons que cette préparation repose sur trois éléments : une connaissance éclairée des différents types de données en jeu, une initiation aux ressources nécessaires pour traiter chaque type et une compréhension de l'impact qu'aura chacun sur la discipline des sciences de l'information et sur la pratique des professionnels de l'information.

The Theoretical and Practical Impact of Data on Information Professionals

Various forms of data are presented in the current information environment : confidential commercial or government data, big data, government open data, and linked open data of the Semantic Web. How should information professionals prepare themselves to handle or process the different types of data ? We suggest that this preparation be based on three aspects : a clear understanding of the different types of data, an initiation to the resources required to process each type of data and an understanding of the impact that each type will have on information science as a discipline and on the practice of information professionals.

Introduction

Nous voici à l'ère de la révolution des données (voir notamment [Borgman 2015 ; Kitchin 2014a ; Hey, Tansley & Tolle 2009]). Elles sont présentes dans l'environnement informationnel actuel sous différentes formes : données confidentielles hébergées par les institutions financières ou gouvernementales, grands ensembles de données (données massives, ou mégadonnées – *big data*) de diverses sources, données ouvertes (comme celles publiées par les gouvernements municipaux ou nationaux), données ouvertes liées (*Linked Open Data*) du Web sémantique, données de la recherche universitaire, industrielle ou gouvernementale. Devant le déluge des publications reliées au thème des données, il peut être difficile de s'y retrouver. L'objectif de cet article est de dresser un panorama des types de données auxquelles sont confrontés les professionnels de l'information afin de bien distinguer chacun des types, ainsi que d'étudier les conséquences des particularités de chaque type sur la théorie et la pratique de la gestion de l'information. Spécifiquement, nous tentons de répondre à la question suivante : comment les professionnels de l'information (bibliothécaires, archivistes, gestionnaires de documents) devraient-ils se préparer pour traiter ces divers types de données ? Nous allons faire valoir que cette préparation doit reposer sur trois éléments. Le premier est une connaissance éclairée des différents types de données qui sont en jeu. Le deuxième est une initiation aux ressources nécessaires pour traiter chaque type de données. Le troisième élément clé est une compréhension de l'impact qu'aura chacun de ces types distincts sur la discipline des sciences de l'information et sur la pratique des professionnels de l'information, ainsi que l'impact que ces professionnels peuvent avoir à leur tour sur la gestion des données. Le texte qui suit reprend ces trois éléments, s'appuyant sur une recension des écrits sur les données.

Nous nous intéressons principalement aux quatre types de données suivants : les mégadonnées (*big data*), les données

de recherche, les données ouvertes et les données liées. Ces types de données peuvent facilement être confondus, alors qu'ils se distinguent selon un nombre de dimensions que nous présentons ci-dessous.

La démarche de description et de définition de ces divers types de données a son origine dans un constat que nous avons fait en voulant nous documenter sur la notion de la science des données : d'une part, nous avons relevé plusieurs confusions dans la perception de ce qui relève (ou non) de la science des données ; d'autre part, nous avons constaté que le rôle potentiel des professionnels de l'information pouvait être interprété différemment selon le contexte relatif aux données. Nous avons donc voulu éclaircir le sujet, ce qui a mené à une recension extensive des écrits relatifs aux données et qui s'inscrivent dans le champ des sciences de l'information.

[…] sans une compréhension des différents types de données et des distinctions entre elles, le traitement qui leur est apporté peut être déficient. Également, les compétences que les professionnels de l'information doivent développer pour traiter chaque type de données de manière appropriée ne sont pas les mêmes ; ceux-ci doivent donc s'adapter en fonction du contexte et des objectifs.

La présente mise au point nous apparaît primordiale aujourd'hui : en effet, la prévalence du terme « données » dans les nombreux écrits récents (soulevée entre autres par Frederick 2016a) peut donner l'impression que ces données représentent une entité monolithique. Dans les faits, les différents types de données possèdent des caractéristiques différentes, bien que certains aspects soient partagés partiellement. Les articles qui leur sont consacrés ne précisent pas toujours suffisamment de quel type de données il s'agit. Or, sans une compréhension des différents types de données et des distinctions entre elles, le traitement qui leur est apporté peut être déficient. Également, les compétences que les professionnels de l'information doivent développer pour traiter chaque type de données de manière appropriée ne sont pas les mêmes ; ceux-ci doivent donc s'adapter en fonction du contexte et des objectifs. Voilà la motivation principale du présent article.

La section ci-dessous détaille les types de données visées. La section suivante, pragmatique, identifie les ressources appropriées pour chaque type de données. Enfin, la dernière section sera consacrée à l'étude des impacts pour les sciences de l'information et pour la pratique professionnelle. La conclusion fera un retour sur les expertises mobilisées et suggérera des actions à prendre pour les divers intervenants impliqués.

Différents types de données

Nous définissons ici les quatre types de données visés, en donnant quelques exemples de collections existantes. Une attention spéciale sera accordée au contexte canadien et québécois.

Chacun des types de données dont nous allons discuter a des caractéristiques particulières, qui sont définies dans la section « Définitions ». Des liens et des distinctions sont établis entre les types à la section « Premiers liens et contrastes entre les types ». Les enjeux soulevés par les différents types font l'objet de la section « Les enjeux », alors que la section « Retour sur les distinctions et mises en garde » fait un retour sur les distinctions entre les types pour motiver davantage la pertinence de la mise au point présentée dans cet article.

Définitions

Bien que certaines caractéristiques puissent être partagées par les différents types de données, il est plus utile ici de les présenter séparément.

Mégadonnées

Les mégadonnées (Gandomi & Haider 2015 ; Chen & Zhang 2014 ; Kim, Jeong & Kim 2014 ; Boyd & Crawford 2012) peuvent être définies comme suit : « Ensemble des données produites en temps réel et en continu, structurées ou non, et dont la croissance est exponentielle. » (Office de la langue française 2015b) On les dénote également par le terme de données massives ou *big data*. Typiquement, elles sont produites par des méthodes automatiques, plutôt que manuelles ; par exemple, elles peuvent être le produit d'appareils de captation de données (satellites, sondes, etc.) ou de logiciels[1]. Voici quelques exemples d'ensembles de données massives :

- Données océanographiques captées par des sondes sous-marines.
- Données polaires colligées de diverses sources.
- Données astronomiques provenant d'observatoires ou de satellites.
- Données géologiques issues d'analyses tectoniques, géomorphologiques, structurelles et sismiques.
- Données économiques extraites de transactions financières ou commerciales.

1. Notons que (Klapwijk & IFLA Big Data Special Interest Group 2016) identifie toutefois trois modes de création de données : celles recueillies à dessein (surveillance), celles générées par des appareils et celles fournies volontairement, notamment via les réseaux sociaux.

- Données météorologiques produites par les senseurs, capteurs, thermomètres, etc.
- Données de réseaux complexes émanant des médias sociaux.

La nomenclature des « trois V » (Laney 2001) a été proposée pour caractériser les mégadonnées : volume, vitesse et variété. Le volume fait référence à la quantité de données ; la vitesse, au rythme d'ajout de nouvelles données ; la variété dénote l'éventail de sources et de types de données. Certains ajoutent d'autres « V » pour caractériser les mégadonnées : véracité (liée à l'incertitude des données) ou valeur (potentiel d'avantages liés à leur utilisation) (Marr 2014).

Les réseaux sociaux tels que Twitter peuvent être considérés comme une source de mégadonnées, alimentées en continu par les contributions de millions d'utilisateurs et composées de données de types variés (textes et images, notamment). Par le passé, des ensembles importants de données ont été constitués (par exemple, des données personnelles liées aux clients des institutions financières, des corpus linguistiques, des dossiers médicaux, des bases de données bibliographiques) sans en faire pour autant des mégadonnées. Ces ensembles ne satisfont pas aux critères de définition des mégadonnées, soit les trois V, en particulier la variété et la vitesse. Elles sont alimentées par des ajouts manuels et non automatiques, ce qui limite dans les faits le potentiel de croissance incontrôlée caractéristique des mégadonnées. Le propre des données massives, c'est le fait qu'elles croissent de manière importante, en continu. Cela soulève des problèmes quant à leur gestion (voir la section « Mégadonnées » sous la rubrique « Les ressources pertinentes ».).

Les applications des données massives comprennent typiquement la recherche scientifique fondamentale, l'administration publique et le développement de stratégies marketing ou économiques (Chen & Zhang 2014), et aussi l'évaluation du risque et les analyses prévisionnelles, par exemple en météorologie ou en séismologie. Une bonne partie des efforts de développement se réalisent en entreprise et non dans le monde académique (systèmes de recommandation pour les achats en ligne, évaluation des cotes de crédit). Un néologisme révélateur de l'importance que prennent les données à l'heure actuelle est « datafication » (ou « mise en données » [Office de la langue française 2014]), qui désigne la transformation de « plusieurs aspects de notre vie quotidienne[2] » en données.

Données de recherche

Les données de recherche (Erway *et al.* 2016 ; Strasser 2015 ; Ray 2014 ; Guindon 2013) font référence aux données générées à l'intérieur d'un projet de recherche, en milieu académique, gouvernemental ou industriel : par exemple, des observations sur le terrain, des réponses à des sondages ou questionnaires, des données créées par des processus de simulation par ordinateur, etc.

Certaines données sont faciles à répliquer, par exemple des phénomènes physiques courants (chutes de corps, jeux d'optique, etc.). D'autres demandent de l'équipement spécial ou une collecte sur une longue période. Ce sont pour ces dernières en particulier que la préservation et la diffusion seraient les plus profitables à la recherche subséquente (Heidorn 2011, 664). Il est reconnu que la présente ère est celle de la recherche axée sur les données (*data-intensive research*) (Hey, Tansley & Tolle 2009).

Un précurseur, l'ICPSR (Inter-university Consortium for Political and Social Research[3]) collige depuis 1962 des données provenant de projets de recherche en politique et en sciences sociales ; leur collection dépasse les 250 000 fichiers[4].

Parfois, les données de recherche sont des mégadonnées (Meyer & Schroeder 2014) comme celles colligées pour l'année polaire internationale (dont certaines sont disponibles sur le site de la NASA[5] et sur le site du Canadian Cryospheric Information Network[6]) et les données générées par le grand collisionneur de hadrons (LHC) du CERN[7], le plus puissant accélérateur de particules du monde. Cependant, elles peuvent aussi être des « petites données » (*little data*) (Borgman 2015) : des jeux de données individuels liés à un individu en particulier, ou de petits ensembles de données liés à des activités de recherche, des enquêtes, etc. On peut donc y retrouver des ensembles comme les suivants :

- Données statistiques générées automatiquement par des appareils de mesure ou par ordinateur.
- Ensembles de réponses courtes ou fixes à des questions.
- Ensembles de textes (p.ex., réponses longues à des questions ou transcriptions verbatim d'entrevues).
- Texte balisé comme des journaux de consultation de sites Web ou d'exécution de logiciels.
- Contenu multimédia (p.ex., vidéos d'observations ou enregistrements d'entrevues).
- Code informatique généré dynamiquement.

2. Notre traduction à partir de l'entrée Wikipedia pour « *datafication* » en anglais.

3. <www.icpsr.umich.edu/>.

4. <www.icpsr.umich.edu/icpsrweb/content/about/>.

5. <gcmd.gsfc.nasa.gov/KeywordSearch/Home.do?Portal=ipy&MetadataType=0>.

6. <www.polardata.ca/>.

7. <home.cern/fr/topics/large-hadron-collider>.

Les formes sont aussi variées que les méthodologies de recherche possibles pour chaque discipline. Un exemple digne de mention en informatique consiste en ReScience, un dépôt de codes sources disponible pour répliquer les expériences et ainsi mettre à l'épreuve les résultats annoncés par les chercheurs originaux (Rougier *et al*. 2017).

Les données de recherche sont souvent la base de publications comme des articles de revues ou des actes de congrès. Ainsi, elles sont naturellement accompagnées de texte correspondant (les articles publiés) ; elles sont d'ailleurs de plus en plus publiées avec ces mêmes articles (voir une discussion des problèmes relatifs aux liens entre données et publications dans [Mayernik, Phillips & Nienhouse 2016]).

L'objectif invoqué pour ouvrir les données est souvent l'innovation : la possibilité pour différents acteurs de faire davantage avec des données existantes détenues par un organisme.

Parmi les caractéristiques importantes des données de recherche, on note qu'elles sont liées à un chercheur, à un projet de recherche, à une méthodologie spécifique et souvent à des politiques institutionnelles ou gouvernementales. Par exemple, la *Déclaration de principes sur la gestion des données numériques*[8] des organismes subventionnaires CRSNG, CRSH et IRSC du Canada a été publiée en décembre 2016. Une politique « sur la gestion des données numériques s'appliquant à la recherche financée par l'entremise de conseils subventionnaires[9] » est également en cours d'élaboration. Aux États-Unis, la National Science Foundation exige depuis 2011 qu'un plan de gestion des données[10] soit inclus dans toute demande de subvention (Heidorn 2011, 665).

Les liens entre les données et le projet de recherche ont des conséquences importantes sur le traitement adéquat des données. Le lien à un chercheur entraîne des préoccupations liées à la propriété intellectuelle : le chercheur est en droit de se voir attribuer la « paternité » des données. Le projet de recherche duquel les données sont issues leur confère le contexte d'interprétation nécessaire à une bonne compréhension de leur nature et de leur portée ; il est donc primordial que les données soient accompagnées de documentation qui décrit le projet et la méthodologie de collecte ou de création des données. Également, le projet peut impliquer des sujets humains à propos desquels les données doivent être protégées. Enfin, les politiques de gestion des données peuvent par exemple dicter le mode de diffusion ou d'accès, la durée de conservation, etc.

Données ouvertes

Le terme « données ouvertes » (Dickner 2017 ; Comité OGGO 2014 ; Peugeot 2014 ; Janssen, Charalabidis & Zuiderwijk 2012 ; Mercier 2011) fait référence à des « [d]onnées qu'un organisme met à la disposition de tous sous forme de fichiers numériques afin de permettre leur réutilisation » (Commission d'enrichissement de la langue française [France], France-Terme, 2014). C'est typiquement de l'information, surtout sous forme de statistiques, chiffriers ou autres formats tabulaires, qui provient d'un organisme public (mais parfois privé) et qui est rendue disponible publiquement sur le Web. De plus en plus, ces données ouvertes sont associées aux administrations publiques (comme les gouvernements municipaux, provinciaux ou fédéraux), notamment le gouvernement britannique[11], américain[12] (voir notamment [Holdren, Orszag & Prouty 2009]), canadien[13] et autres. Également, des organismes comme les Nations Unies publient de telles données[14].

La diffusion de données s'inscrit dans un mouvement de « gouvernement ouvert » :

> [...] *au début de 2009 paraît aux États-Unis ce que plusieurs considèrent comme un jalon dans le mouvement des données ouvertes : le mémorandum présidentiel sur la transparence et le gouvernement ouvert (Executive Office of the President, Holdren, Orzag & Prouty 2009). Ce document, publié peu après l'élection de Barack Obama, déclenche un important mouvement d'ouverture des données aux États-Unis et donnera une visibilité sans précédent au phénomène.* (Dickner 2017, 14)

Quelques exemples typiques de données ouvertes sont énumérés ci-dessous :

- Budgets d'un gouvernement
- États financiers d'une société sans but lucratif
- Horaires de services municipaux
- Infrastructures publiques comme les piscines, parcs et bibliothèques
- Statistiques de fréquentation d'un événement
- Données météorologiques nationales pour une année entière

Aux données ouvertes publiées par les administrations publiques s'ajoutent un nombre croissant de données

8. <www.science.gc.ca/eic/site/063.nsf/fra/h_83F7624E.html>.
9. <portagenetwork.ca/fr/comment-gerer-vos-donnees/politiques-et-declarations-gdr/>.
10. <www.nsf.gov/eng/general/dmp.jsp>.
11. <data.gov.uk/>.
12. <www.data.gov/>.
13. <ouvert.canada.ca/fr/donnees-ouvertes>.
14. <data.un.org/Default.aspx>.

provenant d'autres sources : celles d'entreprises privées et des organisations non gouvernementales comme le réseau de transport en commun de Longueuil[15] et le service de bicyclette en autopartage Bixi à Montréal[16], ou encore des compilations d'images (Google Earth) ou des contributions individuelles d'utilisateur sur le Web (sous la forme de liens dans les réseaux sociaux ou d'entrées dans Wikipédia par exemple). L'objectif invoqué pour ouvrir les données est souvent l'innovation : la possibilité pour différents acteurs de faire davantage avec des données existantes détenues par un organisme.

À strictement parler, les données sont « Ouvertes » (*Open Data*, avec majuscules initiales) quand elles sont non seulement disponibles librement, mais aussi en formats informatiques qui peuvent être traités aussi aisément par ordinateur que par les humains. Comme contre-exemple, un fichier comprimé (ZIP) qui contient un fichier PDF-image d'un document DOCX que l'on pourrait télécharger d'un site Web, même gratuitement, serait certes « ouvert », mais pas « Ouvert ».

Il est largement reconnu que pour la recherche financée sur fonds publics, les données issues de cette recherche devraient être diffusées librement. Elles représenteraient donc une forme de données ouvertes.

Données liées

Enfin, les données liées (Harth, Hose & Schenkel 2016 ; Bizer, Heath & Berners-Lee 2011) correspondent à une infrastructure technologique utilisée pour encoder des données à l'aide de normes spécifiques développées par le World Wide Web Consortium[17] (W3C). La définition affichée sur le site Web de la communauté des données liées est la suivante : « […] *a term used to describe a recommended best practice for exposing, sharing, and connecting pieces of data, information, and knowledge on the Semantic Web using URIs and RDF*[18]. »

Cette définition introduit donc les noms des normes spécifiques pertinentes (dont RDF et URI, définies à la section « Donnée liées » ci-dessous) et elle fait également référence à la notion de Web sémantique, intrinsèquement associée à la notion de données liées (et que nous abordons à la section « Web sémantique » de la rubrique « Concepts connexes » ci-dessous).

