

LES DONNÉES ET LES SCIENCES DE L'INFORMATION



Volume 63, n° 4 | OCTOBRE – DÉCEMBRE 2017

DOCUMENTATION  BIBLIOTHÈQUES

Volume 63, n° 4
OCTOBRE – DÉCEMBRE 2017

Documentation et bibliothèques
2065, rue Parthenais, bureau 387
Montréal (Québec) H2K 3T1
Téléphone : 514 281-5012
Télécopieur : 514 281-8219

Abonnement

www.asted.org
info@asted.org

Direction

Réjean Savard

Coordination à la rédaction

Charlotte de Celles

Comité de rédaction

Jean-Philippe Accart · Gaston Bernier
Denis Boisvert · Bernard Dione · Dominique
Gazo · Alex Guindon · Cristina Ion · Stéphane
Labbé · Dominique Lahary · Patrice Landry
Élisabeth Lavigueur · Sabine Mas
Heather Lea Moulaison

Graphisme

Marquis Interscript

Photo couverture

Bibliothèque du cégep de Limoilou
(campus de Charlesbourg) (Ville de Québec)

Révision et correction d'épreuves

Charlotte de Celles

Traduction

Charlotte de Celles

Traduction des résumés

Edward Collister

Publicité

Lionel Villalonga
lvillalonga@asted.org

Production

ASTED
En partenariat avec
AIFBD

Impression

Marquis Imprimeur

Les articles de *Documentation et bibliothèques* sont accessibles en version numérique sur Érudit (documentation.erudit.org).

Ils sont répertoriés dans les bases de données FRANCIS de l'Institut de l'information scientifique et technique (INIST), Canadian Periodical Index – Index de périodiques canadiens, Library and Information Science Abstracts (LISA), Library Literature & Information Science, Repère. Les articles plus récents sont aussi référencés dans Google Scholar, Primo Central, Summon, WorldCat, JournalTOCs et EBSCO.

Nous reconnaissons l'appui du gouvernement du Canada par l'entremise du Programme d'aide aux publications pour nos dépenses d'envoi postal.

Dépôt légal

Bibliothèque et Archives nationales
du Québec
Bibliothèque et Archives Canada

Imprimé au Canada

érudit
Canada

ISSN-0315-2340 | Envoi de Poste-Publications n° 40005041
Canada : **35,00 \$**, Étranger : **45,00 \$ CAN**



DOCUMENTATION BIBLIOTHÈQUES

SOMMAIRE

Éditorial	4
Réjean Savard	
Les données et leurs impacts théoriques et pratiques sur les professionnels de l'information	5
Lyne Da Sylva	
Étapes pour le développement d'un projet de données ouvertes et liées en bibliothèque	35
Marielle St-Germain	
Technologies d'exploitation du <i>big data</i> dans les organisations et transformations organisationnelles: une étude de cas au sein du Service de santé des armées françaises	46
Marc Tanti	
Résultats d'une enquête sur les pratiques et attitudes des chercheurs de l'Université Concordia en matière de gestion des données de recherche ...	59
Danielle Dennie et Alex Guindon	
Valoriser les données d'enquêtes qualitatives en sciences sociales: le cas français de la banque d'enquête beQuali	73
Selma Bendjaballah, Sarah Cadorel, Émilie Fromont, Guillaume Garcia, Émilie Groshens et Émeline Juillard	
Présentation des auteurs	86

LES DONNÉES

Réjean SAVARD

Directeur
Documentation et bibliothèques
rejean.savard@umontreal.ca

La question des données en sciences de l'information a pris une importance accrue ces dernières années. Ce phénomène extrêmement lié au développement des technologies se décline sous de nombreuses facettes : données ouvertes, *big data*, métadonnées, données liées, données de recherche, fouille de textes (*text mining*), etc. Il connaît un impact certain dans le domaine des bibliothèques, des archives et de la documentation, allant même jusqu'à transformer la société.

Documentation et bibliothèques a voulu faire le point sur ce sujet. Vous trouverez dans ce numéro des articles essentiels pour comprendre la problématique des données et le lien à faire avec les bibliothèques et les autres services d'information. On trouvera aussi des études de cas qui viendront démontrer l'application des concepts présentés en début de numéro.

Bonne lecture !

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.

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.

[...] 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.

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 logiciels¹. 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, structurales 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 quotidienne² » 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 Research³) collige depuis 1962 des données provenant de projets de recherche en politique et en sciences sociales ; leur collection dépasse les 250 000 fichiers⁴.

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 NASA⁵ et sur le site du Canadian Cryospheric Information Network⁶) et les données générées par le grand collisionneur de hadrons (LHC) du CERN⁷, 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.

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

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

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

6. <www.polardata.ca/>.

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

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

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]).

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 subventionnaires⁹ » est également en cours d'élaboration. Aux États-Unis, la National Science Foundation exige depuis 2011 qu'un plan de gestion des données¹⁰ 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.

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.

Données ouvertes

Le terme « données ouvertes » (Dickner 2017; Comité OGGO 2014; Peugeot 2014; Janssen, Charalabidis & Zuijderwijk 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 britannique¹¹, américain¹² (voir notamment [Holdren, Orszag & Prouty 2009]), canadien¹³ et autres. Également, des organismes comme les Nations Unies publient de telles données¹⁴.

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, Orszag & 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 Longueuil¹⁵ et le service de bicyclette en autopartage Bixi à Montréal¹⁶, 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 Consortium¹⁷ (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.

TABEAU 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 DBPedia¹⁹ : 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).

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 terme²⁰. 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 Data²¹ 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 disposition 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.

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

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.

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.

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.

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.

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 donnée²² 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 étoiles²⁵ 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 Foundation²⁶ et l'Open Data Network²⁷ 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/>.

Handbook²⁸, le site Web de la Open Knowledge Foundation²⁹ et le portail d'Open Knowledge International³⁰ sont des ressources d'intérêt général. Un logiciel comme OpenRefine³¹ permet de nettoyer et transformer des données avant de les rendre disponibles. Un autre, LODRefine³², 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 GéoGratis³⁶ (données géospatiales du ministère des Ressources naturelles du Canada) et les données polaires hébergées sur le site du Canadian Cryospheric Information Network³⁷ (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 schèmes³⁹) ; 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'IETF 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 pré-existants (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 « résolubles »). 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](http://isni.org/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 Friend⁴¹) 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).

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 (SPARQL⁴³, 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 diagram⁴⁴. Le portail des données ouvertes de l'Union européenne⁴⁵ 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 France⁴⁶, les données de WorldCat⁴⁷ de l'OCLC, le thésaurus AAT sur l'art et l'architecture du Getty Research Institute⁴⁸, le plan de classification IconClass pour les œuvres d'art⁴⁹, ainsi que des notices d'autorité et divers vocabulaires de la Library of Congress⁵⁰ tels que son répertoire de vedettes-matière.

Divers «vocabulaires⁵¹» 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 Vocabularies⁵² de la Open Knowledge Foundation présente une collection de ces vocabulaires.

Plus spécifiquement pour les bibliothèques, le regroupement Linked Data for Libraries⁵³ 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 Persée⁵⁴ fournit l'inter-

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.

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 Sparklis⁵⁵, 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 RDE, 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 Initiative⁵⁶ 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 préservation⁵⁷.»

Les organismes subventionnaires de nombreux pays dont le Royaume-Uni⁵⁸ 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 citation⁵⁹. 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 Archive⁶⁰ au Royaume-Uni regroupant les données en sciences humaines et sociales) ou encore disciplinaires (par exemple, DATAONE⁶¹ 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 DataCite⁶², qui est utilisé notamment pour citer cette ressource.
- Documentation, guides, etc. : plusieurs documents sont disponibles auprès du Digital Curation Centre, au Royaume-Uni⁶³. 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 recherche⁶⁴ « 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'experts⁶⁵ : 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 Canadien⁶⁶ ». Il s'agit d'une « organisation dirigée et appuyée par des intervenants déterminés à améliorer la gestion des données scientifiques au Canada⁶⁷ ». 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 recherche⁶⁸ (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 québécois⁶⁹ ». La Journée de réflexion sur la gestion des données de recherche⁷⁰ (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 PGD⁷¹ (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.cihir-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. <oneresearch.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 Dataverse⁷³. 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 données⁷⁴. 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.

Une plateforme nationale fédérant les dépôts de données de recherche canadiens⁷⁵ est en cours de préparation ; son lancement est prévu pour janvier 2018 et une version bêta est disponible⁷⁶. 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

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

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 ges-

TABEAU 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/>.

tion 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

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 Analytics⁷⁷ à l'Université Dalhousie et IVADO (Institut de valorisation des données, pôle scientifique et économique⁷⁸), 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 numérique⁷⁹, 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 données⁸⁰ ».

Par ailleurs, les chercheurs canadiens peuvent compter sur CANARIE⁸¹, 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 Canada⁸² 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 catalogue) 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 MODS⁸³, tous deux particulièrement utiles pour décrire des documents, certains schémas de métadonnées spécialisés 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 Attribution Licence*⁸⁹ : 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.ddalliance.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 : Fosse 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 Canada⁹⁴; ASIS&T⁹⁵; 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 recherche⁹⁶. Enfin, voici quelques exemples de vidéos en ligne :

- « Le Web des données ouvertes et liées. Qu'est-ce que c'est ? » par EuropeanaEU⁹⁷

- « RDC Webinar : What is RDC⁹⁸ ? » et « RDC Webinar : Persistent IDs : Best Practices⁹⁹ » par CANARIE Inc.
- Chaînes YouTube : Le Réseau canadien des centres de données de recherche¹⁰⁰ (RCCDR), l'ABRC¹⁰¹

La vidéo « Introduction à la gestion des données de recherche » sur le site du service des bibliothèques de l'Université de Montréal¹⁰² 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_2YI-OIM>.

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-Neige¹⁰³ 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. <info-neige.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.

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 AGROVOC¹⁰⁶ et le thésaurus de l'UNESCO¹⁰⁷. 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 1998¹¹⁰. 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.

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).

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.

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 examinées

		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	Volume	Qualité	Volume
		Partage	Qualité		Qualité
		Qualité			
	Éthiques/juridiques	Confidentialité PI ¹¹²	Confidentialité PI		
		Technologiques	Préservation	Préservation	Description
	Sécurité		Sécurité	Publication	Sécurité
	Description		Encodage		Analyse
	Organisation		Stockage		Stockage
	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 (<i>curation</i>)	Encodage <i>Mapping</i>	Documentation Diffusion	Analyse Visualisation
Technologies et outils principaux		PGD ¹¹³ Entrepôts	RDF URI <i>Triplestores</i>	Plateformes de diffusion Formats	Apprentissage automatique Intelligence artificielle Fouille
Rôles principaux du PI ¹¹⁴		Gestion à long terme (<i>curation</i>)	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-sémantique-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-Éditions.

- 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.ijinfo-mgt.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.mylibrary.com/ProductDetail.aspx?id=621856>.
- Heery, Rachel. 2004. Metadata Futures: Steps Toward Semantic Interoperability. In *Metadata in Practice*, par D.I. Hillman & E.L. Westbrook. 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*. Memorandum. 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 framework (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/ijdw.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/marclid-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/open-page16.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
- Mayermik, 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
- Mayermik, 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, Silvére. 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 21^e 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*. 1^{re} 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 & Roeksana Sheikh Alibaks. 2012. Socio-technical Impediments of Open Data. *Electronic Journal of e-Government* 10 (2): 156-72.

ÉTAPES POUR LE DÉVELOPPEMENT D'UN PROJET DE DONNÉES OUVERTES ET LIÉES EN BIBLIOTHÈQUE

Marielle St-Germain

Étudiante au doctorat
École de bibliothéconomie et des sciences de l'information (EBSI)
Université de Montréal
marielle.st-germain@umontreal.ca

RÉSUMÉ | ABSTRACT

On constate depuis quelques années un intérêt de plus en plus marqué pour le Web sémantique et les avantages qu'il présente pour les institutions documentaires. Cependant, un flou persiste quant au potentiel, au fonctionnement et aux étapes de développement liées à des projets qui mettent en œuvre les différents standards et composantes qui le caractérisent. L'objectif du présent article est de proposer une liste d'étapes pour la mise en œuvre d'un projet de données ouvertes et liées (Web de données) en bibliothèque afin de faciliter l'appropriation de ces technologies par les professionnels de l'information.

The Stages in the Development of an Open Data Project Linked to the Library

A marked interest in the Semantic Web and its advantages for libraries has been observed in recent years. However, the potential, the operations and the stages in the development of such projects using different standards and components remains nebulous. This article suggests a list of stages in the implementation of open and linked data (Web data) in the library in order to facilitate the appropriation of these technologies by information professionals.

S'adapter aux évolutions technologiques: une réalité bien connue des professionnels de l'information

L'arrivée du Web, et du numérique par le fait même, a non seulement modifié les comportements informationnels, mais aussi notre façon de consommer et de traiter les connaissances. De plus, la multiplication des formats de documents a eu pour conséquence directe la prolifération des données les décrivant, nécessitant de nouvelles compétences et connaissances de la part des professionnels de l'information (Stuart 2011). On a aussi vu la perception du rôle de bibliothécaire changer: les tâches ont été redéfinies au fil des ans pour laisser plus de place à l'aide à la recherche ainsi qu'à une collaboration étroite entre professionnels de l'information et informaticiens. Les qualités reconnues aux professionnels de l'information telles que le fait d'accorder une importance particulière aux besoins des usagers et au partage de connaissances, la capacité d'identifier et d'organiser des ressources physiques et numériques ainsi que la connaissance des concepts relatifs à la gestion leur ont

permis de saisir rapidement les avantages découlant de l'arrivée du Web traditionnel (Rao & Babu 2001; Stuart 2011). Malgré les différentes évolutions technologiques auxquelles elles ont dû faire face, la raison d'être et les missions des bibliothèques restent les mêmes: offrir l'accès à une collection de documents pour leur communauté, acquérir des ressources ainsi que produire des services et, finalement, agir comme des intermédiaires entre l'utilisateur et les ressources (Leroux *et al.* 2009). On comprend ainsi aujourd'hui l'importance que peuvent prendre les données relatives aux ressources et l'on saisit le potentiel d'utilisation qu'elles présentent et qui permettrait de répondre à ces missions plus efficacement et en innovant. Le Web sémantique et ses applications viennent répondre à cette nécessité et proposent une nouvelle méthode visant à traiter de façon automatique l'information et les données qui la constituent sur le Web.

Quelques définitions

Pour toute définition complémentaire, nous invitons le lecteur à se référer au glossaire.

Web sémantique

L'expression « Web sémantique » (*Semantic Web*) est attribuée à Berners-Lee, Hendler & Lassila (2001) qui le décrivent non pas comme un Web distinct de celui que l'on connaît, mais plutôt comme une extension de celui-ci. Il s'agit d'un ensemble de technologies dont l'objectif est d'exploiter plus efficacement la quantité importante d'informations que l'on retrouve sur le Web, c'est-à-dire de trouver, partager, réutiliser ou modifier de l'information de manière plus rapide et efficace. Dans cet espace virtuel, l'information est disponible dans une grande variété de langues naturelles et dont le public est humain tandis que, dans le Web sémantique, l'information est exprimée dans un langage destiné à être compris et interprété par des machines (DeWeese & Segal 2015). De grandes quantités de données structurées sont présentement conservées dans des bases de données isolées du Web, que l'on pourrait comparer à des silos d'informations, dont font partie les métadonnées des catalogues de bibliothèques. L'interprétation par les machines permet de mieux gérer l'information de manière automatique, de permettre la création de nouveaux services et applications ainsi que d'assurer l'accessibilité aux connaissances à la fois par l'humain et par la machine. D'ailleurs, dès les balbutiements du Web, on évaluait déjà la possibilité d'organiser l'information afin que celle-ci soit « comprise » par les machines (Berners-Lee, Cailliau, Luotonen, Nielsen & Secret 1994; Shadbolt, Berners-Lee & Hall 2006).