Les données liées peuvent représenter n'importe quel type d'information : des données bibliographiques à des listes de lieux géographiques en passant par des entités biomédicales :

- Données catalographiques d'une bibliothèque
- Données sur des artistes, leurs œuvres et autres informations reliées
- Grille horaire des programmes d'un télédiffuseur
- Données relatives aux génomes d'espèces spécifiques
- Sites touristiques et informations sur leur géolocalisation, leurs spécialités, etc.
- Informations statistiques sur divers pays
- Ontologies géopolitiques
- Données et relations extraites des articles de Wikipédia

L'intérêt de construire un jeu de données liées, c'est lorsqu'elles entretiennent des liens avec d'autres données, ce qui permet d'étoffer les descriptions. Les données liées consistent en de larges jeux d'informations élémentaires (jeux de données ou *datasets*) sur des « entités » individuelles, entités qui peuvent être tout autant des objets du monde réel que des concepts abstraits ou des documents disponibles sur le Web. Par exemple, pour une entité « livre », les informations élémentaires utilisées dans les jeux de données bibliographiques comprennent le titre d'un ouvrage, son auteur, sa date de publication (en d'autres termes, toutes les métadonnées bibliographiques traditionnelles), chaque information encodée de manière individuelle et non rassemblée dans une fiche, contrairement à la tradition bibliothéconomique. Dans le cas d'informations géographiques, pour une entité « lieu » on pourra avoir le nom de ce lieu, sa latitude, sa longitude, le nom de la région dont il fait partie, le climat qu'on y retrouve, etc. Chaque pièce d'information représente une unité distincte dans le jeu de données. Ce morcellement des informations est une caractéristique importante des données liées.

Les données sont encodées à l'aide de triplets, dont les trois éléments sont les suivants : l'entité décrite (ou le « sujet »), une de ses propriétés (ou la « relation ») et la valeur de cette propriété (ou l'« objet »). Des exemples sont présentés au Tableau 1.

TABLEAU 1

Exemple de triplets sujet-relation-objet

Sujet	Relation	Objet
Molière	a-écrit	Amphitryon
Molière	a-pour-année-de-naissance	1622
Molière	a-pour-ville-de-naissance	Paris
Paris	est-la-capitale-de	France

Une caractéristique de la technologie utilisée pour les données liées permet de relier les entités entre elles, d'où le nom « données liées ». Par exemple, on pourra lier Molière à la France à partir des informations présentes dans le Tableau 1 (cela représente une information nouvelle, non représentée explicitement dans les données du tableau, mais que l'on peut inférer à partir de celui-ci). Plusieurs autres types de

15. <www.rtl-longueuil.qc.ca/fr-CA/donnees-ouvertes/>.

16. <bixi.com/fr/donnees-ouvertes>.

17. <www.w3.org/>.

18. <linkeddata.org/>.

liens sont possibles, par exemple relier ce jeu de données sur Molière à un autre jeu de données sur les dramaturges ou sur le mythe d'Amphitryon ou encore sur les événements survenus en 1622.

Le but de l'entreprise est que les données liées soient Ouvertes aussi et donc l'idéal visé est les données Ouvertes liées (*Linked Open Data*).

Un exemple important de données liées est DBPedia[19] : il s'agit de l'encodage d'informations tirées de Wikipédia (par exemple, des informations sur les populations des villes, les dates et lieux de naissance d'individus et bien d'autres choses encore).

Sur le plan des traits partagés, notons d'abord que les données ne représentent pas des documents dans le sens traditionnel du terme. Il est donc normal qu'elles soulèvent, en sciences de l'information, des questionnements nouveaux.

Premiers liens et contrastes entre les types

On pourrait faire le résumé caricatural suivant :

- Les mégadonnées, c'est quand il y en a une grande quantité qui est augmentée en continu.
- Les données ouvertes, c'est quand un gouvernement veut se donner bonne conscience.
- Les données liées, c'est quand un groupe veut profiter des ressources disponibles sur le Web, afficher les siennes et se faire reconnaître comme une autorité en la matière ; ou encore, quand il veut apporter de la valeur ajoutée à ses propres données.
- Les données de recherche, c'est là où les institutions exigent que le financement de la recherche soit bien administré et mieux géré.

Il devrait être clair à la lecture des sections précédentes que les différents types de données décrits possèdent des caractéristiques propres distinctes et soulèvent des problématiques différentes, ce qui sera étoffé dans la suite. Soulignons d'abord néanmoins les caractéristiques qui les unissent.

Sur le plan des traits partagés, notons d'abord que les données ne représentent pas des documents dans le sens traditionnel du terme[20]. Il est donc normal qu'elles soulèvent, en sciences de l'information, des questionnements nouveaux.

Le premier rapprochement qu'on peut faire aisément, c'est de regrouper les données de recherche et les mégadonnées d'une part, et les données ouvertes et les données liées d'autre part. En effet, il existe une intersection non nulle entre les mégadonnées et les données de recherche : certaines données de recherche sont également des mégadonnées. Cependant, le défi lié aux mégadonnées réside dans leur gestion en cours de traitement (Chen & Zhang 2014, 318) : la collecte et l'analyse (Gandomi & Haider 2015), la fouille (Kim, Jeong & Kim 2014), alors que pour les données de recherche il est davantage dans la planification de leur gestion (Digital Curation Centre 2013 ; Guindon 2013, 194), dans leur préservation et dans leur diffusion (Guindon 2013, 191), une fois la recherche terminée. Les données liées, étant pratiquement toujours ouvertes, peuvent être rapprochées des données ouvertes, bien que leur provenance diffère considérablement.

Selon une autre dimension, les données de recherche et les données ouvertes peuvent être rapprochées, en ce qu'elles sont naturellement regroupées en ensembles cohérents. Pour les données liées, bien que des jeux de données soient constitués, il y a à la fois un morcellement des données dans les triplets RDF et un éclatement des données à cause des liens établis entre les jeux. Les ensembles sont donc moins bien clairement définis. Dans le cas des mégadonnées, elles sont bien sûr regroupées en ensembles, mais la taille de ceux-ci défie en quelque sorte leur gestion. Pour ces raisons, il est souvent plus simple de décrire de manière cohérente et utile les données ouvertes et les données de recherche que les deux autres types.

Enfin, les données (ouvertes) liées sont plus spécifiques que les trois premières, puisqu'elles impliquent des technologies spécifiques ; toutefois, leurs propriétés d'ouverture et d'interrelations sont désirables pour tous les types. Les motivations pour l'ouverture sont différentes : pour les données de recherche, c'est la conduite plus efficace de la recherche qui est visée ; pour les données ouvertes (gouvernementales ou publiques), c'est la transparence d'une administration ou encore le potentiel d'innovation.

Cette présentation des différentes caractéristiques distinctes et partagées des types de données soulève la question des enjeux posés par la gestion de ces données, que nous abordons maintenant.

Les enjeux

Comme la nature des jeux de données ainsi que les modes de gestion et les objectifs diffèrent selon les types de données, il est naturel que les enjeux soient différents aussi. Les défis posés par les mégadonnées font l'objet d'un grand nombre de travaux récents dans différentes disciplines (par exemple [Hilbert 2016 ; Chen & Zhang 2014 ; Marx 2013 ; Tole 2013 ; Bizer *et al.* 2012 ; Labrinidis & Jagadish 2012]). Les enjeux posés par les données de recherche préoccupent particulièrement les chercheurs et les professionnels de l'information en milieu universitaire (Koltay 2017 ; Weller & Monroe-Gulick 2014 ; Guindon 2013 ; Borgman 2012 ; Swan

19. <wiki.dbpedia.org/>.

20. Sauf dans certains cas de données ouvertes comme des procès-verbaux de réunions, etc.

& Brown 2008). Pour les données ouvertes, les communautés particulièrement impliquées dans l'étude des enjeux sont celles intéressées par le gouvernement ouvert et la participation citoyenne (Janssen, Charalabidis & Zuiderwijk 2012 ; Zuiderwijk *et al*. 2012) ou celle de la recherche où les impacts sociaux sont importants (Reichman, Jones & Schildhauer 2011). Enfin, les données liées soulèvent divers enjeux technologiques et sociologiques (Bizer, Heath & Berners-Lee 2011) et leur développement, bien qu'alimenté davantage par les informaticiens, est perçu de plus en plus désirable par des communautés d'utilisateurs différents, par exemple en éducation (Dietze *et al.* 2013) et pour les milieux documentaires (Gracy 2015 ; Bermès, Isaac & Poupeau 2013 ; Stuart 2011 ; Hannemann & Kett 2010) ; ce qui unit tous ces utilisateurs des données liées, c'est la plateforme du Web.

Cependant, certains enjeux sont partagés par tous les types de données. Nous en énumérons un certain nombre ci-dessous, déclinés selon les dimensions suivantes : enjeux pratiques, éthiques ou juridiques, technologiques, épistémologiques et enfin économiques.

Enjeux pratiques

Nous avons regroupé ici toutes les considérations pratiques relatives à la gestion des données, qui exigent une certaine planification ou qui demandent de tenir compte de certaines propriétés des données : le défi du volume important (ou non) des données, l'importance de l'effort de collecte, les exigences du partage et de l'accessibilité et les enjeux liés à l'évaluation de la qualité des données.

- Volume : le volume des données représente un défi pour des raisons différentes selon les types de données. Bien que certains jeux de données ouvertes soient de taille modeste, les données de recherche peuvent être de très grands ensembles alors que les mégadonnées sont énormes par définition. Ce n'est pas tant le stockage de quantités énormes de données qui pose problème, mais bien les outils de traitement utilisés ; ceux qui sont disponibles ne réussissent pas toujours le passage « à l'échelle » pour des téraoctets de données. Cela est pourtant primordial si l'on veut extraire les informations pertinentes ou produire de la connaissance à partir des données. Enfin, les jeux de données liées croissent souvent de manière imprévue, étant donné la fragmentation des informations en triplets élémentaires, dont plusieurs sont nécessaires pour décrire une entité donnée. De plus, la taille sans cesse croissante du réseau Linked Data[21] complexifie l'établissement de relations entre les jeux de données.
- Efforts de collecte : les différents types de données ne demandent pas le même type d'effort pour la collecte. Les données de recherche proviennent d'initiatives individuelles ou collectives, sujettes aux aléas de la recherche (disponibilité des subventions, activités de chercheurs individuels ou d'équipes à géométrie variable, etc.), qui sont aussi influencées par des facteurs contextuels variés. Par exemple, les données pour l'année polaire internationale seraient incomplètes : celles datant d'avant 1882 ne seraient pas disponibles, certaines ont été détruites lors des conflits mondiaux dans les années 1930 et 1940, et une bonne partie des données de 1957-1958 auraient été perdues (Brown 2009, 115). Pour les données liées, ancrées dans le Web, il n'existe aucune autorité centralisée pour assurer que la collecte soit homogène, complète, équilibrée, ni démocratique. La collecte ou la mise à disponibilité des données ouvertes, elle, est tributaire des organismes qui publient les données, qui obéissent à leurs propres motivations, selon leur calendrier soumis à différentes contraintes internes politiques, économiques, etc. Enfin, dans les cas où les mégadonnées sont générées de manière automatique, l'effort de collecte dépend de l'appareil utilisé ; peu important pour les clics des réseaux sociaux, il est extrêmement coûteux pour un appareil comme le collisionneur du CERN.
- Partage et accessibilité : en règle générale, tous les types de données sont vouées à être accessibles, diffusées, partagées. Par contre, pour les données de recherche, l'objectif de partage peut aller à l'encontre, jusqu'à un certain point, des intérêts des chercheurs, qui pourraient par exemple vouloir être assurés de pouvoir publier leurs résultats avant que leurs données ne soient réutilisées par d'autres (Borgman 2012) ; du coup, comme ce sont les publications (articles ou communications) qui sont davantage valorisées, les chercheurs ont peu de motivation pour préparer leurs données afin de les partager (Guindon 2013, 191).
- Qualité des données : la qualité des données est un enjeu primordial pour tous les types étudiés ici. La problématique a été étudiée plus en détail pour les données de recherche (voir notamment National Academy of Sciences, National Academy of Engineering & Institute of Medicine 2009). Des jeux de données ouvertes sont ajoutés sans nécessairement qu'il y ait une autorité pour valider leur intégrité et pour les mettre à jour lorsque nécessaire. Cela entraîne des exigences quant au nettoyage des données : la préparation des données pour la diffusion peut révéler des erreurs, des incohérences, des duplicata, etc., problèmes qui doivent être corrigés avant la publication (van Hooland & Verborgh 2014).

L'évaluation de la qualité doit tenir compte du processus de production des jeux de données et d'éventuelles transformations qu'elles auraient pu subir (anonymisation, compilations statistiques, etc.) qui auraient pu altérer les données,

21. <lod-cloud.net/>.

ce qui peut être impossible à déterminer par les responsables de la publication. Comme solution à ce problème, la documentation disponible devrait pouvoir fournir les informations nécessaires pour rétablir, au besoin, l'information correcte.

Dans tous les cas, la prolifération des données est à redouter. Pour les données de recherche, même des données liées à des résultats négatifs, qui ne seront vraisemblablement pas publiés, feront l'objet d'une gestion dans un plan de gestion des données. Cela représente une mutation importante dans l'organisation des données (Heidorn 2011, 665). En fin de compte, la disponibilité des données n'assure en rien leur qualité ni l'utilité pour un utilisateur donné. Cependant, des données de mauvaise qualité minent de manière importante la confiance que pourraient avoir les utilisateurs envers les jeux de données et les organismes qui les produisent.

Les données sont des objets numériques, soumis aux contraintes de gestion informatiques : soucis de préservation à long terme, garantie de sécurité, choix d'outils de traitement automatique combinés à des approches manuelles.

Enjeux éthiques et juridiques

Les données à gérer sont peut-être confidentielles ou la propriété d'ayants droit à considérer.

- Confidentialité : certaines données issues de la recherche, surtout en sciences sociales (Parry & Mauthner 2004), sont liées à des individus particuliers (comportements personnels, lieu de résidence, etc). Il n'est pas souhaitable que ces individus puissent être identifiés à partir des données récoltées. Différentes solutions sont utilisées pour préserver la confidentialité des personnes et des organismes visés, dont l'anonymisation et l'agrégation des informations.
- Propriété intellectuelle : le respect de la propriété intellectuelle est une problématique partagée par toutes les données amenées à être ouvertes ou partagées. Des moyens doivent être mis en place par les diffuseurs pour respecter ces droits (Carroll 2015 ; Maurel 2012 ; McGeever 2007).

L'ouverture des données exige également un comportement éthique de la part de l'utilisateur : « *Just because it is accessible does not make it ethical.* » (Boyd & Crawford 2012, 671)

Enjeux technologiques

Les données sont des objets numériques, soumis aux contraintes de gestion informatiques : soucis de préservation à long terme, garantie de sécurité, choix d'outils de traitement automatique combinés à des approches manuelles.

- Pérennité ou préservation : l'information numérique est encodée sur des médias instables, lisibles par des logiciels particuliers développés pour des plateformes informatiques particulières. Sans une attention portée au mode de stockage et aux méthodes de préservation à long terme, les données peuvent devenir illisibles dans un horizon temporel relativement court (c'est tout le problème de la préservation de l'information numérique) (Bachimont 2017 ; Corrado & Moulaison Sandy 2017). Pour les données de recherche en particulier, l'IWGDD a identifié les problèmes inhérents à l'accès et à la préservation des données numériques de la recherche (Interagency Working Group on Digital Data 2009). Dans le cas des données ouvertes, une situation particulière se présente. D'un côté, ce sont des données produites par une organisation que celle-ci voudra protéger. D'un autre côté, comme elles sont diffusées souvent pour des raisons politiques liées à la transparence ou pour témoigner du dynamisme d'une administration, leur durée de vie utile peut être jugée limitée par les décideurs. Leur valeur est davantage dans l'immédiat que dans la durée. Elles seront conservées à long terme, sans doute, mais dans une base de données dédiée, dans un format qui peut être différent de celui diffusé aujourd'hui. Ainsi, les opérations liées à la préservation de ces données ne seront pas nécessairement appliquées aux jeux repérables en ligne.
- Sécurité des données : à cause des enjeux éthiques et juridiques, il est essentiel que les données soient conservées de manière sécuritaire, à l'abri d'utilisateurs mal intentionnés. On pense également aussi, ici, à l'intégrité des données : avec des quantités impossibles à inspecter de manière exhaustive, comment s'assurer qu'il n'y a pas de corruption dans les données ? Il y a un lien à faire ici avec la question de la qualité des données.
- Traitement : pour être utiles (et utilisées), les données doivent être décrites, organisées, analysées, stockées de manière appropriée. Les défis se présentent de manière différente pour les mégadonnées (défi de mise à l'échelle des techniques, pour traiter de très grandes quantités de données), pour les données de recherche (défis liés à la documentation pour la diffusion et le partage subséquent) ou pour les données liées (qui reposent souvent sur la transformation dans un nouveau format de données existantes). Dans le cas des données ouvertes, c'est la publication (plutôt que le stockage) qui représente le traitement crucial.

Une question connexe est celle du type de traitement, humain vs automatique, approprié pour les données : si certains traitements exigent une intervention humaine (par exemple documenter un jeu de données), d'autres

sont plus aisément faits par des logiciels de traitement automatique, notamment les traitements d'analyse sur les mégadonnées afin d'en extraire de la connaissance (traitement automatique de la langue, apprentissage automatique, fouille de données ou de documents, extraction de métadonnées). Ces derniers risquent cependant d'introduire des avaries dans les données (voir un exemple bien décrit dans Nunberg 2008), ce qui peut remettre en question leur qualité.

Enjeux épistémologiques

La prévalence des données dans toutes les sphères professionnelles et personnelles aujourd'hui soulève bien des questions sur leur statut ontologique (voir notamment Borgman 2015, 17-30). Si on peut reconnaître que, dans le cas des données ouvertes ou des données de recherche, la nature des données n'est pas différente de ce qu'elle a pu être dans le passé, dans le cas des mégadonnées la situation est différente. En particulier, celles-ci amènent des modifications importantes dans la façon de mener la science (Boyd & Crawford 2012 ; Hey, Tansley & Tolle 2009) et dans la définition même de la connaissance (Boyd & Crawford 2012, 665). Les observations encodées dans les données revêtent un caractère objectif et une précision qui occultent en réalité la subjectivité et l'imprécision inhérentes à la sélection et à la préparation des données (Boyd & Crawford 2012, 666-68). Tout cela réactualise la discussion sur la distinction entre les données, l'information et la connaissance (Rowley 2007 ; Zins 2007).