Web de données

Le Web de données (*Linked Data*), aussi appelé « données liées », pour sa part, est l'une des composantes du Web sémantique et probablement l'application la plus connue. Afin d'éviter toute confusion entre Web sémantique et Web de données, il est possible de dire que le Web de données est l'une des applications qui entrent dans la grande famille du Web sémantique. Par exemple, les données structurées internes, qui permettent d'ajouter du contenu sémantique à même les pages HTML, sont aussi considérées comme l'une des applications du Web sémantique. Le Web de données doit donc être considéré comme l'ensemble des pratiques et des standards permettant de publier des données structurées sur le Web afin que celles-ci puissent être liées entre elles et interrogées. Il s'agit aussi de promouvoir l'idée selon laquelle les données sont comparables aux documents et qu'il est possible de les relier entre elles de la même façon que les documents sont liés entre eux grâce aux liens hypertextes sur le Web traditionnel (Bermès, Isaac & Poupeau 2013; Bizer & Heath 2011). Berners-Lee (2006) a défini quatre principes de base du Web de données sur

Il s'agit d'un ensemble de technologies dont l'objectif est d'exploiter plus efficacement la quantité importante d'informations que l'on retrouve sur le Web, c'est-à-dire de trouver, partager, réutiliser ou modifier de l'information de manière plus rapide et efficace.

lesquels la majorité des acteurs du domaine s'appuiera par la suite pour déployer les technologies qui les caractérisent :

1. Nommer les ressources avec des *Uniform Resource Identifier* (URI);
2. Utiliser des URI déréférencables (protocole HTTP afin qu'il soit possible d'accéder à des informations sur les ressources;
3. S'assurer que les URI déréférencables fournissent des informations pertinentes à l'aide des standards tels que *Resource Description Framework* (RDF) et *SPARQL Protocol and RDF Query Language* (SPARQL);
4. Créer un réseau de liens avec d'autres URI provenant d'autres bases de données.

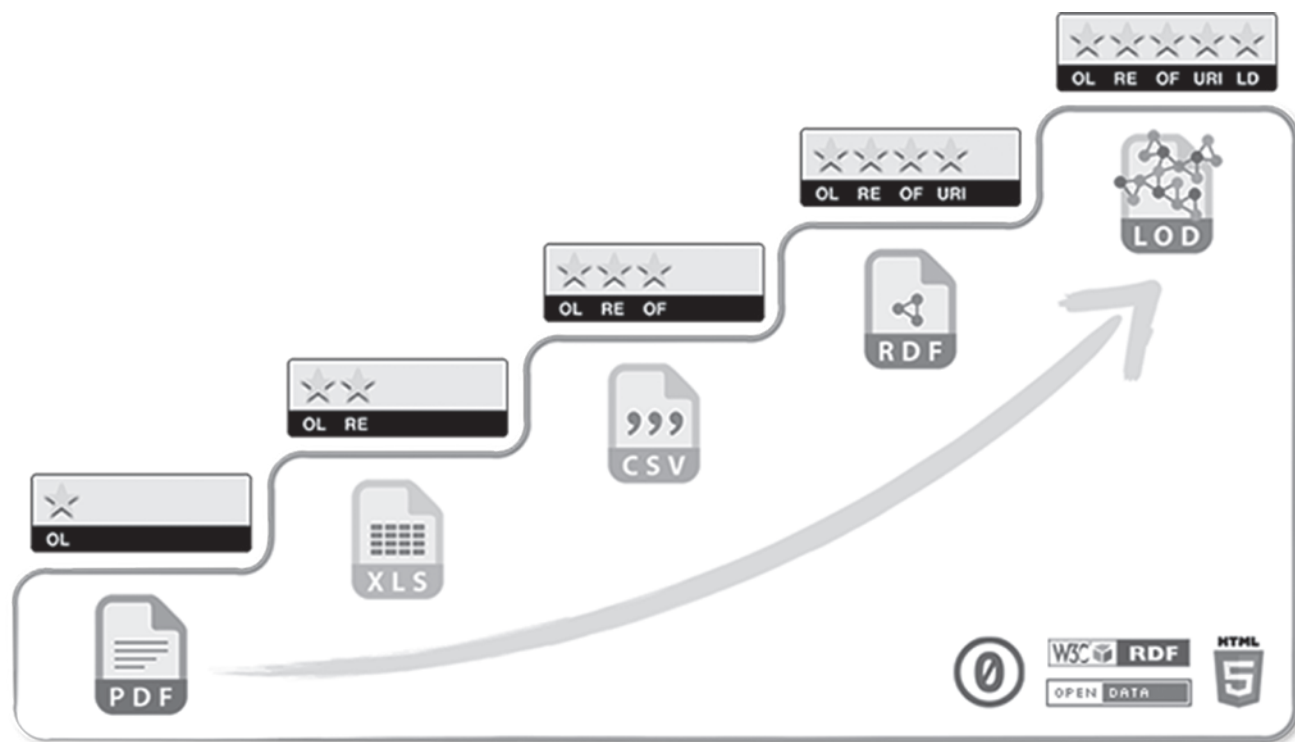
Dans le milieu des bibliothèques, on constate la possibilité de jumeler les principes du Web de données à ceux des données ouvertes et, ainsi, non seulement rendre accessibles les données, mais aussi les lier entre elles. C'est ce qu'on appelle les données ouvertes et liées (DOL) (plutôt qu'uniquement « Web de données » ou encore « données liées »). Dans sa note sur les données liées (2006), Berners-Lee a, en 2010, ajouté une échelle de qualité basée sur cinq étoiles qui permet d'évaluer jusqu'à quel point l'information est facilement réutilisable. Les cinq étapes (ou étoiles) sont les suivantes et la figure 1 permet de les illustrer :

1. Rendre vos données disponibles sur le Web (quel que soit leur format) en utilisant une licence ouverte (*Open Licence*);
2. Rendre vos données disponibles sous forme de données structurées (par exemple, en format Excel [.xlsx] plutôt que sous forme d'image numérisée d'un tableau) (*Reusable*);
3. Utiliser des formats non-propriétaires (par exemple utiliser le comma-separated values [.csv] plutôt que le format Excel [.xlsx]) (*Open Format*);
4. Utiliser des URI pour identifier vos données afin que les autres utilisateurs puissent pointer vers elles (*URI*);
5. Relier vos données à d'autres données pour fournir du contexte (*Linked Data*).

Potentiel des données ouvertes et liées en bibliothèque et impacts

Les professionnels de l'information sont entre autres formés de façon à être aptes à analyser l'information structurée, à effectuer la description de contenus en suivant des standards et des normes et à participer à l'évolution et à la

FIGURE 1

Échelle de qualité des données ouvertes propose par Berners-Lee (2010)¹

définition de formats informatiques qui permettront de conserver, manipuler et échanger les données (Bermès, Isaac & Poupeau 2013). Il est important de miser sur cette habileté déjà acquise quant à l'encodage des données structurées afin de permettre une évolution des catalogues de bibliothèques. Il est en effet maintenant nécessaire de mettre de plus en plus de l'avant l'aspect d'interopérabilité, non seulement dans le contexte bibliothéconomique, mais aussi dans le contexte infiniment plus large qu'est le Web. Les données ouvertes et liées se présentent comme l'une des solutions envisagées pour répondre aux nouveaux défis associés aux évolutions technologiques.

D'abord, les données ouvertes et liées impliquent la réutilisation de données par les moteurs de recherche, donc une meilleure visibilité. En effet, les ressources des catalogues de bibliothèques seraient ainsi disponibles pour les usagers du Web, ce qui ferait en sorte que l'expertise des bibliothécaires et professionnels de l'information pourrait être mise de l'avant grâce à la qualité et la fiabilité de ces ressources et de leurs descriptions (Coyle 2010; Gonzales 2014). Étant donné que l'accès aux ressources dans les bibliothèques numériques dépend des métadonnées qui les décrivent, le Web de données permettrait de naviguer d'une ressource à l'autre, même si ces ressources sont externes au catalogue de bibliothèque. Contrairement aux catalogues en ligne qui sont basés sur le fait d'effectuer une recherche textuelle

pour obtenir des résultats (par exemple, avec le nom d'un auteur ou le titre d'une ressource), le Web de données ouvre la porte à des recherches plus larges et qui peuvent être basées sur des concepts plutôt que sur des informations précises. En effet, l'utilisateur pourrait faire des découvertes inattendues et naviguer d'une bibliothèque numérique à une autre ou encore d'une bibliothèque numérique à des sources externes telles que Wikipédia, l'exposition virtuelle d'un musée, une base de données thématique, etc. On constate ainsi du même coup la possibilité d'effectuer des recherches fédérées ainsi que des recherches interdisciplinaires plus performantes, qui permettraient d'accéder à l'information provenant d'un plus grand nombre de bases de données ouvertes.

Ensuite, contrairement à la pratique établie où les données sont partagées sous la forme de notices, il s'agit donc ici de la possibilité de créer des graphes sur des ressources données dont les informations proviennent de sources différentes. Par le fait même, on constate la possibilité de diminuer la redondance en ce qui a trait aux descriptions bibliographiques et aux métadonnées. Il s'agit ainsi d'une occasion d'optimiser les coûts liés aux activités de description des fonds. Conséquemment, les bibliothèques peuvent partager entre elles les données qui sont déjà accessibles

1. Source : <5stardata.info/en>.

par l'intermédiaire de sources fiables et ainsi éviter la duplication du travail (Alemu, Stevens, Ross & Chandler 2012; Tillet 2013).

Le fait d'avoir un format universel pour le partage de données présente des avantages, car celui-ci permettrait une meilleure interopérabilité d'une institution à une autre, mais aussi d'une institution à d'autres instances publiant

leurs données, et ce, peu importe le système utilisé. Les technologies développées pour la gestion des données en bibliothèques sont présentement produites par les fabricants de systèmes intégrés de gestion de bibliothèque (SIGB) et les normes bibliographiques sont conçues uniquement pour la communauté des professionnels de l'information, ce qui limite les possibilités de partages et contribue largement à l'isolement de leurs données.

Finalement, Baker *et al.* (2011) soulignent aussi qu'étant donné que les données liées sont caractérisées par le fait qu'elles décrivent la sémantique, et ce, indépendamment des formats et de la syntaxe, même si un nouveau format voit le jour, les données garderont leur signification et sont par conséquent plus pérennes.

Dans le monde, plusieurs bibliothèques nationales ont déjà incorporé les technologies relatives aux données liées à leurs catalogues, soit la Bibliothèque nationale de France (BnF) avec son projet dédié², la Bibliothèque nationale d'Allemagne³, la Library of Congress⁴, la Bibliothèque nationale d'Espagne⁵ et la Bibliothèque nationale du Royaume-Uni⁶. Au Québec et au Canada, notons entre autres l'existence de l'Initiative canadienne sur les données liées⁷ (ICDL), mise sur pied en 2015 à la suite de la constatation que certaines institutions américaines et européennes ont déjà une longueur d'avance dans le développement de projets de données liées en bibliothèques. L'ICDL est née d'une collaboration entre l'Université de Toronto, l'Université McGill, l'Université de Montréal, l'Université de l'Alberta et l'Université de la Colombie-Britannique, et est menée en partenariat avec Bibliothèque et Archives Canada (BAC) et Bibliothèque et Archives nationales du Québec (BANQ). En octobre 2016, le Sommet canadien sur les données liées⁸ a eu lieu à Montréal. Celui-ci visait à sensibiliser différents

acteurs du domaine à l'importance des données liées en misant sur le partage d'expériences et à permettre aux membres de se rencontrer. L'ICDL est composée de différents groupes de travail, dont le Groupe de travail francophone qui a pour mission d'étudier les défis que pose l'adoption des technologies du Web sémantique en milieu francophone.

De plus, quelques projets ont été développés ou sont en cours de développement dans le but de démontrer que les données ouvertes liées représentent une bonne option pour le partage de données et de ressources d'une institution à l'autre. Prenons par exemple le projet Au-delà des tranchées⁹ mené par le Réseau pancanadien du patrimoine documentaire (RPCPD), le dépôt des notices de BANQ dans le fichier d'autorité international virtuel (VIAF), 150 années d'art canadien¹⁰ mené par Artefacts Canada et quelques projets toujours en cours de développement liés au Plan culturel numérique du Québec¹¹. Bien que l'on constate un intérêt de plus en plus marqué au sein de différentes institutions et sur le plan de la recherche, il reste du travail à effectuer et il est nécessaire d'encourager la diffusion des connaissances, le partage de compétences et le retour sur les expériences vécues.

Il est aussi nécessaire de mentionner le modèle de données pour la description bibliographique *Bibliographic Framework*¹² (BIBFRAME) initié par la Library of Congress, développé dans l'objectif clair de remplacer les formats MARC et de fournir un format qui serait compatible avec le Web de données. En constante évolution, il est basé sur RDE, sur les principes des données liées et sur trois niveaux d'abstraction : l'œuvre qui reflète l'essence conceptuelle de la ressource à cataloguer, l'instance qui est la réalisation matérielle d'une œuvre et l'item qui est la copie physique ou électronique de l'instance. Comme on peut le constater, le modèle n'est pas sans rappeler un autre modèle, le *Functional Requirements for Bibliographic Records* (FRBR). Il pourrait possiblement se présenter comme une solution à l'implémentation de projets de Web de données dans les institutions documentaires par le fait qu'il a été réfléchi expressément pour répondre aux besoins des bibliothèques.

2. <data.bnf.fr/>.

3. <www.dnb.de/EN/Service/DigitaleDienste/LinkedData/linkedata_node.html>.

4. <id.loc.gov/>.

5. <datahub.io/dataset/datos-bne-es>.

6. <datahub.io/dataset/bluk-bnb>.

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

8. <www.mcgill.ca/clds/fr/propos>.

9. <www.canadiana.ca/rpcpd-dol>.

10. <chin-rcip.canadiana.ca/acld/?lang=fr>.

11. <culturenumerique.mcc.gouv.qc.ca/>.

12. <www.loc.gov/bibframe/docs/bibframe2-model.html>.

Étapes pour la mise sur pied d'un projet de données ouvertes et liées

Parmi la littérature disponible sur le Web de données et le Web sémantique, on retrouve quelques recommandations et méthodologies sur lesquelles se baser afin de publier des jeux de données en respectant les normes et standards (notamment Bizer & Heath 2011; Van Hooland & Verborgh 2014; W3C 2014; Zengenene, Casarosa & Meghini 2014). Cependant, il est difficile d'obtenir une version claire et condensée d'une marche à suivre afin d'appliquer ces technologies aux données bibliographiques. Nous proposons ici les étapes nécessaires à la mise en place d'un projet de Web de données. La formulation de ces étapes est le résultat d'une synthèse effectuée à la suite d'une revue de la littérature approfondie. Les informations étant disséminées à travers différents textes, un travail d'analyse, de regroupement, de classement et de fédération de celles-ci a été effectué. Le but de l'exercice est de proposer un processus clair de publication dans le Web de données en se basant sur les standards préconisés par le World Wide Web Consortium (W3C) afin de faciliter l'appropriation des technologies du Web sémantique par les professionnels de l'information. Les deux premières étapes ne sont pas spécifiques à un projet de Web de données et constituent plutôt des exigences liées au facteur humain. En effet, elles pourraient s'appliquer à n'importe quelle gestion de projet d'implémentation nouvelle. Cependant, elles sont tout de même primordiales à son fonctionnement et à son acceptation au sein de l'institution.