D'autres questions épistémologiques soulevées dans ce dossier des mégadonnées portent sur l'impact de la taille des jeux de données sur la conduite de la recherche (Leonelli 2014 ; Boyd & Crawford 2012, 668), avec l'illusion que les grands volumes de données impliquent que celles-ci soient complètes ; sur l'avantage conféré aux institutions et organismes de grande taille (en moyens matériels et humains) pour le traitement des mégadonnées et pour la direction qu'elles pourront ainsi donner au développement de la science et au déploiement de la technologie (Boyd & Crawford 2012, 674) ; ou encore sur les nouvelles formes d'empirisme centrées uniquement sur les données (« la fin des théories ») (Frické 2015 ; Kitchin 2014b ; Anderson 2008), ce qui peut être sévèrement critiqué : « […] *data-driven science, the "fourth paradigm," is a chimera. Science needs problems, thoughts, theories, and designed experiments. If anything, science needs more theories and less data.* » (Frické 2015, 661) En sciences de l'information, l'impact potentiel des mégadonnées sur l'organisation des connaissances a été évoqué par Ibekwe-Sanjuan & Bowker (2017).

Les enjeux épistémologiques des données de recherche gravitent autour des questions du partage et de la réutilisation des données (Irwin 2013 ; Mauthner & Parry 2013), questions aussi pertinentes pour les données ouvertes.

Enfin, les aspects épistémologiques du Web sémantique et de ses données liées ont été moins étudiés jusqu'à présent ; ils sont abordés dans d'Aquin et Motta (2016), qui relèvent notamment le fait que de tout encoder de la même manière en RDF suggère une homogénéité dans les données qui n'existe tout simplement pas.

> *Somehow, there is one aspect of scalability which is much harder to address by means of purely technical means. In d'Aquin et al. [2014a] we called it diversity : the fact that data and knowledge not only come in different formats and subscribe to different modeling principles, but also that they originate from different sources, might be of different scope and quality, and might be distributed under different constraints, with different regulations applying to them, etc.* (d'Aquin & Motta 2016, 53)

Ajoutons également deux réflexions : d'abord, le fait que les URI (identifiants d'entités) sont utilisés non seulement pour décrire les objets (ce qui tombe sous le sens), mais aussi pour représenter les relations ; or, le statut d'entité rattaché à la notion de relation est pour le moins surprenant (Da Sylva 2017). Ensuite, notons que les représentations sont basées sur une vision simpliste de la sémantique (dont la démonstration dépasse la portée de cet article, mais voir Almeida, Souza & Fonseca 2011).

L'interprétation de la valeur des données doit tenir compte de leur contexte (Boyd & Crawford 2012, 670-71). Pour les professionnels de l'information (en particulier les archivistes), cela va de soi. Les données ne sont, finalement, que la représentation d'une réalité dont les paramètres sont déterminés par le fournisseur du jeu de données.

Enjeux économiques

Les bénéfices économiques de l'information, et plus particulièrement des données ouvertes (Dickner 2017, 30-31), s'ajoutent à leurs avantages politiques (associés à la transparence et à la bonne gestion).

En termes de coûts, la gestion des données de recherche soulève minimalement la question de la formation des intervenants impliqués. Pour les données de recherche, des coûts sont à prévoir à la fois pour les chercheurs, les centres de données et les bibliothèques (Brown 2009, 114).

On peut soutenir que les bibliothèques se doivent de participer au mouvement d'ouverture et de création de biens communs que représentent les données ouvertes (Peugeot 2014). Cependant, des dérives potentielles sont à surveiller (Mercier 2011).

En conclusion : les différents types de données décrites ici possèdent des caractéristiques différentes et soulèvent des

enjeux qui, s'ils peuvent s'apparenter, prennent des formes distinctes selon le cas.

Retour sur les distinctions et mises en garde

Certaines publications apportent bien les distinctions nécessaires entre certains types de données : par exemple, données ouvertes *vs* mégadonnées, dans le contexte des archives (*records management*) (McDonald & Léveillé 2014) ; ou encore, données de recherche et mégadonnées (Federer 2016).

Par contre, la distinction et les parallèles qui peuvent être faits entre les types de données sèment parfois de la confusion dans les écrits. Une confusion fréquente consiste à amalgamer les données de recherche et les mégadonnées (par exemple Koltay 2014), et même les données ouvertes, liées et massives (Janssen & Kuk 2016). Plusieurs auteurs (par exemple [Klapwijk & IFLA Big Data Special Interest Group 2016 ; Zetterlund 2016]) utilisent le terme de « mégadonnées » pour de grands ensembles qui ne sont pas des mégadonnées ; cela démontre une incompréhension du phénomène qui se répercute sur l'évocation de techniques et outils inappropriés pour traiter les données de manière adéquate. En particulier : les algorithmes d'apprentissage automatique développés pour les mégadonnées n'offrent pas des performances optimales sur des ensembles qui sont de taille inférieure à celle des mégadonnées. Notons également la présentation faite par Pouyllau (2013) des données du Web, des données ouvertes et des mégadonnées : l'auteur propose de considérer aussi bien les données de la physique que celles de l'histoire parmi les mégadonnées, bien que la taille des ensembles ne soit pas comparable ; or, même les grands jeux de données de l'histoire ne répondent pas à la définition présentée ci-dessus pour les mégadonnées. Notons enfin que, bien que l'ouverture des données soit un thème récurrent dans la gestion des données, les enjeux ne sont pas les mêmes pour les données gouvernementales (publiques), pour les données de recherche (parfois hautement confidentielles et pas toujours publiques) ou pour les données liées (habituellement ouvertes, pratiquement par définition).

À notre connaissance, aucun article à ce jour n'a été consacré à la caractérisation et à la comparaison systématique des quatre types de données que nous présentons ici.

À notre connaissance, aucun article à ce jour n'a été consacré à la caractérisation et à la comparaison systématique des quatre types de données que nous présentons ici.

Nous reviendrons sur l'importance de bien clarifier le type de données dans un contexte donné lorsque nous aborderons, dans la section « La science des données », la question de la science des données.

Les ressources pertinentes

Les professionnels de l'information doivent être outillés pour traiter les différents types de données. Nous recensons dans les quatre prochaines sous-sections ci-dessous les ressources principales dans chaque cas. Une cinquième sous-section couvrira les ressources générales pertinentes pour tous les types de données, notamment les schémas de métadonnées, les licences encadrant l'utilisation et les occasions de formation pour les professionnels de l'information. Cette présentation sera forcément assez succincte, pour nous permettre de couvrir la base dans chaque cas.

Données ouvertes

Les données ouvertes sont habituellement regroupées en ensembles de données de diverses natures : des rapports financiers annuels, des procès-verbaux de réunions périodiques, des plans ou cartes géographiques situant les services gouvernementaux, des chiffriers contenant des données financières ou statistiques, etc. Chaque ensemble constitue un « jeu de données », soit une collection d'informations de même nature sur un sujet unique. Ainsi, le *Calendrier de la cuisine de rue* à Montréal pour une année donnée[22] représente un premier jeu de données, la *Liste des lots, camions et types de cuisine*[23] en est un deuxième, et le *Taux d'occupation des sites de cuisine de rue pour l'été 2016*[24], un troisième. Les sites Web de données ouvertes peuvent être caractérisés selon le nombre de jeux de données qu'ils contiennent (263 jeux dans le cas de la Ville de Montréal en date du 13 juillet 2017).

La notion du format des données est importante ici. Le troisième jeu de données présenté ci-dessus est contenu dans un fichier XLSX (format Excel de Microsoft) alors que les deux premiers sont disponibles en format JSON (JavaScript Object Notation). Certains sont « plus ouverts que d'autres ». L'organisme The Linux Information Project précise les caractéristiques suivantes pour les formats ouverts (The Linux Information Project 2017) : un format ouvert est un format de fichier pour stocker des données numériques, défini par une spécification publiée (habituellement maintenue par un organisme de normalisation) et qui peut être utilisé et mis en œuvre par quiconque. Un format ouvert

22. <www.bestfoodtrucks.com/api/events/events/?when=today&where=57>.

23. <www.lotmom.com/api/events/what_options/57>.

24. <donnees.ville.montreal.qc.ca/dataset/49a62ebe-25da-4414-95b2-a2fc94100755/resource/0096c4b2-77fe-4c2a-966c-309b908741e8/download/taux-doccupation-ete-2016.xlsx>.

TABLEAU 2

Données ouvertes d'administrations publiques au Canada

Gouvernement	Date de lancement	Nbre de jeux de données	Adresse du site Web
Canada	2011-03-17	118 346	http://ouvert.canada.ca/fr
Provinces			
Alberta	2013-05-28	2 311	http://open.alberta.ca/
Colombie-Britannique	2011-06-19	3 010	https://data.gov.bc.ca/
Nouvelle-Écosse	2016-02-05	433	https://data.novascotia.ca/
Ontario	2012-11-08	2 352	https://www.ontario.ca/fr/page/acces-aux-donnees-gouvernementales
Québec	2012-06-28	844	https://www.donneesquebec.ca/fr/
Terre-Neuve-et-Labrador	2014	91	http://opendata.gov.nl.ca/public/opendata/page
Villes			
Burlington (Ontario)	2011-12-15	36	http://www.burlington.ca/en/services-for-you/open-data.asp
Calgary (Alberta)	2013-12-02	342	https://data.calgary.ca/OpenData/Pages/DatasetListingAlphabetical.aspx
Edmonton (Alberta)	2010-01-13	1 405	https://data.edmonton.ca/
Fredericton (N.-B.)	2011-01-01	34	http://data.fredericton.ca/fr
Gatineau (Québec)	2013-01-01	109	http://gatineau.ca/donneesouvertes/default_fr.aspx
Laval (Québec)	n. d.	26	https://www.laval.ca/Pages/Fr/Citoyens/donnees-ouvertes.aspx
London (Ontario)	2013-06-11[2]	84	http://www.london.ca/city-hall/open-data/Pages/default.aspx
Montréal (Québec)	2011-10-27	263	http://donnees.ville.montreal.qc.ca/
Ottawa (Ontario)	2010-05-12	139	http://ottawa.ca/fr/donnees-ouvertes-ottawa
Québec (Québec)	2012-06-26	65	http://donnees.ville.quebec.qc.ca/
Régina (Saskatchewan)	2012-02-27	235	http://open.regina.ca/
Sherbrooke (Québec)	n. d.	39	http://donnees.ville.sherbrooke.qc.ca/
St-Jean (N.-B.)	2015-11-02	112	http://www.saintjohn.ca/fr/Accueil/hoteldeville/finance/informationtechnology/opendatasj_catalogue/default.aspx
Toronto (Ontario)	2009-11-02	245	http://www1.toronto.ca/wps/portal/contentonly
Vancouver (C.-B.)	2009-09-16	189	http://vancouver.ca/your-government/open-data-catalogue.aspx
Winnipeg (Manitoba)	2014-07-01	477	https://data.winnipeg.ca/

peut être mis en place par des logiciels propriétaires ou par des logiciels gratuits et libres, en utilisant les licences de logiciels typiques utilisées par chacun. Contrairement aux formats ouverts, les formats fermés sont considérés comme des secrets commerciaux. Les formats ouverts sont également appelés formats de fichiers libres s'ils ne sont pas assujettis à des droits d'auteur, des brevets, des marques déposées ou d'autres restrictions afin que quiconque puisse les utiliser sans frais pour un but souhaité.

Les formats ouverts rendent l'accès aux données et leur réutilisation plus faciles. Spécifiquement, les formats ouverts sont préconisés pour les données ouvertes (Obama 2013 ; Braunschweig *et al*. 2012 ; Open Government Working Group 2007). Berners-Lee a proposé un programme de déploiement en cinq étoiles[25] pour les données d'ouverture croissante : une étoile signifie que les données sont simplement disponibles (même en format image) avec une licence ouverte ; deux étoiles, qu'elles sont encodées dans un format de données structurées lisible facilement par ordinateur (par exemple, des données tabulaires en format Excel [XLS] plutôt qu'une image de la même table) ; pour mériter trois étoiles, des formats ouverts (comme CSV plutôt que XLS) doivent être utilisés ; les spécifications « quatre étoiles » et « cinq étoiles » sont réservées aux jeux de données qui exploitent les technologies des données liées, qui seront présentées à la section « Données liées ».

Ressources globales principales

L'Open Knowledge Foundation[26] et l'Open Data Network[27] sont des organismes internationaux impliqués dans la problématique des données ouvertes.

Pour qui veut s'informer davantage ou constituer des jeux de données ouvertes, les ressources de l'Open Data

25. <5stardata.info/fr/>.

26. <okfn.org/>.

27. <www.opendatanetwork.com/>.

Handbook[28], le site Web de la Open Knowledge Foundation[29] et le portail d'Open Knowledge International[30] sont des ressources d'intérêt général. Un logiciel comme OpenRefine[31] permet de nettoyer et transformer des données avant de les rendre disponibles. Un autre, LODRefine[32], ajoute aux fonctionnalités de OpenRefine celles d'ajouter facilement des liens avec DBpedia, d'extraire des entités nommées (ou noms propres de divers types) et d'exporter les données en RDF.

Ressources canadiennes ou québécoises

Une bonne présentation de la situation des données ouvertes de l'administration publique au gouvernement du Québec et à la Ville de Montréal est donnée dans Dickner (2017) : format, qualité, couverture et licence des jeux de données disponibles ainsi que discussion des politiques et des pratiques.

En matière de jeu de données : dans le Tableau 2 ci-dessus sont présentés des jeux de données publiés par différentes administrations publiques du Canada, les plus anciennes datant de 2009, dans le cas de Vancouver et Toronto.

Notons de plus les jeux de données que sont GéoGratis[36] (données géospatiales du ministère des Ressources naturelles du Canada) et les données polaires hébergées sur le site du Canadian Cryospheric Information Network[37] (CCIN).

Données liées

Les données ouvertes liées reposent sur les technologies du Web sémantique (une extension du Web à l'aide de standards définis par le W3C ; voir la section « Web sémantique » de la rubrique « Concepts connexes ») : ces technologies sont principalement les identifiants URI et le modèle de données RDF. La notion d'identifiant n'est pas nouvelle : il s'agit d'un numéro ou d'une chaîne alphanumérique qui sert à identifier de manière unique une entité, comme un numéro d'assurance sociale (pour une personne), un ISBN (pour un livre) ou le numéro d'immatriculation d'un véhicule. Les identifiants URI (*Uniform Resource Identifier*[38]) sont des chaînes alphanumériques qui peuvent prendre plusieurs formes (selon différents schèmes[39]) ; le plus souvent, il s'agit d'une adresse URL (*Uniform Resource Locator*) qui est un lien permettant de localiser la ressource sur le Web. D'ailleurs, la ressource décrite par un URI est souvent une page Web ou un document, mais il peut s'agir aussi d'une entité du monde réel, extérieure au Web (une personne, un lieu, un événement, une date). En fait n'importe quelle entité peut être représentée par un URI, dont quelques exemples sont donnés au Tableau 3.

TABLEAU 3

Exemples d'URI

Identifiant (URI)	Ressource
http://tools.ietf.org/html/rfc3986	document RFC de l'IEFT décrivant les URI
tel : +1-416-555-1212	numéro de téléphone
urn: issn : 1082-9873	la revue D-Lib Magazine
urn : isni : 0000000073659798	l'artiste Gino Vanelli
urn : isni : 0000000123197131	Molière
http://viaf.org/viaf/181287382	l'œuvre Amphitryon de Molière

Certains URI empruntent à des schèmes d'identifiants préexistants (par exemple les numéros ISSN ou ISBN). De plus, alors que certains sont des chaînes ininterprétables générées aléatoirement ou séquentiellement, d'autres sont construits à l'aide de règles qui permettent d'interpréter leur contenu, et même de localiser la ressource sur le Web (on les dit alors « résolvables »). Les URI servent ainsi à identifier les entités ou les données. Celles-ci entretiennent des relations entre elles ; rappelons les exemples présentés ci-dessus, comme le fait que l'œuvre « Amphitryon » soit le titre d'une pièce écrite par Molière. Dans les technologies des données liées, ce triplet est exprimé par une syntaxe particulière, RDF (*Resource Description Framework* [W3C 2004 ; Powers 2003]). Dans l'exemple suivant, encodé en XML, on voit une description RDF où la relation « a-écrit » (ou « auteur ») est exprimée par la relation *creator* puisée au schéma de métadonnées *Dublin Core*[40] (identifié par dcterms).

```
<rdf:Description rdf:about= "http://viaf.org/
viaf/181287382/"/>
    <dcterms:creator rdf:resource= "urn:isni:
    0000000123197131"/>
</rdf:Description>
```

28. <opendatahandbook.org/resources/>.
29. <okfn.org/opendata/>.
30. <dataportals.org/>.
31. <openrefine.org/>.
32. <github.com/sparkica/LODRefine>.
33. En date du 13 juillet 2017.
34. Pour les provinces non représentées, les informations concernant leurs données liées n'étaient pas clairement disponibles.
35. Date supposée, selon le nombre de jeux de données publiés à cette date.
36. <www.rncan.gc.ca/sciences-terre/geographie/information-topographique/donnees-gratuites-geogratis/11043>.
37. <www.polardata.ca/>.
38. <www.w3.org/TR/webarch/#identification>.
39. <www.iana.org/assignments/uri-schemes/uri-schemes.xhtml>.
40. <dublincore.org/>.

On y lit donc que l'entité http://viaf.org/viaf/181287382/ (« Amphitryon ») a été écrite (creator) par l'entité urn :isni : 0000000123197131 (Molière). Bien qu'intimidante au premier abord pour un non-initié, cette technologie ne représente finalement que des encodages de paires de valeurs reliées par une relation, comme Molière-a-écrit-Amphitryon, et qui peuvent par exemple exprimer l'équivalent des données des enregistrements d'un format comme le MARC.

Plusieurs autres technologies participent à l'entreprise des données liées (et du Web sémantique). Parmi celles-ci, on trouve les ontologies, qui revêtent une certaine importance pour les milieux documentaires. On peut définir les ontologies comme un ensemble structuré des termes et concepts qui décrivent un domaine, ou encore « une spécification formelle d'une conceptualisation partagée » (Gruber 1993). Par exemple, l'ontologie FOAF (Friend of a Friend[41]) décrit les propriétés générales des personnes et les liens que celles-ci peuvent entretenir ; elle peut être utile pour établir des liens entre des instances de personnes spécifiques. L'ontologie exprime les relations en général et permet de faire des inférences sur des cas particuliers. Le langage de définition d'ontologies, OWL (*Web Ontology Language*[42]), qui est une recommandation du W3C, est construit sur le modèle de données de RDF.

La notion d'URI ou d'identifiant unique peut être rapprochée de celle du contrôle d'autorité familier aux institutions documentaires. Une différence notable entre les deux, c'est qu'avec les URI, il n'y a pas d'ambition de proposer une forme canonique des étiquettes utilisées (les noms), ni d'en repérer toutes les variantes, mais simplement de définir un endroit précis où l'existence d'une entité est déclarée et à laquelle on peut rattacher les propriétés que l'on veut. Des duplicata peuvent d'ailleurs être créés par des fournisseurs différents de données liées ; par exemple, on recense différentes déclarations d'auteurs ou d'œuvres dans différents jeux de données bibliographiques (ceux de la BnF et de la bibliothèque nationale d'Allemagne contiennent tous les deux des données sur l'œuvre « Amphitryon » de Molière). Une relation définie dans le langage OWL, « sameAs », joue un rôle important dans l'association d'URI distincts qui réfèrent néanmoins à la même entité. Cette relation, on l'espère, devrait permettre d'établir les liens nécessaires, mais son usage polysémique pose en fait plusieurs problèmes (Halpin, Herman & Hayes 2010).

Plusieurs autres technologies participent à l'entreprise des données liées (et du Web sémantique). Parmi celles-ci, on trouve les ontologies, qui revêtent une certaine importance pour les milieux documentaires.

Les jeux de données liées sont emmagasinés dans des bases de données spéciales, appelées *triplestores* (ou entrepôts de triplets). Celles-ci sont essentiellement des implémentations spéciales de bases de données relationnelles. Un langage de requêtes spécialisé (SPARQL[43], semblable à SQL) permet de repérer des données dans ces entrepôts.

Ressources globales principales

Plusieurs jeux de données sont disponibles. Une présentation globale est offerte sur la page The Linking Open Data cloud diagram[44]. Le portail des données ouvertes de l'Union européenne[45] donne accès à des milliers de jeux de données. Parmi les jeux de données liées associées aux milieux documentaires, citons les informations bibliographiques de la Bibliothèque nationale de France[46], les données de WorldCat[47] de l'OCLC, le thésaurus AAT sur l'art et l'architecture du Getty Research Institute[48], le plan de classification IconClass pour les œuvres d'art[49], ainsi que des notices d'autorité et divers vocabulaires de la Library of Congress[50] tels que son répertoire de vedettes-matière.

Divers « vocabulaires[51] » sont utilisés pour décrire les données liées : l'ensemble *Dublin Core* utilisé pour décrire des documents numériques ; l'ontologie FOAF mentionnée ci-dessus, qui sert à décrire des personnes (leur nom, leurs coordonnées, leur titre, etc.) et les liens entre elles (par exemple, le fait qu'une personne en connaît une autre) ; divers vocabulaires contrôlés de la *Library of Congress*, mentionnés ci-dessus. La page Linked Open Vocabularies[52] de la Open Knowledge Foundation présente une collection de ces vocabulaires.

Plus spécifiquement pour les bibliothèques, le regroupement Linked Data for Libraries[53] recense différentes ressources sur son site Web : des ontologies, des sources de données liées, du code, et d'autres ressources utiles pour les bibliothèques.

Pour apprivoiser la recherche par SPARQL dans son entrepôt de données RDF, le portail Persée[54] fournit l'inter-

41. <www.foaf-project.org/>.

42. <www.w3.org/2001/sw/wiki/OWL>.

43. <www.w3.org/TR/rdf-sparql-query/>.

44. <lod-cloud.net/>.

45. <data.europa.eu/euodp/fr/data/>.

46. <data.bnf.fr/>.

47. <www.oclc.org/developer/develop/linked-data.en.html>.

48. <www.getty.edu/research/tools/vocabularies/lod/index.html>.

49. <www.iconclass.org/help/lod>.

50. <id.loc.gov/>.

51. Le terme de « vocabulaire » est utilisé, dans les travaux sur les données liées, pour couvrir des notions qu'on pourra avoir l'habitude de décrire plutôt des schémas de métadonnées, des ontologies, des vocabulaires contrôlés, des terminologies, etc.

52. <lov.okfn.org/dataset/lov/>.

53. <ld4l.org/>.

54. <www.persee.fr/>.

face Sparklis[55], plus conviviale que SPARQL. L'interface peut soutenir également l'apprentissage de ce langage, puisque chaque requête construite à l'aide de l'interface peut être ensuite visualisée en SPARQL natif.

BIBFRAME (Library of Congress 2016, 2012), un nouveau modèle de données pour la description bibliographique développé par la Bibliothèque du Congrès américain, vise à remplacer les formats MARC (*MAchine Readable Cataloging*) et à fournir un format compatible avec le Web de données : « […] l'arrivée de BIBFRAME 2.0 joue inévitablement un rôle important dans la mesure où elle permet d'envisager l'encodage des métadonnées de bibliothèque de manière à ce qu'elle soit [*sic*] déjà prêtes à la publication dans le Web sémantique. » (St-Germain 2017, 148) Exprimé en RDF, BIBFRAME reprend en partie l'approche à la description de ressources proposée par FRBR (IFLA Study Group on the Functional Requirements for Bibliographic Records 2009). BIBFRAME est basé sur trois catégories de base, soit « œuvre » (qui correspond à peu près à la notion d'œuvre de FRBR), « instance » (qui en fusionne les notions d'expression et de manifestation) et « item » (soit un exemplaire particulier). BIBFRAME y ajoute trois classes : agent (pour désigner les auteurs par exemple), sujet et événement, chacune se rapportant aux catégories de base. L'introduction de ce modèle de données est assez récente et il n'est pas encore très répandu ; il n'est d'ailleurs pas tout à fait au point, des dires mêmes de la Bibliothèque du Congrès :

> *BIBFRAME is far from an environment that you could move to yet. The model and its components are still in discussion and development—a work in progress. When it is more mature, vendors and suppliers will need time to adjust services to accommodate it. And then we can expect a mixed environment for some time.* (Bibliographic Framework Transition Initiative, Library of Congress 2017, paragr. #q09).

Il est donc trop tôt pour évaluer l'impact qu'il pourra avoir dans les milieux documentaires.

Ressources canadiennes ou québécoises

Le groupe Canadian Linked Data Initiative[56] regroupe les services des bibliothèques de cinq universités canadiennes, Bibliothèque et Archives Canada ainsi que Bibliothèque et Archives nationales du Québec. Son objectif est de planifier et de trouver un financement pour une série de projets qui coordonneront les activités de production de métadonnées de leurs unités de services techniques, en anglais et en français. Le groupe francophone de cette initiative se penche sur l'identification de ressources en français et vise potentiellement à augmenter le nombre de ces ressources.

55. <data.persee.fr/explorer/sparklis/>.

56. <connect.library.utoronto.ca/display/U5LD/Canadian+Linked+Data+Initiative+Home>.

Données de recherche

La ressource fondamentale nécessaire à la gestion des données de recherche est le plan de gestion des données (PGD), que l'on peut définir comme suit : « […] un plan formel qui décrit comment les données de recherche sont gérées au cours du cycle de vie d'un projet de recherche. Les plans portent sur des thèmes comme la collecte de données, les métadonnées, la documentation, l'échange de données et la préservation[57]. »

Les organismes subventionnaires de nombreux pays dont le Royaume-Uni[58] et le Canada sont en voie d'exiger des candidats de développer des plans de gestion des données, comme c'est déjà le cas aux États-Unis. Des outils en ligne soutiennent les chercheurs dans la rédaction de leur plan (voir des exemples ci-dessous). Perrier *et al.* (2017) offrent un panorama détaillé des écrits sur la gestion des données de recherche dans les institutions universitaires.

Ressources globales principales

Diverses ressources sont disponibles pour gérer et repérer les données de recherche, dont des dépôts disciplinaires ou multidisciplinaires, des moteurs de recherche fédérée d'ensembles de données dans plusieurs dépôts et des normes de citation[59]. Plus précisément :

- Dépôts de données : certaines données sont entreposées par le chercheur lui-même, sur un poste de travail personnel par exemple. Cela peut être suffisant, selon l'ampleur du jeu de données, et à condition qu'une infrastructure adéquate soit disponible (suffisamment d'espace de stockage, bonnes descriptions et organisation des fichiers, copies de sauvegarde fréquentes, etc.). Il existe également des dépôts collectifs utiles aux chercheurs ne détenant pas l'expertise ou le temps pour maintenir un dépôt de données (Cragin *et al.* 2010) : dépôts institutionnels (au sein d'un centre de recherche ou d'une université), nationaux (par exemple le UK Data Archive[60] au Royaume-Uni regroupant les données en sciences humaines et sociales) ou encore disciplinaires (par exemple, DATAONE[61] pour les données environnementales).

  Plusieurs plateformes sont disponibles pour gérer les données de recherche, chacune comportant des forces et des faiblesses (Amorim *et al.* 2015). Entre autres, Dataverse est une application Web de logiciel libre qui permet de partager, de préserver, de citer, d'explorer et

57. <portagenetwork.ca/fr/comment-gerer-vos-donnees/foire-aux-questions/>.

58. <www.dcc.ac.uk/resources/data-management-plans>.

59. <guides.biblio.polymtl.ca/c.php?g=590745&p=4084488>.

60. <www.data-archive.ac.uk/>.

61. <www.dataone.org/>.

d'analyser des données de recherche (Crosas 2011). Le projet est hébergé par l'Université Harvard et est développé par l'équipe Dataverse à l'Institute for Quantitative Social Science (IQSS). Des plateformes génériques développées au départ pour des bibliothèques numériques comportant davantage des documents textuels numériques, comme DSpace et Invenio, peuvent également gérer des données.

- Identification : les jeux de données reçoivent un identifiant unique DOI (*Digital Object Identifier*), fourni par DataCite[62], qui est utilisé notamment pour citer cette ressource.
- Documentation, guides, etc. : plusieurs documents sont disponibles auprès du Digital Curation Centre, au Royaume-Uni[63]. D'autres guides ont été publiés par l'organisme américain NISO (Strasser 2015) et par les académies nationales de recherche américaines (National Academy of Sciences, National Academy of Engineering & Institute of Medicine 2009).

Ressources canadiennes ou québécoises

Les organismes suivants œuvrent à la gestion des données recherche :

- Le réseau Portage de l'Association canadienne des bibliothèques de recherche[64] « rassemble le milieu des bibliothèques pour coordonner l'expertise, les services et la technologie dans le domaine de la gestion des données de recherche et, dans ce contexte, cherche à collaborer avec d'autres acteurs de ce domaine ». Il comprend plusieurs groupes d'experts[65] : sur la planification de la gestion des données, sur la préservation, sur la découverte des données, sur la formation en gestion des données de recherche, sur « la recherche et l'intelligence » (ou planification stratégique), et sur l'organisation des données ;
- Research Data Canada (Données de recherche Canada) « veille à ce que les données scientifiques engendrent des innovations dont profitera chaque Canadien[66] ». Il s'agit d'une « organisation dirigée et appuyée par des intervenants déterminés à améliorer la gestion des données scientifiques au Canada[67] ». Les partenaires incluent plusieurs universités canadiennes, l'Association canadienne des bibliothèques de recherche, certains ministères et agences gouvernementaux, des fournisseurs d'infrastructure de recherche et de services (dont CANARIE et Calcul Canada, présentés à la section « Mégadonnées - Ressources canadiennes ou québécoises ») et plusieurs autres membres.
- Le Réseau canadien des centres de données de recherche[68] (RCCDR) « permet aux chercheurs d'accéder à une multitude de microdonnées sociales, économiques et liées à la santé, qui sont recueillies et administrées par Statistique Canada ».
- Le Sous-comité des bibliothèques du Bureau de coopération interuniversitaire (BCI) du Québec « a pour mandat spécifique de favoriser le développement concerté des collections et des services des bibliothèques universitaires québécoises, en appui à la mission d'enseignement et de recherche des établissements universitaires québécois[69] ». La Journée de réflexion sur la gestion des données de recherche[70] (organisée par le Sous-comité des bibliothèques du BCI) a réuni à Montréal le 21 novembre 2016 une centaine de participants, provenant des quatre coins du pays afin d'aborder spécifiquement les défis de la gestion des données de recherche au Québec et au Canada.

Mis de l'avant par Portage, l'Assistant PGD[71] (ou *DMP Assistant*) est un outil bilingue d'aide à la préparation d'un plan de gestion des données (PGD). Le canevas de base (*template*) incorporé dans l'Assistant amène les chercheurs à définir les aspects suivants du PGD :

- Collecte de données
- Documentation et métadonnées
- Stockage et sauvegarde
- Conservation
- Partage et réutilisation
- Responsabilités et ressources
- Conformité aux lois et à l'éthique

Une plateforme pour la gestion des données a été développée par les bibliothèques universitaires de l'Ontario. Le portail Scholars (Scholars Portal développé par l'Ontario Council of University Libraries [OCUL]) fournit une infrastructure technologique partagée et des collections partagées pour les 21 bibliothèques universitaires de la province. Il a été certifié comme un dépôt numérique fiable (*Trustworthy Digital Repository*[72]). C'est un ensemble de services, contenant des publications aussi bien que des

62. <www.datacite.org/>.
63. <www.dcc.ac.uk/>.
64. <portagenetwork.ca/>.
65. <portagenetwork.ca/fr/collaborer-avec-portage/reseau-dexperts/>.
66. <www.rdc-drc.ca/>.
67. <www.rdc-drc.ca/fr/qui-nous-sommes/>.
68. <www.cihr-irsc.gc.ca/f/48922.html>.
69. <www.bci-qc.ca/comites/affaires-academiques/sous-comite-des-bibliotheques/>.
70. <cbpq.qc.ca/formation/journee-de-reflexion-sur-la-gestion-des-donnees-de-la-recherche-gdr>.
71. <assistant.portagenetwork.ca/fr>.
72. <onesearch.library.utoronto.ca/library-media-release/ocul%E2%80%99s-scholars-portal-canada%E2%80%99s-first-certified-trustworthy-digital-repository>.

données. Il incorpore une base de données en sciences sociales (odesi pour *Ontario Data Documentation, Extraction Service and Infrastructure*) ainsi qu'une implémentation de Dataverse[73]. Il a été conçu principalement pour les données de recherche recueillies par des chercheurs et des organisations affiliées aux universités de l'Ontario, bien que n'importe qui dans le monde soit invité à utiliser le portail Scholars Dataverse pour déposer, partager et archiver des données[74]. Le Sous-comité des bibliothèques du BCI, au Québec, tente d'amorcer des travaux visant la mise sur pied d'une plateforme analogue pour le Québec, mais les travaux sont toujours en cours.

La manipulation de très grands ensembles de données requiert un équipement informatique important et de très grandes capacités de calcul.

Une plateforme nationale fédérant les dépôts de données de recherche canadiens[75] est en cours de préparation ; son lancement est prévu pour janvier 2018 et une version bêta est disponible[76]. Une recherche effectuée avec un terme donné permet d'y repérer des jeux de données provenant de différents dépôts. Par exemple, avec le terme « *polar bear* » on peut repérer des jeux de données provenant du réseau *Polar Data Network*, du UBC Circle et du ministère des Ressources naturelles du Canada.

Mégadonnées

La manipulation de très grands ensembles de données requiert un équipement informatique important et de très grandes capacités de calcul. Les outils et technologies pour traiter les mégadonnées peuvent être regroupés dans les quatre groupes suivants (Chen & Zhang 2014, 321-31) : techniques « directes », techniques appliquées en lots, techniques appliquées en continu et techniques interactives. Le premier groupe représente en fait les techniques de base appliquées aux données : optimisation (relevant de la recherche opérationnelle), statistiques avancées, fouille de données (éventuellement fouille de texte), apprentissage automatique, techniques de visualisation et analyses de réseaux sociaux (ancrées dans la théorie des réseaux et des graphes). Dans le deuxième groupe de techniques de traitement, on trouve des logiciels et plateformes spécifiques (comme Hadoop, Dryad ou Pentaho) qui sont particulièrement robustes et sont optimisés pour fonctionner sur de très grands ensembles. Le troisième groupe d'outils et de technologies sont ceux conçus pour appliquer des traitements continus, en temps réel et de manière incrémentale. Enfin, les technologies interactives permettent aux utilisateurs d'ajouter leur propre analyse en interagissant avec le système. Des principes fondamentaux pour soutenir la conception d'outils de traitement sont énoncés dans Chen & Zhang (2014, 331-32).

Ressources globales principales

Les ressources principales pour les mégadonnées sont plus intéressantes pour les informaticiens et les mathématiciens appelés à appliquer des traitements sur les données (présentées brièvement ci-dessus). Certaines bibliothèques universitaires, en particulier celles associées à des chercheurs en sciences axées sur des mégadonnées, pourront toutefois être amenées à appliquer leur expertise en termes de description et d'organisation des données (voir alors la section sur les données de recherche). Graduellement, des bibliothèques reconnaissent que les bénéfices de traiter de grandes quantités de données peuvent s'appliquer à la gestion des bibliothèques elles-mêmes : ainsi, à partir par exemple de statistiques de fréquentation ou de prêt, des connaissances pourraient être extraites pour améliorer les services offerts (Klapwijk & IFLA Big Data Special Interest Group 2016 ; Zetterlund 2016). Bien que ces données ne

TABLEAU 4

Schémas de métadonnées pour les données de recherche et les données ouvertes

Type de données	Schéma de métadonnées	Description
De recherche	DDI (Data Documentation Initiative)[84]	Développé initialement pour les ensembles de données en sciences sociales et comportementales
	DataCite[85]	Pour la publication et la citation des données de recherche
	Métadonnées incluses dans les assistants de PGD	Spécifiques à chaque plateforme
Ouvertes	Data Catalog Vocabulary (DCAT)[86]	Vocabulaire RDF conçu pour faciliter l'interopérabilité entre les catalogues de données publiés sur le Web
	Project Open Data Metadata Schema[87]	Schéma de métadonnées basé sur DCAT

73. <dataverse.scholarsportal.info/>.

74. <guides.scholarsportal.info/dataverse>.

75. <portagenetwork.ca/frdr-dfdr/francais/>.

76. <beta.frdr.ca/repo/>.

répondent pas à la définition des mégadonnées données ci-dessus, nous mentionnons ces initiatives ici.