Comprendre la motivation et obtenir l'autorisation des parties prenantes

Avant de commencer la mise en place d'un tel projet, il est nécessaire que tous les acteurs impliqués puissent avoir une vision globale de ce que signifient les données ouvertes et liées et les technologies qui s'y rattachent, mais aussi les avantages et défis qui en découlent. Ce type d'initiative permet souvent de répondre à un besoin spécifique en lien avec un jeu de données pertinent. L'importance réside dans la compréhension du rôle des sciences de l'information dans cette nouvelle extension du Web, dans l'amélioration de l'expérience de l'utilisateur, mais aussi dans le perfectionnement du processus de traitement de l'information en général (Zengenene, Casarosa & Meghini 2014). La vulgarisation ainsi que la capacité à communiquer les bénéfices pour les institutions sont primordiales afin de permettre aux instances de s'approprier les technologies du Web de données et de constater les bénéfices à courts, moyens et longs termes. L'un des défis les plus importants est le fait qu'il est, pour le moment, difficile, voire impossible, d'évaluer les retombées économiques d'un tel projet et même d'en mesurer l'efficacité de manière concrète. Ainsi, il est nécessaire

de se concentrer sur les autres avantages tels que l'augmentation du nombre de visiteurs sur les plateformes Web de l'institution, la présence et la visibilité des données dans les résultats présentés par les moteurs de recherche, l'amélioration des services aux usages, les possibilités de réutilisation des données ouvertes permettant la création de nouvelles applications et la participation à un mouvement collaboratif d'envergure internationale visant, entre autres, l'évolution du catalogue de bibliothèque.

Une fois les différentes instances préparées, le projet présenté de façon claire et simple et le document rédigé, il devient plus facile de communiquer les objectifs et d'évaluer les ressources nécessaires. Pour obtenir l'autorisation des parties prenantes, il est aussi possible de présenter des exemples de projets de publication de jeux de données dans le Web de données qui permettront de constater concrètement ce que ces applications peuvent apporter comme nouveaux services et nouvelles applications. Afin de rendre possible une association aux approches traditionnelles de gestion de l'information, il est recommandé de présenter des modèles identifiants le cycle de vie des données ouvertes et liées.

Établir une licence d'utilisation

Il est primordial de statuer sur les droits d'utilisation des données publiées en indiquant le propriétaire des données et dans quelle mesure il est possible de les réutiliser, surtout lorsqu'il s'agit de données provenant d'institutions culturelles publiques (Bermès 2013; Bizer & Heath 2011; Hyvönen 2012; Villazón-Terrazas, Vilches-Blázquez, Corcho & Gómez-Pérez 2011; W3C 2014; Zengenene, Casarosa & Meghini 2014). Les usagers et les développeurs auront d'ailleurs plus tendance à s'approprier les données si une licence présente les spécificités légales de manière claire et précise.

De plus en plus, on constate l'utilisation du modèle de licences *Creative Commons*¹³. Par exemple, Données Québec¹⁴, la Bibliothèque nationale d'Espagne, la Bibliothèque nationale d'Allemagne et plusieurs autres utilisent les licences *Creative Commons* pour encadrer légalement la réutilisation de leurs données. Il existe plusieurs autres types de licences disponibles. On soulignera au passage le projet RightsStatements.org¹⁵, une initiative de la Digital Public Library of America¹⁶ (DPLA) et de Europeana¹⁷, qui propose un ensemble d'énoncés de droits normalisés réfléchis spécifiquement pour les institutions du patrimoine culturel. Ceux-ci ont été conçus en gardant en tête à la fois

13. <creativecommons.org>.

14. <www.donneesquebec.ca/fr/licence>.

15. <rightsstatements.org/en/>.

16. <dp.la/>.

17. <www.europeana.eu/portal/en>.

les usagers et les machines et déploient les technologies du Web sémantique.

Lorsque vient le temps de choisir une licence, il faut tenir compte de trois concepts différents, soit l'attribution, la restriction de l'usage commercial et la redistribution à l'identique (Bermès, Isaac & Poupeau 2013). Les clauses d'une licence liées à l'attribution («by») sont celles qui précisent comment doit être reconnu le détenteur de la propriété intellectuelle de la ressource. Ainsi, dans certains cas, l'utilisateur peut être dans l'obligation de citer la source des données dans le contexte d'une réutilisation. La restriction de l'usage commercial («nc») implique que l'utilisateur ne peut réutiliser des données à des fins commerciales. Une licence peut aussi préciser les différentes possibilités pour l'usage commercial. Finalement, la redistribution à l'identique («share-alike») précise si l'utilisateur peut ou non réutiliser et transmettre les données en utilisant une licence semblable à celle utilisée par l'institution.

Évaluer les compétences

À cette étape, il est nécessaire de commencer à évaluer quel processus de conversion sera le plus pertinent pour les jeux de données à publier. Pour ce faire, une analyse de la situation qui présente les compétences des différents acteurs impliqués dans le processus doit être effectuée. Selon Zengenene, Casarosa & Meghini (2014), les compétences nécessaires pour les bibliothécaires et autres professionnels de l'information sont triples :

- Systèmes d'information :
 - Avoir des connaissances en ce qui a trait au téléchargement, à l'installation et à la configuration de systèmes d'information, de bases de données et de serveurs ;
 - Être aptes à écrire et lire les formats XML et RDF.
- Métadonnées :
 - Connaître le processus de catalogage, l'importance et la signification des métadonnées.
- Modélisation :
 - Comprendre la structure des données et être aptes à évaluer la meilleure façon de convertir les données d'une structure donnée vers RDF.

Ensuite, dans la mesure où l'institution se voit dans l'obligation d'engager des techniciens à l'externe (développeurs de logiciels, par exemple), les bibliothécaires et professionnels de l'information doivent être aptes à communiquer leurs besoins et à décrire les ontologies et vocabulaires qu'ils souhaitent utiliser pour décrire leurs données. Mettre sur

pied une équipe dont l'objectif sera de mettre en place un projet de Web de données peut s'avérer particulièrement ardu. En effet, on constate d'un côté le besoin de développeurs de logiciels et de programmeurs qui seront aptes à construire l'architecture, mais aussi la nécessité d'avoir au sein de l'équipe des individus qui maîtrisent parfaitement bien le contenu qui doit être manipulé. Ainsi, la place des professionnels de l'information au sein de l'équipe de développement est absolument nécessaire. Les compétences doivent donc être hétérogènes, diversifiées et bien développées.

Évaluer les jeux de données

En ce qui a trait à l'identification des jeux de données à publier, on doit analyser lesquels on souhaite rendre disponibles, les formats dans lesquels ils sont encodés ainsi que les technologies nécessaires à chaque situation. On recommande en général d'accorder la priorité à la publication de données qui sont uniques à l'institution et qui présentent un intérêt de réutilisation par des usagers. Trois types de données sont à considérer (Bermès, Isaac & Poupeau 2013) :

- Notices se trouvant dans le catalogue ou la base de données ;
- Référentiels ou notices d'autorité ;
- Éléments de métadonnées ou ontologies développées.

Les référentiels ou les données d'autorité présentent pour leur part un avantage intéressant, dans la mesure où ils sont souvent le résultat d'efforts importants de normalisation et offrent une grande possibilité de réutilisation par d'autres institutions. Par exemple, comme mentionné précédemment, les thésaurus et listes d'autorité RAMEAU (BnF) et Library of Congress Subject Headings (LCSH) ont été rapidement publiés sur le Web de données. Pour ce qui est de la troisième option, soit les ontologies ou les éléments de métadonnées développés au sein de l'institution, il est plutôt recommandé d'utiliser des ontologies existantes plutôt que de procéder à la création de celles-ci afin d'assurer une

meilleure interopérabilité. Il peut cependant arriver que certaines institutions ressentent le besoin de créer leur ontologie lorsque leurs données présentent des spécificités particulières. Finalement, il est intéressant d'envisager la publication de données administratives et techniques afin de s'inscrire dans le mouvement des données

ouvertes, telles que des données de nature statistique et transactionnelle. Dans tous les cas, il est important de statuer quant au choix des jeux de données à publier, car cela facilitera les tâches suivantes et l'analyse du travail qui sera à effectuer.

[...] la place des professionnels de l'information au sein de l'équipe de développement est absolument nécessaire. Les compétences doivent donc être hétérogènes, diversifiées et bien développées.