Ressources canadiennes ou québécoises

Les organismes suivants sont actifs au Canada et au Québec en particulier : Institute for Big Data Analytics[77] à l'Université Dalhousie et IVADO (Institut de valorisation des données, pôle scientifique et économique[78]), associé à l'Université de Montréal, à l'École Polytechnique de Montréal et HEC Montréal. Notons également, le Leadership Council for Digital Infrastructure ou Conseil du leadership sur l'infrastructure de recherche numérique[79], qui « regroupe plusieurs organisations et institutions qui veillent à ce que les chercheurs du Canada aient un accès à des technologies numériques de pointe et à des compétences reconnues dont ils ont besoin pour réaliser des projets qui comportent des calculs complexes et un imposant volume de données[80] ».

Par ailleurs, les chercheurs canadiens peuvent compter sur CANARIE[81], organisme à but non lucratif responsable d'un réseau de télécommunications à haut débit qui fédère les réseaux scientifiques provinciaux et ainsi les universités, instituts de recherche, laboratoires publics et écoles au pays. Enfin, Calcul Canada[82] est responsable de la plateforme nationale de calcul à haute performance employée pour la recherche au pays. Elle permet aux chercheurs canadiens d'effectuer des travaux de recherche de calibre mondial à l'aide de stratégies de calcul informatique de pointe.

Ressources générales pour tous les types de données

Nous avons reporté jusqu'ici la description de trois types de ressources qui sont utiles pour tous les types de données et qui s'ajoutent aux ressources énumérées dans les sections qui précèdent. Il s'agit des métadonnées, des licences encadrant l'utilisation des données (ouvertes ou partagées) et des occasions de formation (webinaires, ateliers, vidéos, etc.).

Les métadonnées et les schémas de métadonnées

Les métadonnées (Riley 2017) consistent en des ensembles de données structurées dont l'objectif est de décrire un document ou un jeu de données. Elles correspondent essentiellement aux informations de description (ou de catalogage) familières aux bibliothécaires et aux archivistes (il y a donc plusieurs parallèles à faire avec les vocabulaires contrôlés).

À côté de schémas de métadonnées généraux comme celui du *Dublin Core* ou de MODS[83], tous deux particulièrement utiles pour décrire des documents, certains schémas de métadonnées spécialisées ont été créés spécifiquement pour les données de recherche et les données ouvertes (voir le Tableau 4).

Des métadonnées seraient utiles pour les mégadonnées, mais elles ne sont pas toujours disponibles (Rousidis *et al.* 2014). Dans le cas des données liées, la question des métadonnées sera abordée à la section « Les métadonnées : une réflexion ciblée ».

Les licences encadrant l'utilisation

Les données ouvertes ou disponibles peuvent être protégées par des formes de licences encadrant la propriété intellectuelle, sans pour autant empêcher leur (ré)utilisation, ce qui est particulièrement important pour les données de recherche. Les plus importantes sont les suivantes :

- Licences *Creative Commons*[88] (CC) : le groupe CC, une association à but non lucratif, a développé une solution juridique alternative pour les auteurs d'œuvres, afin de libérer ces œuvres des droits de propriété intellectuelle standard. Plusieurs licences sont incluses dans l'offre de CC, selon quatre critères : mention de l'auteur original (« *attribution* »), usage commercial interdit (« *non commercial* »), modification de l'œuvre interdite (« *no derivative works* ») et transmission obligatoire de la licence originale dans toute réutilisation (« *share alike* »). Les licences sont définies sur le Web et leur adresse URL peut être incluse aisément dans du code HTML ou RDF et donc associée sans difficulté à des jeux de données.
- *Open Data Commons*[89] *Attribution Licence*[90] : cette ressource regroupe trois licences ayant le caractère des licences *Creative Commons*, mais conçues spécifiquement pour les données et les bases de données.

Pour exprimer les licences et les associer aux données, divers mécanismes sont disponibles (Ball 2014, 12-15), dont l'utilisation d'un langage d'expression des droits (*Rights Expression Language*) comme les suivants :

- Langage ODRL (*Open Digital Rights Language*[91]) : fruit d'un effort international visant à développer et pro-

77. <bigdata.cs.dal.ca/>.
78. <ivado.ca/>.
79. <digitalleadership.ca/fr/>.
80. <digitalleadership.ca/fr/a-propos-du-conseil/>.
81. <www.canarie.ca/fr/>.
82. <www.computecanada.ca/ ?lang=fr>.
83. <www.loc.gov/standards/mods/>.
84. <www.ddialliance.org/>.
85. <schema.datacite.org/>.
86. <www.w3.org/TR/vocab-dcat/>.
87. <project-open-data.cio.gov/v1.1/schema/>.
88. <creativecommons.org/>.
89. <opendatacommons.org/>.
90. <opendatacommons.org/licenses/>.
91. <www.w3.org/community/odrl/>.

mouvoir un standard ouvert pour l'expression de politiques relatives à la publication, à la distribution et à la consommation de contenu, applications et services.

- Langage MPEG-21 REL (International Organization for Standardization 2004)
- *METSRights schema*[92] de la Library of Congress.

Un guide utile portant sur les licences pour les données de recherche est publié par le Digital Curation Centre au Royaume-Uni (Ball 2014).

Les formations disponibles

Pour la formation initiale, dans l'information disponible en ligne sur les programmes de maîtrise en sciences de l'information au Canada, nous avons recensé quelques exemples de cours liés à la gestion des données : Indexation de collections numériques[93] (U. de Montréal) ; *Data Mining* (McGill) ; *Data Management* et *Introduction to Data Science*, *Business Analytics and Data Visualization* (Dalhousie) ; *Data Analytics : Introduction, Methods and Practical Approaches* et *Data Librarianship* (Toronto) ; *Information Visualization and Visual Analytics* et *Research Data Management for Information Professionals* (UBC). Il y a aussi des cours sur des thématiques reliées : Fouille de documents (U. de Montréal) ; *Digital Curation* et *Digital Libraries* (McGill) ; *Data Modeling and Database Design* et *Digital Preservation and Curation* (Toronto) ; *Digital Libraries* (Western) ; *Text Analytics* (UBC).

Sur le plan de la formation continue, un grand nombre d'occasions sont offertes aux professionnels de l'information pour se former sur l'une ou l'autre des thématiques abordées dans cet article. D'abord, les associations professionnelles offrent des webinaires sur ces thématiques : la Corporation des bibliothécaires professionnels du Québec (dont « Web sémantique et bibliothèques numériques ») ; Données de recherche Canada[94] ; ASIS&T[95] ; l'ASTED (dont « Quand les sciences de l'information se mettent au service de la recherche »). Ensuite, des ateliers sont organisés par l'Association des bibliothèques de recherche du Canada sur les données de recherche[96]. Enfin, voici quelques exemples de vidéos en ligne :

- « Le Web des données ouvertes et liées. Qu'est-ce que c'est ? » par EuropeanaEU[97]
- « RDC Webinar : What is RDC[98] ? » et « RDC Webinar : Persistent IDs : Best Practices[99] » par CANARIE Inc.
- Chaînes YouTube : Le Réseau canadien des centres de données de recherche[100] (RCCDR), l'ABRC[101]

La vidéo « Introduction à la gestion des données de recherche » sur le site du service des bibliothèques de l'Université de Montréal[102] fournit une initiation utile à la problématique ; des initiatives semblables se retrouvent dans les bibliothèques d'autres universités de recherche au Canada.

Une recherche sur le Web utilisant les mots-clés « formation », « webinaire » ou « atelier », combinés aux concepts clés identifiés dans cet article, révélera d'autres occasions de formation imminentes.

Cette section a présenté les ressources de base utiles pour gérer les différents types de données. L'exercice visait à fournir un point de départ pour explorer les problématiques et s'outiller selon le cas.

Impact des données sur les professionnels de l'information ou vice versa

La prolifération des données a un impact sur plusieurs groupes : les citoyens de manière générale, les chercheurs, les gouvernements, les sociétés privées ; cet impact se fait ressentir de manière particulière à chaque groupe. Ce qui nous préoccupe particulièrement ici, c'est l'incidence sur les professionnels de l'information. Nous examinons la question pour chaque type de données, relevant en particulier les différents intervenants et leurs tâches respectives, ainsi que les connaissances et compétences des professionnels de l'information mobilisées pour la gestion des données. Enfin, nous abordons des concepts qui sont connexes à notre propos et qui seront examinés à la lumière de la présentation que nous avons faite des données.

Données de recherche

Les organismes subventionnaires en sont venus à reconnaître l'importance de préserver les données issues de la recherche pour permettre leur réutilisation, rendant ainsi plus performantes les différentes initiatives de recherche subventionnée. Pour ce faire, la description et l'indexation de ce type de données sont cruciales, car elles sont bénéfiques pour les chercheurs, pour les gestionnaires des

92. <www.loc.gov/standards/rights/METSRights.xsd>.

93. Ce cours axé sur l'indexation et la description des collections numériques aborde la gestion des jeux de données ainsi que les données liées du Web sémantique.

94. <www.rdc-drc.ca/fr/activites/webinaires/>.

95. <www.asist.org/events/webinars/>.

96. <www.carl-abrc.ca/fr/accroitre-la-capacite/ateliers-et-formation/atelier-sur-la-gestion-des-donnees/>.

97. <www.youtube.com/watch?v=oEuDaJjEFos>.

98. <www.youtube.com/watch?v=yaf_2Yl-OlM>.

99. <www.youtube.com/watch?v=lBvqPDfRGHg>.

100. <www.youtube.com/user/TheCRDCN>.

101. <www.youtube.com/channel/UCK59-sdDLfQgUUoAuiOVQeQ>.

102. <youtu.be/ZaqfEcqy0kU>.

systèmes et pour les organismes subventionnaires. Les bibliothèques (en particulier les bibliothèques universitaires) se sont trouvées à être les institutions tout indiquées pour préserver et diffuser les données. Cela s'inscrit dans leur mission d'acquisition et de diffusion de l'information pour répondre aux besoins de ces institutions d'enseignement et de recherche (Heidorn 2011, 663). Les bibliothèques et les centres d'archives « sont des institutions reconnues pour le rôle dans la préservation et le partage des connaissances » (Guindon 2013, 191).

Les professionnels de l'information seront graduellement appelés à soutenir les chercheurs, non seulement dans le stockage de leurs propres données, mais aussi dans la réutilisation de données provenant d'ailleurs (Lucic & Blake 2016).

Plusieurs bibliothèques universitaires développent des services liés aux données de recherche (*Research Data Services* ou RDS) (Tenopir *et al*. 2015), particulièrement pour les opérations de planification de la gestion des données, de réutilisation, de visualisation et de partage des données (Federer 2016, 39). Une problématique semblable est associée aux dépôts institutionnels (Guindon 2013, 189).

Intervenants et tâches respectives

Parmi les rôles joués dans la gestion de données de recherche (Swan & Brown 2008) se retrouvent les créateurs de données (les chercheurs), les « *data scientists* », les gestionnaires de données (y compris le bureau des technologies de l'information et la direction institutionnelle de la recherche) et les bibliothécaires de données. Ce dernier groupe, issu de la communauté des bibliothèques, est constitué de professionnels formés et spécialisés dans la préservation et l'archivage des données.

Un rapport rédigé pour JISC (organisme britannique qui soutient la recherche universitaire) dès 2008 souligne que les compétences nécessaires à la gestion des données concernent le stockage et la préservation :

> *Data managers, by our definition, are individuals with specialist skills in computational science, experts in database technologies, and are responsible for ensuring that data produced and needed by the research team are properly stored, curated and preserved.* (Swan & Brown 2008, 13)

Diverses tâches liées à chaque étape du cycle de vie de la recherche sont identifiées dans Guindon (2013, 194) ; dans cette analyse, on voit que les bibliothécaires participent à chaque étape, que ce soit dans l'identification de sources de données secondaires, dans le choix d'un schéma de métadonnées, dans la vérification finale des données, dans la gestion des citations, etc. Ray (2014) explore les stratégies que peuvent déployer les professionnels de l'information dans la gestion des données de recherche.

Connaissances et compétences mobilisées

La gestion de ces données nécessite à la fois une vue d'ensemble du processus de recherche et une compréhension de ce qu'il est maintenant convenu d'appeler le « cycle de vie » de la recherche (Inter-university Consortium for Political and Social Research [ICPSR] 2012 ; Higgins 2008). Le modèle pour ce cycle de vie doit être pris en compte ; les pratiques ne sont pas uniformes entre les disciplines (Weller & Monroe-Gulick 2014), ce qui exige des connaissances spécialisées de la part des professionnels de l'information.

Pour assurer la gestion efficace des données de recherche, les bibliothécaires peuvent s'appuyer sur leurs compétences dans les domaines suivants :

- Organisation des données : l'organisation repose sur leur description. Un défi de taille qui se pose est que chaque jeu de données sera différent, et les métadonnées nécessaires vont varier beaucoup (Heidorn 2011). Il importera de maintenir des liens avec les communautés expertes (soit les utilisateurs premiers des données) d'où vont émaner les normes et pratiques applicables.
- Préservation des données : les connaissances en préservation de l'information numérique seront particulièrement sollicitées pour les données, essentiellement de l'information « née numérique » (Jacobs & Humphrey 2004).
- « Curation » des données (Mayernik 2016) ou gestion à long terme (Guindon 2013) : cette expertise repose à la fois sur la description, la planification et la préservation.
- Pratiques de citation : des formes normalisées pour la citation de données ont été développées (Ball & Duke 2015) ; l'aisance qu'ont les bibliothécaires avec la description de ressources en fait de bons défenseurs des pratiques de citations correctes.

L'apport potentiel des bibliothécaires et l'impact sur leur travail sont bien décrits par Guindon :

> *Pour les bibliothèques universitaires, les spécialistes des données et tous ceux qui s'intéressent à la préservation du patrimoine numérique, il s'agit d'une occasion unique de faire valoir leur expertise et d'étendre le champ de collaboration avec les professeurs. Ignorer ce nouveau paradigme ou ne s'y engager qu'avec réticence comporte des risques, notamment celui d'être pris de vitesse par d'autres acteurs (centres de recherche, entreprises commerciales) et d'être relégués à un rôle marginal dans le nouveau monde de la recherche… le candidat idéal pour s'occuper des services de gestion des données devrait avoir une expérience pratique de recherche, ce qui implique au moins un autre diplôme d'études supérieures en plus de la maîtrise en sciences de l'information.* (Guindon 2013, 199)

Le paradoxe des données de recherche

Les données de recherche, administrées par les bibliothèques universitaires, présentent un certain paradoxe. Elles ont pourtant certaines caractéristiques qui les rapprochent des archives :

- Elles sont produites par un organisme (un chercheur ou un groupe de recherche) dans l'exercice de ses fonctions.
- Elles sont un produit unique, comme une pièce d'archive ; toute production ultérieure n'est pas une « réédition », mais un nouveau produit unique.
- Quand de multiples versions sont créées, le chercheur peut être appelé à décider quelle version est la version officielle, laquelle doit être conservée, etc.
- Elles ne sont pas publiées en exemplaires multiples (bien qu'elles puissent être téléchargées de multiples fois) ; en fait, elles ne sont pas généralement publiées dans le sens traditionnel de la publication de livres ou d'articles.
- Elles doivent être interprétées dans leur contexte de production.
- Elles sont vouées à être conservées et gérées à long terme (*curated*) pour une période déterminée (ou indéterminée).
- Comme préconisé par l'approche d'archivistique intégrée (Rousseau & Couture 1994), on reconnaît que la gestion des données doit être pensée dès leur conception (Guindon 2013, 192 ; Higgins 2008, 18).

> *D'ailleurs, plusieurs concepts archivistiques, dont celui du cycle de vie des documents et celui de provenance, sont repris et adaptés par cette discipline émergente.* (Guindon 2013, 192)

Cependant, la gestion des données de recherche est confiée aux services de bibliothèques des universités ou des centres de recherche. Il est à prévoir que l'expertise dans la gestion de produits organiques des archivistes viendra influencer les efforts dans ce domaine.

Données ouvertes

Les données ouvertes sont particulièrement présentes dans les organismes publics et ainsi auront davantage d'impact sur les citoyens ordinaires. En général, la quantité de données disponibles de sources variées permet aux citoyens d'utiliser celles-ci pour générer des idées, des outils et des services, par exemple l'application Info-Neige[103] à Montréal, qui permet aux utilisateurs de suivre les périodes d'interdiction de stationnement dans les rues de la ville, selon les opérations de déneigement en cours, et ainsi savoir où garer leur voiture. Pour les administrations publiques, la charte des données ouvertes (*G8 Open Data Charter*[104]) a été écrite pour encourager les décideurs à faire la promotion de la transparence, de l'innovation et de la responsabilité (*accountability*). Ici, l'impact pour les professionnels de l'information se situe au niveau de ce qu'ils peuvent apporter à l'entreprise : typiquement, les données ouvertes souffrent d'un manque d'organisation et de métadonnées, ce qui limite ou complique leur découverte et leur compréhension de la part des utilisateurs (Comité OGGO 2014, 11).

Intervenants et tâches respectives

On peut identifier comme parties prenantes pour les données ouvertes : les gouvernements, les entreprises et les citoyens (Deloitte LLP 2012) ; il peut être utile d'y ajouter les responsables des technologies de l'information (TI) ainsi que les gestionnaires de données (habituellement un sous-ensemble des membres du gouvernement ou des entreprises). Ce sont ces gestionnaires qui sélectionnent et décrivent les données (sujettes à l'approbation des dirigeants), qui sont ensuite mises en ligne par les TI.

Les données ouvertes ne concernent pas tant les bibliothécaires que les archivistes ou les gestionnaires de documents à l'emploi d'administrations publiques comme les villes, par exemple (si on fait abstraction des données de recherche qui peuvent être ouvertes également). À titre d'illustration, le seul jeu de données provenant d'une institution documentaire au portail Données Québec est Bibliothèques et Archives nationales du Québec et il ne représenterait que 5,6 % des données (Dickner 2017, 60). Si on ajoute les musées, la contribution des institutions patrimoniales s'élève à 8,6 %. Bien sûr, par contre, le Tableau 2 illustre le fait que les villes s'y mettent graduellement, habituellement grâce à la collaboration de leurs archivistes. Par ailleurs, Mercier (2011) incite les bibliothèques à participer à l'effort d'ouverture des données : « On se prend à imaginer que le mouvement de l'*Open data* force les bibliothèques à ouvrir leurs données et que puisse se construire un catalogue commun de grande qualité largement disséminé sur le Web. » (Mercier 2011, 8)

Connaissances et compétences mobilisées

Les professionnels de l'information peuvent faire deux types de contributions importantes. Du côté de la mise en ligne, ils peuvent élaborer la description des données et l'ajout de métadonnées. Du côté de l'utilisation : pour le public, les professionnels de l'information peuvent aider à faire l'évaluation nécessaire de la qualité des données ou de la pertinence pour un besoin d'information donné, ce qui peut rappeler la situation au début du Web (Frederick 2016b, 11).

103. <infoneige.ca/>.

104. <www.gov.uk/government/publications/open-data-charter>.

Données liées

Le mouvement de diffusion de données liées provenant des milieux documentaires est déjà bien amorcé. En effet, des données ouvertes liées sont publiées par certaines institutions documentaires. La publication de ces données représente d'une part une occasion pour les bibliothèques et les archives de s'établir en figure d'autorité en publiant leurs données propres, et d'autre part d'augmenter la visibilité de leurs données en les interconnectant avec celles d'autres organismes et institutions. Il y a en fait une grande urgence à faire ce pas : les technologies actuelles de recherche d'information (par les moteurs de recherche destinés au grand public) font de plus en plus usage de jeux de données liées disponibles et ainsi contournent les institutions d'information que sont les bibliothèques et les archives : « Les catalogues en ligne restent dans le "Web profond", c'est-à-dire qu'ils sont des silos dont les données sont isolées et ne sont accessibles que dans la mesure où un usager est au courant de leur existence. » (St-Germain 2017, 46)

Intervenants et tâches respectives

Dans une optique technocentrée, Rietveld a identifié quatre parties prenantes pour les données liées : les créateurs (qui créent les données à publier), les fournisseurs de contenu (responsables de la publication), les développeurs (qui construisent les interfaces et autres applications utiles) et les scientifiques (qui analysent et consomment les données liées) (Rietveld 2016, 3). Dans un contexte documentaire, on devrait tenir compte également d'autres intervenants, comme les experts de contenus, en l'occurrence les bibliothécaires ou archivistes qui sont les architectes et les gardiens des données documentaires destinées à être publiées en données liées. Il faut tenir compte également dans ce domaine de l'influence majeure du W3C dans la définition de technologies et de pratiques recommandées.

La publication de ces données [liées] représente d'une part une occasion pour les bibliothèques et les archives de s'établir en figure d'autorité en publiant leurs données propres, et d'autre part d'augmenter la visibilité de leurs données en les interconnectant avec celles d'autres organismes et institutions.

Comme la publication de données liées est une activité relativement nouvelle, le rôle des professionnels de l'information comprend normalement la sensibilisation et la formation aussi bien des gestionnaires ou décideurs que des utilisateurs.

Connaissances et compétences mobilisées

Comme mentionné ci-dessus, plusieurs triplets RDF correspondent à l'expression de métadonnées ; la correspondance possible avec le format MARC ou avec RDA (*Resource Description and Access*) permet aux professionnels de l'information d'appliquer leur expertise à une nouvelle technologie. Le passage d'un format à l'autre (par exemple, de MARC à RDF) exige de faire un *mapping* ou transcodage, qui peut être établi par une personne compétente dans les deux formalismes ; en l'occurrence, un professionnel de l'information formé aux technologies du Web sémantique. La méthodologie proposée par St-Germain pour migrer des données bibliothéconomiques à des données liées détaille davantage les étapes et les compétences nécessaires (St-Germain 2017, chap. 6). De plus, certains jeux de données liées représentent des métadonnées descriptives, destinées à être utilisées par exemple comme termes d'indexation ; il y a là un lien important à faire avec les vocabulaires contrôlés existants. D'ailleurs, le cadre SKOS (*Simple Knowledge Organization System*[105]) a été élaboré pour représenter en RDF les systèmes d'organisation des connaissances comme les thésaurus. Parmi les thésaurus disponibles en RDF/SKOS se retrouvent AGROVOC[106] et le thésaurus de l'UNESCO[107]. D'autres thésaurus pourraient être migrés ou transcodés ainsi.

Des compétences générales en recherche d'information doivent être mobilisées, plus précisément en requêtes (à l'aide du langage SPARQL). En règle générale, les professionnels de l'information qui se tournent vers les données liées doivent avoir une aisance à manipuler un formalisme initialement rebutant. À terme, des interfaces plus conviviales devraient être développées, qui masqueront les détails inutiles aux professionnels de l'information et qui permettront d'appréhender les données à un niveau conceptuel — mais nous n'y sommes pas encore tout à fait (l'interface Sparklis fait exception ici). La non-convivialité de RDF (Anadiotis 2017) représente effectivement un frein à toute tentative de développement.

Mégadonnées

Le rôle des bibliothécaires dans la gestion des mégadonnées a été abordé déjà par plusieurs chercheurs (Huwe 2017 ; Swan & Brown 2008). Crastes (2015) y inclut la description des données, l'ajout de métadonnées et l'évaluation de la fiabilité et de la qualité. Effectivement, les professionnels de l'information ont des compétences en évaluation de l'information et peuvent aider à évaluer les deux « V » que sont la « véracité » et la « valeur » des données.

Certains peuvent être inquiets de la pertinence des bibliothécaires dans un monde où l'information disponible en

105. <www.w3.org/2004/02/skos/>.

106. <aims.fao.org/standards/agrovoc/linked-open-data>.

107. <skos.um.es/unescothes/?l=fr>.

ligne, abondante et facilement accessible, peut être analysée par des algorithmes d'apprentissage automatique appliqués aux données : « […] *there will undoubtedly be those who will claim that in the future there will be no need for libraries because the information we need will be embedded in our environment…* » (Frederick 2016b, 11)

Frederick note cependant que les bibliothèques, par leurs processus et normes bien établis, peuvent fournir des bases de connaissance pour construire les outils automatiques de gestion des données :

> *Perhaps library data, with its mature and robust procedures and standards for creation, has the capacity for providing a baseline from which artificial intelligence can learn to detect critical errors in the growing mass of data on the web. If such a use for library data were to come into being, libraries and information science can be expected to grow in relevance and importance in the near future.* (Frederick 2016b, 11)

Cela pourrait contribuer à des solutions pour désamorcer les « fausses nouvelles », les « faits alternatifs », la désinformation, les fraudes, etc.

Intervenants et tâches respectives

Les intervenants ici comprennent les fournisseurs de données, les utilisateurs (par exemple les chercheurs et citoyens), les « scientifiques des données » (*data scientists)*, les bailleurs de fonds, les éditeurs (responsables de la publication) et les bibliothèques (Margolis *et al*. 2014).

Marchionini (2016) suggère que les professionnels de l'information peuvent jouer un rôle dans l'organisation, dans la description (par les métadonnées) et dans la préservation des mégadonnées ; ils peuvent aussi alimenter la réflexion sur le traitement éthique des données.

Makhlouf Shabou relève à son tour que les archivistes peuvent aider dans l'évaluation des données :

> *Une des raisons d'être de l'évaluation est justement d'éliminer l'information qui perd son importance au fur et à mesure de l'avancement de son cycle de vie et qui est jugée inutile aussi bien pour l'activité humaine que pour la mémoire sociétale. Il va falloir alors investir dans le développement des méthodes, des outils et des critères de cette fonction pour qu'elle fournisse des réponses adéquates qui conviennent à la quantité, à la nature et aux flux des mégadonnées (*big data*). L'enjeu le plus important sera alors le paramétrage et le contrôle de l'élimination des mégadonnées (*big data*). Le combat d'avenir pour les archivistes est l'élimination.* (Makhlouf Shabou 2015)

Cela rappelle ce que nous avons relevé ci-dessus à propos du traitement des données « inutiles » en recherche, parce que stériles en termes de résultats de recherche.

Connaissances et compétences mobilisées

Les mégadonnées interpellent davantage les informaticiens et les statisticiens, d'une part, et les chercheurs, d'autre part. Cependant, les bibliothécaires devraient en savoir suffisamment sur les mégadonnées pour adapter leur pratique en conséquence :

> *Because of its prevalence and potential impacts, librarians need to know the basics of big data and how it affects academic research. Business librarians need to know how companies leverage big data, how such data mining provides a competitive advantage, and how students might need to grapple with big data sets in future employment. Science librarians need to know how big data differs from other scientific data and the impact of emerging software and hardware used for its analysis. Humanities and Social Science librarians should know that big data is becoming more commonplace in their disciplines as well, and is no longer restricted to corpus linguistics. Librarians in all disciplines, in order to facilitate the research process, will need to be aware of how big data is used and where it can be found.* (Bieraugel 2013)

La gestion des mégadonnées exige donc minimalement une bonne connaissance des utilisateurs et des caractéristiques des ressources à gérer, deux compétences qui appartiennent déjà à l'arsenal du professionnel de l'information.

Avant de résumer nos observations, il est utile de définir brièvement (pour les distinguer) quelques concepts connexes aux différents types de données décrits.

Concepts connexes

Ces concepts sont les métadonnées, l'Internet des objets, le Web sémantique, la cyberscience et la science des données.

Les métadonnées : une réflexion ciblée

Les métadonnées, qui sont elles-mêmes des données sur d'autres données, sont fondamentales à la gestion d'ensemble comme les données ouvertes et les données de recherche, mais quel est leur lien avec les autres types de données ? La réponse n'est pas simple dans le cas des données liées.

Les données liées représentent des informations élémentaires au sujet d'une ressource ; on peut les décrire comme une relation binaire entre deux éléments. C'est le cas, par exemple, quand on relie deux personnes qui « se connaissent », notamment par la relation codifiée « *knows* » à l'aide du vocabulaire FOAF mentionné ci-dessus. Dans d'autres cas cependant, les données liées correspondent exactement à la notion de métadonnées : des exemples sont les informations descriptives exprimées par le vocabulaire

du *Dublin Core* (comme le titre et le créateur d'un document). D'ailleurs, plusieurs schémas de métadonnées sont utilisés comme vocabulaires pour les données liées. Heery (2004) souligne les ressemblances entre les données liées et les métadonnées bibliographiques traditionnelles :

> *What is perhaps the most striking aspect of the Semantic Web for the library community is the commonality between traditional information management and library interests (constructing vocabularies, describing properties of resources, identifying resources, exchanging and aggregating metadata) and the concerns that are driving the development of Semantic Web technologies.* (Heery 2004, 270)

On peut alors se demander où se situe la frontière entre les données liées et les métadonnées (Da Sylva 2017).

Internet des objets

« *The Internet of things* » (IoT), ou Internet des objets, « représente l'extension d'Internet à des choses et à des lieux du monde physique », selon Wikipédia, ou encore « [e]nsemble des objets branchés à Internet capables de communiquer avec des humains, mais aussi entre eux, grâce à des systèmes d'identification électronique, pour collecter, transmettre et traiter des données avec ou sans intervention humaine » selon le Grand Dictionnaire terminologique (Office de la langue française 2015a). Ces objets comprennent tout appareil interrogeable ou contrôlable à distance à l'aide de radio-identification (RFID) (Chabanne, Urien & Susini 2013 ; Draetta 2012) : appareils de domotique (Harper 2011 ; Jeuland 2005), dispositifs d'e-santé (ou informatique médicale) (Venot, Burgun & Quantin 2013) et analyse de données personnelles (*personal analytics*) ou de « soi quantifié » (*quantified self*) (Swan 2013). L'Internet des objets tient compte des échanges de données provenant de dispositifs existant dans le monde réel vers le réseau virtuel Internet. Il est donc la source de données potentiellement très nombreuses, diffusées en continu ; soit, des mégadonnées.

Une mauvaise gestion de l'information (et des données) pourrait avoir des répercussions négatives importantes dans le monde réel : « […] *it seems reasonable to assume that for the "fourth industrial revolution*[108]*" to see full fruition, issues regarding data management and integrity on the open web will need to be addressed.* » (Frederick 2016b, 11)

L'impact sur les professionnels de l'information pourrait être considérable : « […] *it is possible that in an "Internet of things" library data could become a part of the infrastructure human beings rely on to carry out their everyday activities. This is just one way in which the role and function of libraries could change as part of a new industrial revolution.* » (Frederick 2016b, 11)

C'est donc un sujet duquel les professionnels de l'information devraient se préoccuper.

Web sémantique

Le Web sémantique (Barrière 2013 ; Vatant 2008 ; Charlet, Laublet & Reynaud 2005) représente un ajout au Web actuel (dit « de documents ») à l'aide d'une infrastructure qui s'y superpose. Le Web sémantique vise à expliciter la sémantique des informations décrites par les ressources sur le Web afin qu'elles puissent être manipulées de manière automatique par des agents logiciels. Ces agents peuvent alors faire des inférences éventuellement très complexes sur les données, des inférences plus élaborées que le simple exemple évoqué ci-dessus où on peut conclure qu'il existe un lien entre Molière et la France, à partir du même jeu de données. Cela permettrait par exemple des recherches sophistiquées où la réponse ne peut être trouvée qu'en combinant des données de sources distinctes, par exemple repérer des gens selon à la fois leur provenance, leurs activités et celles de membres de leur famille, ou encore, proposer un circuit touristique dans une ville, en tenant compte des sites patrimoniaux, de la localisation des musées et de leurs heures d'ouverture ainsi que de l'âge du public cible de leurs expositions, des trajets d'autobus ou de métro, des prévisions météorologiques et d'événements spéciaux en cours. Ce sont toutes des informations sans doute disponibles en ligne, mais il faudrait non seulement qu'elles soient encodées de manière standardisée (par exemple, en RDF), mais aussi qu'on leur ajoute des instructions pour préciser comment prendre en compte chacune des informations et les combiner afin de créer un parcours touristique cohérent qui correspond au profil d'un utilisateur donné. On peut imaginer des applications également complexes pour les milieux documentaires : par exemple, un système de recommandation de lectures qui tiendrait compte à la fois des lectures précédentes d'un utilisateur (les titres, les auteurs et les thèmes qui y sont développés), de la disponibilité des titres à suggérer et d'événements d'actualité (saison en cours, festivals locaux, salons du livre ou de l'automobile…).

Ainsi, afin d'effectuer ce type de traitement complexe, les simples données liées ne sont pas suffisantes ; il faut leur ajouter les autres technologies définies dans le *Semantic Web Stack*[109] (voir aussi Hitzler 2010). Parmi celles-ci se trouve le raisonnement, c'est-à-dire les outils pour définir des règles d'inférence (notamment SWRL ou *Semantic Web Rule Language*). Les projets de données liées n'incluent pas souvent cet aspect ; il présente pourtant un grand intérêt pour les sciences de l'information, offrant la possibilité de participer à l'élaboration d'outils de traitement sémantique pour bonifier considérablement les outils de repérage de l'information.

108. Celle associée à l'utilisation de données massives.

109. <en.wikipedia.org/wiki/Semantic_Web_Stack>.

L'ouvrage de Bermès, Isaac & Poupeau (2013) définit le Web sémantique et y reconnaît le rôle des bibliothèques, dans la continuité de leurs missions et activités traditionnelles. Il s'agirait d'ailleurs d'« un puissant levier pour la visibilité des données de bibliothèques et leur mise en relation avec des données provenant d'autres univers professionnels » (Peyrard & Simon 2014).

Cyberscience ou e-science

Le vocable e-science (ou eScience, ou encore cyberscience selon l'Office de la langue française du Québec) décrit un nouveau modèle scientifique axé sur l'analyse de grandes quantités de données dont le traitement est rendu possible par les capacités informatiques actuelles :

> *Characterized by large-scale, distributed global collaboration using distributed information technologies, eScience is typically conducted by a multidisciplinary team working on problems that have only become solvable in recent years with improved data collection and data analysis capabilities.* (Luce 2008, 42)

Le terme e-science aurait été inventé par John Taylor, alors directeur général du bureau de la science et de la technologie du Royaume-Uni (Office of Science and Technology) en 1999, à peu près en même temps que celui de « *big data* » (mégadonnées), attribuable à John Mashey vers 1998[110]. Les deux concepts sont intimement reliés, l'un (les mégadonnées) décrivant l'objet, et l'autre, l'approche méthodologique (ou le paradigme scientifique). Des exemples de disciplines où se développe la cyberscience : l'astrophysique, la physique nucléaire, le génie des matériaux, les nanotechnologies, la biologie computationnelle, la génomique, la protéomique.

Les professionnels de l'information pourraient donc contribuer utilement à la science des données, malgré les différences entre les approches traditionnelles des sciences de l'information d'une part, et de la science des données et de l'informatique d'autre part.

La science des données

Toutes les considérations explorées dans le présent article sembleraient culminer en ce qui se nomme la science des données. Or, ce n'est pas tout à fait le cas.

La science des données désigne un nouveau champ multidisciplinaire qui connaît un essor considérable depuis quelques années. Il s'appuie sur des outils mathématiques, statistiques ou informatiques (dont des techniques d'intelligence artificielle) (Baškarada & Koronios 2017) pour développer des méthodes, des processus et des systèmes pour gérer de grands ensembles de données, éventuellement pour extraire des connaissances de celles-ci. Le terme par ailleurs est souvent considéré comme une appellation plus accrocheuse des statistiques (Silver 2013).

La science des données est distincte de la cyberscience : cette dernière peut avoir comme objet d'étude (comme nous l'avons vu) la physique, la biologie ou le génie, étudiés à l'aide de données. La science des données, elle, a pour objet d'étude des données elles-mêmes. C'est la science **au sujet** des données et non la science qui utilise les données (qui, faut-il le rappeler, n'est que la science, finalement).

Ainsi définie, la science des données s'intéresse essentiellement aux enjeux du traitement des données. Elle est naturellement à mettre en relation avec les mégadonnées. Elle ne concerne pas les données ouvertes ni les données ouvertes liées (*Linked Open Data*), à moins que celles-ci ne soient des mégadonnées par ailleurs.

Un intérêt accru pour la science des données s'est développé au moment de la parution en 2009 d'un ouvrage phare (Hey, Tansley & Tolle 2009) dans lequel les auteurs consacraient la naissance d'un nouveau paradigme de recherche, axé sur les données (alors que les précédents étaient caractérisés par [i] l'analyse expérimentale directe des phénomènes naturels, puis [ii] la science théorique basée sur le développement de modèles et enfin [iii] la simulation de phénomènes complexes par ordinateur).