Choisir le modèle de publication et évaluer les outils nécessaires

Cette étape est importante dans la mesure où elle permet d'avoir une idée d'ensemble du processus de publication des données. D'abord, si les documents sont en format texte en langue naturelle (par exemple, des rapports ou des articles), il est possible d'utiliser un extracteur d'entités nommées et d'annotations Web sémantique, qui permettra la reconnaissance et l'extraction d'informations dans un corpus donné ainsi que l'annotation en RDF (Bizer & Heath 2011). Le fonctionnement de ce type d'extracteur est le suivant : les documents sont annotés par l'extracteur à l'aide des URI des entités nommées dans le document. La publication de ces documents accompagnés de ces annotations augmente les chances de découverte et permet aux applications du Web sémantique d'effectuer des liens vers ceux-ci. Notons cependant que les extracteurs d'entités nommées ne permettront pas d'extraire les relations entre les entités. Ainsi, on aura la possibilité de définir les URI, mais non les triplets RDF. Un travail supplémentaire est donc nécessaire afin d'obtenir des fichiers RDF qui pourront être versés sur un serveur Web. Cette façon de faire est la plus simple et est principalement utilisée lorsque la quantité de fichiers est limitée, car le travail se fait principalement de façon manuelle.

Ensuite, dans la mesure où les données sont déjà structurées, ce qui est le cas dans la plupart des situations liées au domaine des sciences de l'information, tout dépend alors de la façon dont les données sont stockées. Ainsi, si les données sont stockées dans une base de données relationnelle, deux technologies de conversion peuvent être utilisées : RDB-to-RDF¹⁸ ou l'utilisation d'un système de gestion de contenu qui permet d'exprimer les données en RDFa. Un système de gestion de contenu permet la création et la gestion de contenu numérique via une interface conviviale. Ainsi, certains outils comme Drupal¹⁹, lui-même un système de gestion de contenu, permettent donc de transformer les données structurées qui se trouvent dans sa base de données en RDFa en ajoutant des attributs RDF aux éléments HTML. Mentionnons aussi l'existence des langages de mise en correspondance de données qui permettent de transformer les données provenant de bases de données relationnelles en RDF²⁰. Lorsque les données sont encodées dans des bases de données relationnelles, il est en général préférable de ne pas procéder à un transfert des données vers un *triplestore* afin de conserver l'infrastructure de gestion de l'information déjà mise en place (Bizer & Heath 2011).

Il est aussi possible d'accéder aux données stockées via une interface de programmation applicative (API). Dans ce cas-ci,

la situation est plus complexe, dans la mesure où l'institution devra développer un adaptateur personnalisé (*wrapper*). Un adaptateur permettra de régler certaines limites qui sont associées aux API, comme, par exemple, le fait que leur contenu ne peut être repéré par les moteurs de recherche. En général, les adaptateurs permettent d'assigner des URI aux ressources sur lesquelles l'API fournit des données, de reformuler une requête pour que celle-ci soit comprise par l'API et de transformer les résultats de cette requête en RDF (Bizer & Heath 2011). Le développement d'un tel outil implique des connaissances informatiques qui dépassent en général celles déjà acquises par les professionnels de l'information, donc l'appel à une ressource externe peut alors être nécessaire.

Finalement, les données structurées peuvent aussi être stockées dans des *triplestores*. La plupart des *triplestores* offrent une interface de données liées, ce qui peut faciliter grandement le travail de programmation. On peut donc accéder directement à un point d'accès SPARQL et il est possible d'effectuer la configuration et la gestion du contenu à même le *triplestore*.

Attribuer les URI

Il est nécessaire de porter une attention particulière à la création d'URI pour décrire les ressources et les relations qui les unissent. Cette étape permet aussi de constater la quantité d'entités avec lesquelles on devra travailler et qu'il sera nécessaire d'identifier. Les URI doivent être construits de façon simple, stable, pérenne et gérable (Berners-Lee 2008 ; Villazón-Terrazas, Vilches-Blázquez, Corcho & Gómez-Pérez 2011).

On devra d'abord choisir un nom de domaine qui ne sera pas susceptible de changer au fil du temps. Ensuite, pour ce qui est de la description des ressources, il est préférable d'utiliser des identifiants déjà existants, tels que les référentiels ou les listes d'autorité, qui ont démontré leur persistance dans le temps. Par exemple, un ISBN pour un livre, des codes de bibliothèque ou, comme c'est le cas pour la Bibliothèque nationale de France, le numéro de notice associée à la ressource (Bermès, Isaac & Poupeau 2013 ; Villazón-Terrazas, Vilches-Blázquez, Corcho & Gómez-Pérez 2011 ; Zengenene, Casarosa & Meghini 2014). Dans la mesure où aucun identifiant n'existe, l'institution devra en créer de nouveaux et deux options s'offrent alors à elle. L'utilisation d'identifiants opaques, qui n'ont aucune signification particulière, ou d'identifiants signifiants, qui présentent une information lisible et compréhensible par l'humain.

Il est important de s'assurer de respecter le concept de négociation de contenu et de procéder à la création d'au moins trois URI pour identifier la même ressource (ou ses différentes représentations). Ainsi, il est nécessaire d'accorder une attention particulière à l'origine de l'identifiant (référentiel ou liste d'autorité), à la forme de l'identifiant (opaque ou

18. *Relational database to RDF*.

19. <www.drupal.org>.

20. D2RQ ou R2RML, par exemple.

signifiant) et à la création d'au moins trois identifiants pour une même ressource (un pour le concept, un pour la représentation HTML et un pour la représentation RDF).

Choisir son modèle de données et effectuer la mise en correspondance et le nettoyage des données

Cette étape, qui consiste à développer la structure sémantique, peut présenter certaines difficultés, surtout dans la mesure où les métadonnées sont encodées de manières différentes, dans le cas où, par exemple, elles proviendraient de catalogues différents. Un modèle de données découle lui-même de la modélisation des données, c'est-à-dire le fait d'analyser et de concevoir la structure des données qui sont contenues dans un système, mais aussi leur sémantique. D'ailleurs, la modélisation des données dépendra du mode de publication choisi et de leur nature. Ainsi, cette étape implique les choix relatifs aux classes et aux propriétés qui seront sélectionnées pour décrire les entités et les relations qui les unissent. Pour ce faire, il est donc important de bien connaître le niveau de description et les éléments de métadonnées utilisés pour, par la suite, effectuer une mise en correspondance entre les jeux de données. La plupart du temps, l'institution devra utiliser plusieurs ontologies et vocabulaires pour représenter son information.

La difficulté réside aussi dans le fait que le modèle de données doit permettre de représenter toutes les entités et leurs relations, tout en assurant une logique durant le processus. Le choix des ontologies et des vocabulaires est donc l'une des parties les plus importantes du processus de publication de données ouvertes et liées sur le Web et l'on encourage la réutilisation de standards. Notons que le choix dépend grandement des objectifs de l'institution ainsi que du type de données qu'elle souhaite publier. De plus, il est fort probable que la mise en correspondance des données ait comme conséquence une perte de granularité au niveau des descriptions. Cet aspect peut décevoir, mais il est à noter que les données qui seront publiées dans le Web de données n'ont pas comme objectif de remplacer les notices bibliographiques qui sont souvent plus précises, mais plutôt d'apporter une couche supplémentaire d'interopérabilité et une possibilité de réutilisation.

Ensuite, une fois le ou les vocabulaires et ontologies choisis, et ainsi, le modèle de données défini, il est nécessaire de procéder à la mise en correspondance des données. Il s'agit d'une sous-étape qui peut donner du fil à retordre dans la mesure où elle nécessite une bonne maîtrise des technologies qui seront utilisées, du vocabulaire ou de

l'ontologie et des formats. En général, on commence par la création d'un tableau de conversion qui permettra d'indiquer la source, la cible, la règle de conversion et, si possible, un exemple. Il est possible de créer la conversion selon différents niveaux de précision et de détail, le choix revenant à l'institution d'évaluer les différentes possibilités. Puis, on doit par la suite transformer les données en fonction des règles de conversion. Pour ce faire, on utilise généralement un programme informatique et le choix de cet outil dépend du format source.

Avant de convertir les données en RDF, il est nécessaire de s'assurer qu'on y retrouve le moins d'erreurs possible et qu'on en augmente la qualité. On pense à des erreurs typographiques, des valeurs manquantes, des doublons ou des contradictions (Rahm & Do 2000; Van Hooland & Verborgh 2014). Par exemple, il pourrait y avoir des informations manquantes en ce qui a trait à une notice bibliographique (date de publication, numéro de l'édition, etc.) ou la forme choisie pour le nom de l'auteur pourrait ne pas correspondre à la notice d'autorité. Il est évident que le résultat ne sera pas parfait, mais il est toujours préférable de faire quelques vérifications avant de procéder à la conversion. Il est donc nécessaire de s'assurer que les données sont dans un format structuré et l'on peut par la suite utiliser des outils tels que OpenRefine²¹ ou DataWrangler²² pour procéder au nettoyage.

Enrichir les données en créant des liens et convertir les données en RDF

Cette étape est primordiale et a comme objectifs de définir des triplets qui seront connectés à d'autres et de créer des triplets qui permettront de définir des relations, le tout à l'interne et à l'externe (Zengenene, Casarosa & Meghini 2014). Par exemple, il est possible de lier l'URI d'un auteur à d'autres notices d'autorité d'institutions reconnues comme la Library of Congress ou le Virtual International Authority File (VIAF). Lorsqu'on crée des liens au sein même du jeu de données, il est nécessaire de s'assurer que toutes les ressources seront connectées entre elles. Pour ce qui est des liens vers l'externe, il est préférable de se lier à des jeux de données qui sont fiables, grandement utilisés, stables et bien établis. C'est le cas de DBpedia²³, Geonames²⁴, VIAF et Library of Congress. Le choix de ces jeux de données dépend de plusieurs

Le choix de ces jeux de données dépend de plusieurs facteurs, tels que la valeur ajoutée que ceux-ci peuvent apporter aux données de l'institution, mais aussi la visibilité que ces liens peuvent apporter.

21. <openrefine.org>.

22. <vis.stanford.edu/wrangler/>.

23. <wiki.dbpedia.org/>.

24. <www.geonames.org/ontology/documentation.html>.

TABLEAU 1

Points à vérifier avant de procéder à la publication des données sur le Web de données

Points à vérifier	☑
Est-ce que le jeu de données est lié à d'autres jeux de données ?	
Avez-vous rendu disponible la provenance des métadonnées ?	
Avez-vous rendu disponible une licence d'utilisation ?	
Utilisez-vous des termes provenant de vocabulaires ou d'ontologies bien établis ?	
Les URI sont-ils déréférencables ?	
Avez-vous effectué la mise en correspondance des données entre les termes de votre vocabulaire ou ontologie avec d'autres vocabulaires ou ontologies, si applicable ?	
Avez-vous rendu disponible une description des jeux de données ?	

facteurs, tels que la valeur ajoutée que ceux-ci peuvent apporter aux données de l'institution, mais aussi la visibilité que ces liens peuvent apporter. Notons la nécessité de prendre connaissance de la licence d'utilisation de chacune des entités externes vers lesquelles on souhaite effectuer des liens et, ainsi, réutiliser les données. Selon Bermès, Isaac & Poupeau (2013), la création de liens implique trois tâches liées, soit :

- identifier quels sont les points de contact entre ses propres jeux de données et ceux de l'externe ;
- identifier les liens qui unissent les entités de l'externe et celles de son jeu de données ;
- évaluer quelle est la meilleure méthode pour effectuer ces liens, soit de façon manuelle ou automatique.

Le stade de la conversion des données vers RDF ne pose en général pas trop de difficultés. Il est cependant nécessaire de choisir la ou les syntaxes de sérialisation dans lesquelles les données seront rendues disponibles pour les usagers qui voudraient peut-être les réutiliser, mais surtout pour les machines. Il est important de savoir que la procédure doit se faire de façon automatique ou semi-automatique étant donné que d'effectuer un tel travail manuellement serait beaucoup trop fastidieux. Il existe un bon nombre d'outils qui permettent la conversion.

Valider et publier les jeux de données

Avant de publier les jeux de données sur le Web, il est recommandé d'en vérifier la qualité. Bizer & Heath (2011) proposent une liste de points à vérifier avant de procéder à la publication. Le Tableau 1 propose une liste de vérification.

À cette étape, il est aussi pertinent de vérifier si les triplets RDF sont bien construits à l'aide d'outils de validation et de relire l'échelle de qualité basée sur les cinq étoiles présentée précédemment.

Une fois les données converties et évaluées, il est possible de les rendre accessibles sous forme de fichiers RDF, dans un

triplestore présentant une interface d'interrogation SPARQL, via un point d'accès SPARQL, via un RESTful API (*Representational State Transfer*) ou encore directement dans les pages Web grâce aux données structurées internes.

Conclusion

L'objectif de cet article était de proposer une série d'étapes afin de faciliter l'implantation d'un projet de données ouvertes et liées en bibliothèque. On constate que le Web de données permet de jeter un regard nouveau sur les occasions d'innover qui se présentent aux professionnels de l'information. Les comportements et les besoins informationnels ayant grandement changé depuis l'arrivée du Web, les défis sont nombreux et les bibliothèques et institutions documentaires doivent s'adapter tout en travaillant au processus d'évolution du catalogue. Ce nouvel environnement potentiel présente de nombreux avantages. Cependant, le déploiement de ces technologies n'est pas sans défi. En effet, de tels changements impliquent de nombreuses modifications quant aux habitudes de travail et un temps non négligeable quant à la formation requise pour mettre en pratique ces connaissances. Ainsi, les ressources nécessaires sont importantes et une réflexion s'impose de la part des décideurs et des instances gouvernementales. On peut cependant croire que les projets de Web de données prendront de plus en plus d'ampleur lorsque les pratiques auront été établies de manière plus claire et qu'un plus grand nombre de projets concrets auront vu le jour.

Glossaire

Application Programming Interface (API) : Interface de programmation qui permet d'accéder à l'ensemble de règles qui composent une application. Un développeur de logiciels peut ainsi y accéder et procéder à l'échange de données.

Données ouvertes : Données diffusées de manière structurée et en libre accès selon une licence d'utilisation permettant leur réutilisation par tous.

Données ouvertes et liées : Concept qui jumèle les principes du Web de données aux principes des données ouvertes.

Données structurées internes : Contenu sémantique ajouté à même les pages HTML afin d'en faciliter le repérage par les moteurs de recherche. Il s'agit de l'une des applications du Web sémantique.

Licence d'utilisation : Contrat dans lequel sont stipulées les conditions d'utilisation, de modification et de diffusion des jeux de données.

Métadonnée : Donnée normée présentant de l'information précise à propos d'une autre donnée.

Modèle de données : Modèle visant à définir la structure et la sémantique des données contenues dans un système.

Négociation de contenu : Mécanisme qui permet de proposer pour une même ressource différentes visualisations en fonction de la provenance de la requête.

Ontologie : Ensemble structuré visant à définir les concepts et les relations afin de classifier l'utilisation de certains termes dans une application donnée, caractériser les relations entre les ressources et clarifier les contraintes quant à l'utilisation d'un terme.

Point d'accès SPARQL : Service qui respecte le protocole SPARQL et qui présente la possibilité d'effectuer des requêtes dans un *triplestore*.

RDFa : Recommandation du W3C définissant une méthode permettant l'ajout de données structurées dans une page HTML ou un document XML.

Resource Description Framework (RDF) : Modèle de graphe ayant comme objectif de décrire les ressources et leurs métadonnées. Un document RDF est basé sur la formation de triplets.

SPARQL Protocol and RDF Query Language (SPARQL) : Langage de requête permettant d'effectuer des recherches au sein de données encodées en RDF, de les récupérer et de les manipuler.

Syntaxe de sérialisation : Syntaxe permettant l'encodage de données et le partage de celles-ci entre les systèmes.

Triplestore : Base de données conçue pour stocker et récupérer des données en RDF. On retrouve donc dans un *triplestore* uniquement des triplets.

Triplet RDF : Énoncé créé à partir d'un sujet (ressource décrite), d'un prédicat (propriété de la ressource ou relation) et d'un objet (donnée ou autre ressource).

Uniform Resource Identifier (URI) : Suite alphanumérique qui identifie de manière univoque une ressource physique ou abstraite.

URI déréférençable : URI qui se définit principalement par le fait qu'il met en pratique le mécanisme de négociation de contenu.