Les intervenants identifiés relèvent essentiellement des mathématiques, des statistiques et de l'informatique : expert du domaine, « ingénieur de données », statisticien, informaticien, communicateur et chef de projet (Baškarada & Koronios 2017).

Sur le rôle que peuvent jouer les sciences de l'information dans la science des données, Marchionini souligne d'une part l'apport qui peut être fait sur la reconnaissance du cycle de vie des données et d'autre part sur les fonctions classiques de la gestion de l'information : description (métadonnées), organisation, considérations éthiques, représentation et préservation (Marchionini 2016, fig. 1). L'auteur offre de plus la réflexion suivante, à savoir que les sciences de l'information devraient être intégrées dans les disciplines constitutives de la science des données (et non l'inverse) :

> *It could be argued that data science is a subset of information science and some data science training programs may be housed in information schools, however, it is more strategic to view information science as an essential component of data science so that the emerging field can benefit from the diversity of perspectives that interdisciplinary collaborations bring. Information science programs can participate as key partners to ensure that students are prepared to take the lead on the sociocultural issues noted above.* (Marchionini 2016, 5)

110. <en.wikipedia.org/wiki/Big_data>.

Les professionnels de l'information pourraient donc contribuer utilement à la science des données, malgré les différences entre les approches traditionnelles des sciences de l'information d'une part, et de la science des données et de l'informatique d'autre part.

> *Data science and informatics, as emerging fields, are expanding our understanding of how the massive amount of information currently being generated can be collected, managed and used. While these may not be traditional «library» problems, the contributions of the library and information science communities are critical to help address aspects of these issues.* (Cervone 2016, 7)

Récapitulatif

Nous avons examiné différents types de données, avec leurs caractéristiques distinctives particulières, dont : les domaines d'application, les publics cibles prioritaires ou potentiels, les formats et la taille des jeux de données, leur confidentialité variable, leur qualité incertaine, les technologies utiles ou nécessaires pour leur traitement et la problématique de préservation à moyen ou à long terme. Nous avons également dégagé les rôles qui pouvaient être dévolus aux professionnels de l'information.

La gestion des données vient modifier la pratique des professionnels de l'information. D'abord, comme nous l'avons relevé ci-dessus, les données de recherche ne sont pas des

TABLEAU 5

Récapitulatif des propriétés des données examminées

		Données de recherche	Données liées	Données ouvertes	Mégadonnées
Taille des ensembles		+ - +++	+ - +++	+ - ++	+++
Ouverture		Sous conditions	Habituellement	Par définition	Selon le détenteur des droits
Public cible principal		Communauté scientifique	Utilisateurs du Web	Citoyens Gouvernement	Chercheurs Décideurs
Enjeux dominants	Pratiques	Collecte Partage Qualité	Volume Qualité	Qualité	Volume Qualité
	Éthiques/juridiques	Confidentialité PI[112]	Confidentialité PI		
	Technologiques	Préservation Sécurité Description Organisation Stockage	Préservation Sécurité Encodage Stockage	Description Publication	Préservation Sécurité Analyse Stockage Traitements automatiques
	Épistémologiques	Démarche scientifique (par la réutilisation)	Statut ontologique des données		Démarche scientifique Statut ontologique des données Définition de la connaissance
	Économiques	Coûts de gestion		Bénéfices Coûts	
Traitements primordiaux		Gestion à long terme (*curation*)	Encodage *Mapping*	Documentation Diffusion	Analyse Visualisation
Technologies et outils principaux		PGD[113] Entrepôts	RDF URI *Triplestores*	Plateformes de diffusion Formats	Apprentissage automatique Intelligence artificielle Fouille
Rôles principaux du PI[114]		Gestion à long terme (*curation*)	Description Évaluation	Transcodage Formation	Description Organisation Gestion du cycle de vie

documents dans le sens communément admis en sciences de l'information et exigent de ce fait des compétences nouvelles, notamment dans « l'emploi des métadonnées, l'organisation intellectuelle de l'information et la préservation des documents numériques » (Guindon 2013, 190).

Pour les données de recherche, une autre différence réside dans l'insertion des bibliothécaires dans le processus de recherche plus en amont qu'à leur habitude : « Les bibliothécaires ont l'habitude d'exercer leur expertise en aval du processus de recherche, c'est-à-dire une fois les résultats publiés… pour ce qui est de la gestion des données, le travail commence dès le début du processus de recherche. » (Guindon 2013, 192)

Les professionnels de l'information peuvent avoir à leur tour un impact positif sur la pratique de gestion des données. Notamment, ils peuvent contribuer à garantir la pérennité d'accès aux données (voir Pouyllau 2013). Ils sont en mesure d'établir des balises et des outils pour permettre l'évaluation des données. Ils peuvent jouer un rôle de médiation entre les données et leur utilisateur, conformément à leurs pratiques habituelles aussi bien en bibliothèque qu'en centre d'archives. Enfin, les compétences spécifiques dans la rédaction de dictionnaires de données (nécessaire à la conception d'une base de données) sont directement mises à profit dans la définition et la documentation des métadonnées pertinentes pour une collection donnée.

Les professionnels de l'information ont apporté une contribution à l'organisation de l'information dans l'univers des données liées : notamment, Frederick (2016b, 11) souligne que l'expertise en contrôle d'autorité déployée dans VIAF (*Virtual Internet Authority File*[111]) a été déterminante pour désambiguïser des personnes, des lieux et des objets avec des noms identiques dans Wikipédia. Les données liées permettent des recoupements intéressants, dérivant des connaissances nouvelles imprévues à partir de données colligées au départ pour une autre application (voir notamment Frederick [2016b, 12] sur l'utilisation de données MARC pour une application en sociologie des sciences).

De plus, les archivistes devraient être mis à contribution davantage dans le processus de gestion des données, pour leur expertise non seulement en préservation, mais aussi en planification : leur pratique reconnaît la nécessité d'intégrer dès le départ les préoccupations liées à la description, à la conservation et à l'évaluation de la valeur pérenne des documents, ce qui peut se transposer tout naturellement pour le cas des données.

Le Tableau 5 récapitule les dimensions de l'analyse que nous avons présentée des différents types de données.

Les diverses informations recueillies et synthétisées dans le présent article aideront, nous l'espérons, à permettre aux professionnels de l'information de bonifier leur expertise de manière éclairée.

Conclusion

Les professionnels de l'information, dans leur pratique aujourd'hui, doivent considérer non seulement les documents et l'information, mais aussi différents types de données. Cela aura un impact également sur le développement de la discipline des sciences de l'information, sur la recherche qui est menée et sur la formation offerte.

Nous avons relevé quelles compétences des professionnels de l'information seront interpellées :

- Pour les données de recherche : des compétences surtout en organisation et en préservation, en plus des compétences proprement archivistiques liées à la planification dès leur conception de la gestion des données ;
- Pour les données ouvertes : des compétences en description des données, éventuellement dans l'élaboration de schémas de métadonnées, ainsi qu'en évaluation de l'information et en formation des utilisateurs ;
- Pour les données liées : des compétences en description (métadonnées) et en recherche d'information, qu'il faudra adapter à la façon dont se présentent les technologies du Web sémantique ;
- Pour les mégadonnées : des compétences en description et en préservation, semblables donc au cas des données de recherche ; il est probable de plus que les professionnels de l'information doivent acquérir davantage de connaissances et de compétences sur les environnements de stockage et de gestion des mégadonnées, différents de leurs outils habituels.

L'expertise propre aux professionnels de l'information qu'ils peuvent mettre de l'avant dans ces domaines peut être décrite de la façon suivante : « […] *librarians* can *exercise a range of sophisticated skills that occupy the central ground between understanding information user needs on one end and data curation on the other.* » (Stanton 2012) Ils sont particulièrement bien placés pour le faire.

Quelles actions prochaines pouvons-nous suggérer maintenant aux acteurs en sciences de l'information ? Nous proposons trois pistes.

1. Chercheurs : parmi les sujets à explorer, on devrait étudier davantage la convergence des pratiques en bibliothéconomie et en archivistique pour les données de recherche, l'impact du Web sémantique sur le Web des documents (et sur les collections

111. <viaf.org/>.

documentaires classiques), ainsi que l'apport des sciences de l'information dans la science des données.

2. Professionnels : ils devraient d'une part profiter d'occasions de formation continue et d'autre part agir en tant qu'ambassadeurs auprès de leurs collègues et de leurs milieux pour l'intégration de professionnels de l'information dans des équipes pluridisciplinaires : par exemple en contexte académique (données de recherche), dans le cadre d'initiatives d'ouverture des données ou encore dans l'adoption graduelle des données liées par les milieux documentaires.
3. Direction de services documentaires : il importe de maintenir une veille informationnelle sur ces nouveaux enjeux afin d'intégrer les technologies et les approches appropriées au moment opportun. En effet, toute présentation du type que celle présentée ci-dessus sera vite dépassée, en particulier pour les ressources disponibles.

À l'instar de Marchionini, on peut souhaiter que les sciences de l'information contribuent au développement de la science des données, ainsi qu'à la gestion et à l'exploitation des données en général.

SOURCES CONSULTÉES

Almeida, Maurício Barcellos, Renato Rocha Souza & Fred Fonseca. 2011. Semantics in the Semantic Web : A Critical Evaluation. *Knowledge Organization* 38 (3) : 187-203.

Amorim, Ricardo Carvalho, João Aguiar Castro, João Rocha da Silva & Cristina Ribeiro. 2015. A Comparative Study of Platforms for Research Data Management : Interoperability, Metadata Capabilities and Integration Potential. In *New Contributions in Information Systems and Technologies*. Advances in Intelligent Systems and Computing. Springer, Cham, 101-11. doi : 10.1007/978-3-319-16486-1_10

Anadiotis, George. 2017, mai. Graph Databases and RDF : it's a family affair. *ZDNet*. <www.zdnet.com/article/graph-databases-and-rdf-its-a-family-affair/>.

Anderson, Chris. 2008, 30 juin. The End of Theory. Will the Data Deluge Makes the Scientific Method Obsolete ? *Edge*. <www.edge.org/3rd_culture/anderson08/anderson08_index.html>.

Aquin, Mathieu d' & Enrico Motta. 2016, mai. The Epistemology of Intelligent Semantic Web Systems. *Synthesis Lectures on the Semantic Web : Theory and Technology* 6 (1). <www.morganclaypool.com/doi/pdf/10.2200/S00708ED1V01Y201603WBE014>.

Bachimont, Bruno. 2017. *Patrimoine et numérique : technique et politique de la mémoire*. Médias et humanités. Bry-sur-Marne, France : INA.

Ball, Adam & Monica Duke. 2015. *How to Cite Datasets and Link to Publications. DCC How-to Guides*. Edinburgh : Digital Curation Centre. <www.dcc.ac.uk/sites/default/files/documents/publications/reports/guides/How_to_Cite_Link.pdf>.

Ball, Alex. 2014. *How to License Research Data. Guide. A Digital Curation Centre and JISC Legal 'working level' guide*. UK : Digital Curation Centre. <www.dcc.ac.uk/sites/default/files/documents/publications/reports/guides/How_To_License_Research_Data.pdf>.

Barrière, Caroline. 2013. *Le Web sémantique : une ressource à découvrir*. <www.crim.ca/media/publication/texte_complet/wp-web-semantique-sept2013.pdf>.

Baškarada, Saša & Andy Koronios. 2017. Unicorn data scientist : The rarest of breeds. *Program* 51 (1) : 65-74. doi :10.1108/PROG-07-2016-0053

Bermès, Emmanuelle, Antoine Isaac & Gauthier Poupeau. 2013. *Le Web sémantique en bibliothèque*. Paris : Électre/Éditions du Cercle de la Librairie.

Bibliographic Framework Transition Initiative, Library of Congress. 2017, 21 juillet. BIBFRAME Frequently Asked Questions. <www.loc.gov/bibframe/faqs/#q09>.

Bieraugel, Mark. 2013, 19 juin. Keeping Up With... Big Data. Association of College & Research Libraries (ACRL). <www.ala.org/acrl/publications/keeping_up_with/big_data>.

Bizer, Christian, Peter Boncz, Michael L. Brodie & Orri Erling. 2012. The Meaningful Use of Big Data : Four Perspectives-Four Challenges. *SIGMOD Rec.* 40 (4) : 56–60. doi :10.1145/2094114.2094129

Bizer, Christian, Tom Heath & Tim Berners-Lee. 2011. Linked Data : The Story So Far. In *Semantic Services, Interoperability and Web Applications : Emerging Concepts : Emerging Concepts*, par Sheth Amit. IGI Global.

Borgman, Christine L. 2012. The Conundrum of Sharing Research Data. *Journal of the American Society for Information Science and Technology* 63 (6) : 1059-78. doi :10.1002/asi.22634

Borgman, Christine L. 2015. *Big Data, Little, Data, No Data. Scholarship in the Networked World*. Cambridge, MA : MIT Press.

Boyd, Danah & Kate Crawford. 2012. Critical Questions for Big Data. *Information, Communication & Society* 15 (5) : 662-79. doi :10.1080/1369118X.2012.678878

Braunschweig, Katrin, Julian Eberius, Mark Thiele & Wolfgang Lehner. 2012. *The State of Open Data - Limits of Current Open Data Platforms*. Communication présentée à Web Science Track at WWW' 12, Lyon, France.

Brown, David J. 2009. International Council for Scientific and Technical Information (ICSTI). *Annual Conference - Managing Data for Science* 29 (4) : 103-21.

Carroll, Michael W. 2015. Sharing Research Data and Intellectual Property Law : A Primer. *PLOS Biology* 13 (8) : e1002235. doi :10.1371/journal.pbio.1002235

Cervone, H. Frank. 2016. Informatics and Data Science : An Overview for the Information Professional. *Digital Library Perspectives* 32 (1) : 7-10.

Chabanne, Hervé, Pascal Urien & Jean-Ferdinand Susini. 2013. *RFID and the Internet of Things*. ISTE. London : Wiley. <onlinelibrary.wiley.com/book/10.1002/9781118614297>.

Charlet, Jean, Philippe Laublet & Chantal Reynaud. 2005. *Le Web sémantique*. Toulouse : Cépaduès-Editions.

Chen, C.L. Philip & Chun-Yang Zhang. 2014. Data-intensive applications, challenges, techniques and technologies : A survey on Big Data-ScienceDirect. *Information Sciences* 275 : 314-47.

Comité OGGO. 2014. *Les données ouvertes : la voie du futur. Rapport du comité permanent des opérations gouvernementales et des prévisions budgétaires*. Rapport d'un comité parlementaire (Canada) No 5-OGGO (41-2). Ottawa : Chambre des communes du Canada. <www.noscommunes.ca/Content/Committee/412/OGGO/Reports/RP6670517/oggorp05/oggorp05-f.pdf>.

Corrado, Edward M. & Heather Moulaison Sandy. 2017. *Digital Preservation for Libraries, Archives, and Museums*. Vol. Second Edition. Lanham : Rowman & Littlefield Publishers. <search.ebscohost.com/login.aspx ?direct=true&db=nlebk&AN=1442939&lang=fr&site=ehost-live>.

Cragin, Melissa H., Carole L. Palmer, Jacob R. Carlson & Michael Witt. 2010. Data Sharing, Small Science and Institutional Repositories. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences* 368 (1926) : 4023-38. doi :10.1098/rsta.2010.0165

Crastes, Matthieu. 2015. Dessine-moi mon métier ! *I2D – Information, données & documents* me 52 (2) : 4-6.

Crosas, Mercè. 2011. The Dataverse Network® : An Open-Source Application for Sharing, Discovering and Preserving Data. *D-Lib Magazine* 17 (1/2). doi : 10.1045/january2011-crosas

Da Sylva, Lyne. 2017, 11 août. *Vers les données liées : conséquences théoriques et pratiques pour les sciences humaines*. Communication présentée à Digital Humanities 2017, Montréal. <www.conftool.pro/dh2017/sessions.php>.

Deloitte LLP. 2012. Open data Driving growth, ingenuity and innovation. London, UK. <www.mendeley.com/research-papers/open-data-driving-growth-ingenuity-innovation>.

Dickner, Nicolas. 2017. *Comprendre et manipuler les données ouvertes de l'administration publique. La situation au Gouvernement du Québec et à la Ville de Montréal*. (Mémoire de maîtrise, Université de Montréal, Montréal).

Dietze, Stefan, Salvador Sanchez-Alonso, Hannes Ebner, Hong Qing Yu, Daniela Giordano, Ivana Marenzi & Bernardo Pereira Nunes. 2013. Interlinking Educational Resources and the Web of Data : A Survey of Challenges and Approaches. *Program : Electronic Library and Information Systems* 47 (1) : 60-91. doi : 10.1108/00330331211296312

Digital Curation Centre. 2013. *Checklist for a Data Management Plan*. Edinburgh : Digital Curation Centre. <www.dcc.ac.uk/sites/default/files/documents/resource/DMP/DMP_Checklist_2013.pdf>.

Draetta, Laura. 2012. *RFID, une technologie controversée : ethnographie de la construction sociale du risque*. Collection Mondialisation, hommes et sociétés. Cachan : Hermès science publications-Lavoisier.

Erway, Ricky, Laurence Horton, Amy Nurnberger, Reid Otsuji & Amy Rushing. 2016. *Building Blocks : Laying the Foundation for a Research Data Management Program*. Dublin, Ohio : OCLC Research. <www.oclc.org/content/dam/research/publications/2016/oclcresearch-data-management-building-blocks-2016.pdf>.

Federer, Lisa. 2016. Research data management in the age of big data : Roles and opportunities for librarians. *Information Services & Use* 36 : 35-43. doi : 10.3233/ISU-160797.

Frederick, Donna Ellen. 2016a. Data, Open Science and Libraries - The Data Deluge Column. *Library Hi Tech News* 33 (8) : 11-16.

Frederick, Donna Ellen. 2016b. Libraries, data and the fourth industrial revolution – The Data Deluge Column. *Library Hi Tech News* 33 (5) : 9-12.

Frické, Martin. 2015. Big Data and Its Epistemology. *Journal of the Association for Information Science and Technology* 66 (4) : 651-61. doi : 10.1002/asi.23212

Gandomi, Amir & Murtaza Haider. 2015. Beyond the Hype : Big Data Concepts, Methods, and Analytics. *International Journal of Information Management* 35 (2) : 137-44. doi : 10.1016/j.ijinfomgt.2014.10.007

Gracy, Karen F. 2015. Archival Description and Linked Data : A Preliminary Study of Opportunities and Implementation Challenges. *Archival Science* 15 (3) : 239-94. doi : 10.1007/s10502-014-9216-2

Gruber, Tom. 1993. A translation approach to portable ontologies. *Knowledge Acquisition* 5 (2) : 199-220.

Guindon, Alex. 2013. La gestion des données de recherche en bibliothèque universitaire. *Documentation et bibliothèques* 59 (4) : 189-200. doi : 10.7202/1019216ar

Halpin, Harry, Ivan Herman & Patrick J. Hayes. 2010. When owl :sameAs isn't the Same : An Analysis of Identity Links on the Semantic Web. In *RDF Next Steps Workshop*. Palo Alto, CA. <www.w3.org/2009/12/rdf-ws/papers/ws21>.

Hannemann, J. & J. Kett. 2010, août. *Linked Data for libraries*. Communication présentée au World Library and Information Congress : 76th IFLA General Conference and Assembly, Gothenburg, Suède. <conference.ifla.org/past-wlic/2010/149-hannemann-en.pdf>.

Harper, Richard. 2011. *The Connected Home : The Future of Domestic Life*. London ; New York : Springer. doi : 10.1007/978-0-85729-476-0

Harth, Andreas, Katja Hose & Ralf Schenkel. 2016. *Linked Data Management*. Boca Raton, FL : CRC Press, Taylor & Francis Group. <lib.myilibrary.com/ProductDetail.aspx ?id=621856>.

Heery, Rachel. 2004. Metadata Futures : Steps Toward Semantic Interoperability. In *Metadata in Practice*, par D.I. Hillman & E.L. Westbrooks. Chicago : American Library Association, 257-71.

Heidorn, P. Bryan. 2011. The Emerging Role of Libraries in Data Curation and E-science. *Journal of Library Administration* 51 (7-8) : 662-72. doi : 10.1080/01930826.2011.601269

Hey, Tony, Stewart Tansley & Kristin Tolle. 2009. *The Fourth Paradigm : Data-Intensive Scientific Discovery*. Redmond, VA : Microsoft Research.

Higgins, Sarah. 2008. The DCC Curation Lifecycle Model. *International Journal of Digital Curation* 3 (1) : 134-40. doi : 10.2218/ijdc.v3i1.48

Hilbert, Martin. 2016. Big Data for Development : A Review of Promises and Challenges. *Development Policy Review* 34 (1) : 135-74. doi : 10.1111/dpr.12142

Hitzler, Pascal. 2010. *Foundations of Semantic Web Technologies*. Chapman & Hall/CRC Textbooks in Computing. Boca Raton : CRC Press.

Holdren, John P., Peter Orszag & Paul F. Prouty. 2009. *Memorandum for Heads of Departments and Agencies*. Mémorandum. Executive Office of the President. <www.whitehouse.gov/sites/whitehouse.gov/files/omb/memoranda/2009/m09-12.pdf>.

Hooland, Seth van & Ruben Verborgh. 2014. *Linked Data for Libraries, Archives and Museums : How to Clean, Link and Publish Your Metadata*. U.S. Edition. Chicago : Neal-Schuman, an imprint of the American Library Association.

Huwe, Terence K. 2017, 1 mai. Librarians and Data : Curator, Creator, or Both ? <www.highbeam.com/doc/1P4-1907273753.html>.

Ibekwe-Sanjuan, Fidelia & Geoffrey Bowker. 2017. Implications of big data for knowledge organization. *Knowledge Organization*, Special issue on New trends for Knowledge Organization, Renato Rocha Souza (guest editor), 44 (3) : 187-98.

IFLA Study Group on the Functional Requirements for Bibliographic Records. 2009. *Functional Requirements for Bibliographic Records*. Munich : International Federation of Library Associations. <www.ifla.org/files/assets/cataloguing/frbr/frbr_2008.pdf>.

Interagency Working Group on Digital Data. 2009. *Harnessing_Power_Web*. <www.nitrd.gov/About/Harnessing_Power_Web.pdf>.

International Organization for Standardization. 2004. ISO/IEC 21000-5 :2004. Information technology – Multimedia frame- work (MPEG-21) - Part 5 : Rights Expression Language.

Inter-university Consortium for Political and Social Research (ICPSR). 2012. *Guide to Social Science Data Preparation and Archiving : Best Practice Throughout the Data Life Cycle* (5e éd.). Ann Arbor, MI : ICPSR.

Irwin, Sarah. 2013. Qualitative secondary data analysis : Ethics, epistemology and context. *Progress in Development Studies* 13 (4) : 295-306.

Jacobs, James A. & Charles Humphrey. 2004. Preserving Research Data. *Communications of the ACM* 47 (9) : 27. doi : 10.1145/1015864.1015881

Janssen, Marijn, Yannis Charalabidis & Anneke Zuiderwijk. 2012. Benefits, Adoption Barriers and Myths of Open Data and Open Government. *Information Systems Management* 29 (4) : 258-68. doi : 10.1080/10580530.2012.716740

Janssen, Marijn & George Kuk. 2016. Big and Open Linked Data (BOLD) in research, policy, and practice. *Journal of Organizational Computing and Electronic Commerce* 26 (1-2) : 3-13. doi : 10.1080/10919392.2015.1124005

Jeuland, François-Xavier. 2005. *La maison communicante*. Paris : Eyrolles.

Kim, Won, Ok-Ran Jeong & Chulyun Kim. 2014. A Holistic View of Big Data. *International Journal of Data Warehousing and Mining* 10 (3) : 59-69. doi : 10.4018/ijdwm.2014070104

Kitchin, Rob. 2014a. *The Data Revolution : Big Data, Open Data, Data Infrastructures and Their Consequences*. SAGE Publications.

Kitchin, Rob. 2014b. Big Data, New Epistemologies and Paradigm Shifts. *Big Data & Society* 1 (1) : 2053951714528481. doi : 10.1177/2053951714528481

Klapwijk, Wouter & IFLA Big Data Special Interest Group. 2016, 15 juin. *The Library (Big) Data scientist*. Webinaire présenté à IFLA/ALA Webinar. <npsig.files.wordpress.com/2016/04/bd-sig-wouter-klapwijk.pdf>.

Koltay, Tibor. 2014. Research data and libraries. *Konyvtari Figyelo* 24 (2) : 223-35.

Koltay, Tibor. 2017. Data Literacy for Researchers and Data Librarians. *Journal of Librarianship and Information Science* 49 (1) : 3-14. doi : 10.1177/0961000615616450

Labrinidis, Alexandros & H. V. Jagadish. 2012. Challenges and Opportunities with Big Data. *Proc. VLDB Endow.* 5 (12) : 2032-2033. doi : 10.14778/2367502.2367572

Leonelli, Sabina. 2014. What Difference Does Quantity Make ? On the Epistemology of Big Data in Biology. *Big Data & Society* 1 (1) : 1-10. doi : 10.1177/2053951714534395

Library of Congress. 2012. *Bibliographic Framework as a Web of Data : Linked Data Model and Supporting Services*. Washington, D.C. : Library of Congress. <www.loc.gov/bibframe/pdf/marcld-report-11-21-2012.pdf>.

Library of Congress. 2016. Overview of the BIBFRAME 2.0 Model. <www.loc.gov/bibframe/docs/bibframe2-model.html>.

Luce, Richard. 2008. A New Value Equation Challenge : The Emergence of eResearch and Roles for Research Libraries - Council on Library and Information Resources. In *No Brief Candle : Reconceiving Research Libraries for the 21st Century*. Washington, D.C. : Council on Library and Information Resources (CLIR), 42-50. <www.clir.org/pubs/reports/pub142/luce.html>.

Lucic, Ana & Catherine Blake. 2016. Preparing a Workforce to Effectively Reuse Data. In *Proceedings of the 79th Meeting of the Association for Information Science and Technology*. Copenhagen : ASIS&T, 1-10. <www.asist.org/files/meetings/am16/proceedings/openpage16.html>.

Makhlouf Shabou, Basma. 2015. Fonction d'évaluation des archives : bilan sommaire des développements, des enjeux actuels et des défis futurs. In *Panorama de l'archivistique contemporaine : évolution de la discipline et de la profession*, par Louise Gagnon-Arguin et Marcel Lajeunesse. Québec : Presses de l'Université du Québec.

Marchionini, Gary. 2016. Information Science Roles in the Emerging Field of Data Science. *Journal of Data and Information Science* 1 (2) : 1-6.

Margolis, Ronald, Leslie Derr, Michelle Dunn, Michael Huerta, Jennie Larkin, Jerry Sheehan, Mark Guyer & Eric D. Green. 2014. The National Institutes of Health's Big Data to Knowledge (BD2K) initiative : capitalizing on biomedical big data. *Journal of the American Medical Informatics Association* 21 (6) : 957-58. doi : 10.1136/amiajnl-2014-002974

Marr, Bernard. 2014, 6 mars. Big Data : The 5 Vs Everyone Must Know. LinkedIn Pulse. <www.linkedin.com/pulse/20140306073407-64875646-big-data-the-5-vs-everyone-must-know>.

Marx, Vivien. 2013. Biology : The Big Challenges of Big Data. *Nature* 498 (7453) : 255-60. doi : 10.1038/498255a

Maurel, Lionel. 2012. Du web de documents au web de données : la révolution juridique inachevée de l'*Open Data*. In *Le document numérique à l'heure du web*, édité par Lisette Calderan, Pascale Laurent, Hélène Lowinger & Jacques Millet. ADBS, 155-71. <hal.inria.fr/hal-00843783>.

Mauthner, Natasha Susan & Odette Parry. 2013. Open Access Digital Data Sharing : Principles, Policies and Practices. *Social Epistemology* 27 (1) : 47-67. doi : 10.1080/02691728.2012.760663

Mayernik, Matthew S. 2016. Research Data and Metadata Curation as Institutional Issues. *Journal of the Association for Information Science and Technology* 67 (4) : 973-93. doi : 10.1002/asi.23425

Mayernik, Matthew S., Jennifer Phillips & Eric Nienhouse. 2016. Linking Publications and Data : Challenges, Trends, and Opportunities. *D-Lib Magazine* 22 (5/6). doi : 10.1045/may2016-mayernik

McDonald, John & Valérie Léveillé. 2014. Whither the retention schedule in the era of big data and open data ? *Records Management Journal* 24 (2) : 99-121. doi : 10.1108/RMJ-01-2014-0010

McGeever, Mags. 2007. IPR in Databases | Digital Curation Centre. UK : Digital Curation Centre. <www.dcc.ac.uk/resources/briefing-papers/legal-watch-papers/ipr-databases>.

Mercier, Silvère. 2011. Open data et bibliothèques. *Documentaliste-Sciences de l'Information* 48 (3) : 8-13.

Meyer, Eric & Ralph Schroeder. 2014. *Digital Transformations of Research*. Cambridge, MA : MIT Press.

National Academy of Sciences, National Academy of Engineering & Institute of Medicine. 2009. *Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age*. Washington, D.C. : The National Academies Press. doi : 10.17226/12615

Nunberg, Geoffrey. 2008, 31 août. Google's Book Search : A Disaster for Scholars. The Chronicle of Higher Education. <chronicle.com/article/Googles-Book-Search-A/48245/>.

Obama, Barack. 2013. Making Open and Machine Readable the New Default for Government Information. *Federal Register* 78 (93) : 28111-13.

Office de la langue française. 2014. Mise en données. In *Le grand dictionnaire terminologique*. <granddictionnaire.com/ficheOqlf.aspx?Id_Fiche=26523020>.

Office de la langue française. 2015a. Internet des objets. In *Le grand dictionnaire terminologique*. <granddictionnaire.com/ficheOqlf.aspx?Id_Fiche=26529845>.

Office de la langue française. 2015b. Mégadonnées. In *Le grand dictionnaire terminologique*. <granddictionnaire.com/ficheOqlf.aspx?Id_Fiche=26507313>.

Open Government Working Group. 2007. 8 Principles of Open Government Data. <public.resource.org/8_principles.html>.

Parry, Odette & Natasha S. Mauthner. 2004. Whose Data Are They Anyway ? Practical, Legal and Ethical Issues in Archiving Qualitative Research Data. *Sociology* 38 (1) : 139-52. doi : 10.1177/0038038504039366

Perrier, Laure *et al.* 2017. Research data management in academic institutions : A scoping review. *PLOS ONE* 12 (5) : e0178261. doi : 10.1371/journal.pone.0178261

Peugeot, Valérie. 2014. Données publiques ouvertes : or du 21e siècle ou biens communs ? *Documentaliste-Sciences de l'Information* 50 (4) : 48-63.

Peyrard, Sébastien & Agnès Simon. 2014. Le web sémantique en bibliothèque. *Bulletin de bibliothèques de France* (2). <bbf.enssib.fr/consulter/bbf-2014-02-0189-007>.

Pouyllau, Stéphane. 2013. Web de données, Big Data, Open Data, Quels Rôles pour les Documentalistes ? *Documentaliste* 50 (3) : 32.

Powers, Shelley. 2003. *Practical RDF*. 1st Ed. Sebastopol, CA : Farnham : O'Reilly.

Ray, Joyce M. 2014. *Research data management : practical strategies for information professionals*. West Lafayette, Indiana : Purdue University Press.

Reichman, O. J., Matthew B. Jones & Mark P. Schildhauer. 2011. Challenges and Opportunities of Open Data in Ecology. *Science* 331 (6018) : 703-5. doi : 10.1126/science.1197962

Rietveld, Laurens. 2016. *Publishing and Consuming Linked Data : Optimizing for the Unknown*. IOS Press.

Riley, Jenn. 2017. *Understanding metadata*. Baltimore, MD : National Information Standards Organization. <www.niso.org/apps/group_public/download.php/17446/understanding%20metadata>.

Rougier, Nicolas P. *et al*. 2017, juillet. *Sustainable computational science : the ReScience initiative*. <arxiv.org/pdf/1707.04393.pdf>.

Rousidis, Dimitris, Emmanouel Garoufallou, Panos Balatsoukas & Miguel-Angel Sicilia. 2014. Metadata for Big Data : A Preliminary Investigation of Metadata Quality Issues in Research Data Repositories. *Information Services & Use* 34 (3-4) : 279-86.

Rousseau, Jean-Yves & Carol Couture. 1994. *Les fondements de la discipline archivistique*. Québec : Presses de l'Université du Québec.

Rowley, Jennifer. 2007. The Wisdom Hierarchy : Representations of the DIKW Hierarchy. *Journal of Information Science* 33 (2) : 163-80. doi : 10.1177/0165551506070706

Silver, Nate. 2013, 23 août. What I need from statisticians. Statistics Views. <www.statisticsviews.com/details/feature/5133141/Nate-Silver-What-I-need-from-statisticians.html>.

Stanton, Jeffrey M. 2012, juillet. Data Science : What's in it for the New Librarian ? - Information Space. Infospace - The Official Blog of the Syracuse University iSchool. <ischool.syr.edu/infospace/2012/07/16/data-science-whats-in-it-for-the-new-librarian/>.

St-Germain, Marielle. 2017. *Le Web de données et le Web sémantique à Bibliothèque et Archives nationales du Québec : constats et recommandations fondés sur l'initiative de la Bibliothèque nationale de France*. (Mémoire de maîtrise, Université de Montréal, Montréal). <papyrus.bib.umontreal.ca/xmlui/handle/1866/18414>.

Strasser, Carly. 2015. *NISO Primer : Research Data Management*. Baltimore, MD : National Information Standards Organization. <www.niso.org/apps/group_public/download.php/15375/Primer-RDM-2015-0727.pdf>.

Stuart, David. 2011. *Facilitating access to the Web of Data : A guide for librarians*. London : Facet.

Swan, Alma & Sheridan Brown. 2008. *Skills, Role & Career Structure of Data Scientists & Curators : Assessment of Current Practice & Future Needs*. Rapport pour JISC. Truro, UK. <www.webarchive.org.uk/wayback/archive/20140615053226/http://www.jisc.ac.uk/media/documents/programmes/digitalrepositories/dataskillscareersfinalreport.pdf>.

Swan, Melanie. 2013. The Quantified Self : Fundamental Disruption in Big Data Science and Biological Discovery. *Big Data* 1 (2) : 85-99. doi : 10.1089/big.2012.0002

Tenopir, Carol *et al.* 2015. Research Data Services in Academic Libraries : Data Intensive Roles for the Future ? *Journal of eScience Librarianship* 4 (2) : 1-21. doi : 10.7191/jeslib.2015.1085

The Linux Information Project. 2017. What is a free file format ? definition by The Linux Information Project (LINFO). Consulté le 19 juillet. <www.linfo.org/free_file_format.html>.

Tole, Alexandru Adrian. 2013. Big Data Challenges. *Database Systems Journal* 4 (3) : 31-40.

Vatant, Bernard. 2008. Des métadonnées à la description des ressources. Les langages du web sémantique. In *Métadonnées : mutations et perspectives : séminaire INRIA, 29 septembre-3 octobre 2008*, par Lisette Calderan, Bernard Hidoine & Jacques Millet. Paris : ADBS Éditions, 163-94.

Venot, Alain, Anita Burgun & Catherine Quantin (éd). 2013. *Informatique médicale, e-Santé - Fondements et applications* | Alain Venot | Springer. Paris : Springer-Verlag. <www.springer.com/cn/book/9782817803371>.

W3C. 2004. Initiation à RDF. <www.yoyodesign.org/doc/w3c/rdf-primer/>.

Weller, Travis & Amalia Monroe-Gulick. 2014. Understanding Methodological and Disciplinary Differences in the Data Practices of Academic Researchers. *Library Hi Tech* 32 (3) : 467-82.

Zetterlund, Boris. 2016, 20 juin. Big Data and Libraries : Getting the most from your library data. Axiell UK - helping you create the library of the future. <www.axiell.co.uk/getting-the-most-from-your-library-data/>.

Zins, Chaim. 2007. Conceptual Approaches for Defining Data, Information, and Knowledge. *Journal of the American Society for Information Science and Technology* 58 (4) : 479-93. doi : 10.1002/asi.20508

Zuiderwijk, Anneke, Marijn Janssen, Sunil Choenni, Ronald Meijer & Roexsana Sheikh Alibaks. 2012. Socio-technical Impediments of Open Data. *Electronic Journal of e-Government* 10 (2) : 156-72.