Vocabulaire : Schéma permettant l'organisation des connaissances de manière standardisée afin de faciliter la recherche d'information.

Web 2.0 : Mouvement découlant du World Wide Web misant sur la participation des usagers, la création de contenu et l'interopérabilité.

Web de données : Application du Web sémantique visant la publication de données en format RDF afin de permettre la création de liens entre celles-ci et leur compréhension par la machine. On parle aussi de « données liées ».

Web sémantique : Extension du Web tel qu'on le connaît basée sur des standards définis par le W3C qui vise la publication de données dans des formats standardisés afin de les exploiter de manière plus efficace.

World Wide Web Consortium (W3C) : Organisme de standardisation pour le Web.

SOURCES CONSULTÉES

- Alemu, Getaneh, Brett Stevens, Penny Ross & Jane Chandler. 2012. Linked data for libraries: Benefits of a conceptual shift from library-specific record structures to RDF-based data models. *New Library World* 113 (11/12): 549-570.
- Baker, Thomas *et al.* 2011. Library linked data incubator group final report. Consulté le 13 juillet 2017. <www.w3.org/2005/Incubator/ld/XGR-ld-20111025/>.
- Bermès, Emmanuelle. 2013. Enabling your catalogue for the semantic web. In *The future of the library catalogue*, sous la direction de Sally Chambers. Chicago: Neal-Schuman, an imprint of the American Library Association, 117-142.
- Bermès, Emmanuelle, Antoine Isaac & Gautier Poupeau. 2013. *Le web sémantique en bibliothèque*. Paris: Électre/Éditions du Cercle de la Librairie.
- Berners-Lee, Tim. 2006. Linked data. Consulté le 13 juillet 2017. <www.w3.org/DesignIssues/LinkedData.html>.
- Berners-Lee, Tim. 2010. Is your linked open data 5 star? Consulté le 13 juillet 2017. <www.w3.org/DesignIssues/LinkedData.html>.
- Berners-Lee, Tim, Robert Cailliau, Ari Luotonen, Henri Frystyk Nielsen & Arthur Secret. 1994. The world-wide web. *Communications of the ACM* 37 (7): 76-82.
- Berners-Lee, Tim, Jim Hendler & Ora Lassila. 2001. *The semantic web*. Scientific American 284 (5): 28-37.
- Bizer, Christian & Tom Heath. 2011. *Linked data: Evolving the web into a global data space*. San Rafael, Californie: Morgan & Claypool.
- Coyle, Karen. 2010. Library data in a modern context. *Library Technology Reports* 46 (1): 5-13.
- DeWeese, Keith P. & Dan Segal. 2015. *Libraries and the semantic web*. San Rafael, Californie: Morgan & Claypool.
- Gonzales, Brighid M. 2014. Linking libraries to the web: Linked data and the future of the bibliographic record. *Information Technology and Libraries* 33 (4): 10-22.
- Hyvönen, Eero. 2012. *Publishing and using cultural heritage linked data on the semantic web*. San Rafael, Californie: Morgan & Claypool.
- Leroux, Éric *et al.* 2009. Les professions et les institutions. In *Introduction aux sciences de l'information*, sous la direction de Jean-Michel Salaün et Clément Arsenault. Montréal: Presses de l'Université de Montréal, 16-52.
- Rahm, Erhard & Hong Hai Do. 2000. Data cleaning: Problems and current approaches. *IEEE Data Engineering Bulletin* 23 (4): 3-13.
- Rao, K. Nageswara & K. H. Babu. 2001. Role of librarian in internet and world wide web environment. *Information science* 4 (1): 25-34.
- Shadbolt, Nigel, Tim Berners-Lee & Wendy Hall. 2006. The semantic web revisited. *Intelligent Systems, IEEE* 21 (3): 96-101.
- Stuart, David. 2011. *Facilitating access to the web of data: A guide for librarians*. Londres: Facet.
- Tillett, Barbara. 2013. RDA and the semantic web, linked data environment. *JLIS.it* 4 (1): 139-145.
- Van Hooland, Seth & Ruben Verborgh. 2014. *Linked data for libraries, archives and museums: How to clean, link and publish your meta-data*. Chicago: Neal-Schuman, an imprint of the American Library Association.
- Villazón-Terrazas, Boris, Luis Vilches-Blázquez, Oscar Corcho & Asunción Gómez-Pérez. 2011. Methodological guidelines for publishing government linked data. In *Linking government data*, sous la direction de David Wood. New York: Springer, 27-49.
- W3C. 2014. Best practices for publishing linked data. Consulté le 13 juillet 2017. <www.w3.org/TR/ld-bp/>.
- Zengenene, Dydimus, Vittore Casarosa & Carlo Meghini. 2014. Towards a methodology for publishing library linked data. In *Bridging between cultural heritage institutions*, sous la direction de Tiziana Catarci, Nicola Ferro & Antonella Poggi. Berlin: Springer, 81-92.

TECHNOLOGIES D'EXPLOITATION DU *BIG DATA* DANS LES ORGANISATIONS ET TRANSFORMATIONS ORGANISATIONNELLES : UNE ÉTUDE DE CAS AU SEIN DU SERVICE DE SANTÉ DES ARMÉES FRANÇAISES

Marc Tanti, Ph. D.

Adjoint au chef de service de veille sanitaire du Centre d'épidémiologie et de santé publique des armées (CESPA)

Chef de projets de recherches en sciences de l'information CESPA – UMR 912-SESSTIM-INSELM-IRD-AMU

Chercheur associé à l'IRSIC-EA-4262

mtanti@gmx.fr

RÉSUMÉ | ABSTRACT

L'exploitation des données sanitaires du *big data* via des technologies de collecte, de visualisation et de communication a permis au Service de santé des armées françaises de créer une valeur stratégique. Par exemple, après le tsunami de 2004 en Asie, cette exploitation a permis d'anticiper les épidémies pour les forces envoyées en soutien humanitaire. Plus récemment, dans le contexte de l'épidémie d'Ebola en Afrique, des vies humaines ont pu être préservées. L'appropriation de tels outils a également entraîné des bouleversements dans l'organisation et la culture organisationnelle, notamment un découplage spatio-temporel des activités et un gain de coordination.

Operating Technologies of Big Data in Organisations and Organisational Change: A Case Study of the Service de santé des armées française

The exploitation of health-related data of big data using collection, visualisation and communication technologies has enabled the *Service de santé des armées française* to produce strategic value. For example, following the 2004 tsunami in Asia, this exploitation enabled the humanitarian resources to better anticipate epidemics. More recently, during the Ebola epidemic in Africa, human lives were saved. The mastery of such tools also brought about important changes within the organisation and its culture, namely the spatiotemporal decompartmentalisation of activities and improved coordination.

Problème de recherche

Ces dernières années, le nombre de données circulant sur Internet a littéralement explosé. Ce phénomène de données massives, ou *big data*, amène les organisations à se positionner sur l'exploitation de tels volumes d'informations et notamment à mettre en place des processus organisationnels pour les aider à maîtriser cette pléthore d'informations à des fins d'intelligence pour la décision et l'action (Cohen 2013).

Ainsi, face à l'ampleur de ce phénomène, lié notamment à la prolifération des données des réseaux sociaux, téléphones intelligents, applications mobiles, etc., leur croissance exponentielle et les problèmes de repérage, d'accès, de gestion et de traitement de ces données, un certain nombre d'organisations, entre autres des entreprises du domaine privé, se sont engagées dans des travaux d'exploitation de ces données massives (Karoui 2014).

Par exemple, dans le domaine de la finance, notamment des banques et des assurances, les projets d'exploitations des données du *big data* permettent d'aider aux prises de décisions concernant les données à conserver dans le cadre réglementaire (Karoui 2014). Dans le secteur du *trading*, des projets d'exploitation des données massives ont été conçus pour identifier les mouvements de fonds suspects. C'est le cas par exemple pour l'entreprise BourseDirect (Karoui 2014). L'exploitation de ces données peut permettre également de créer de nouveaux services pour les clients. Dans ce cadre, Le Crédit Mutuel Arkéa a été primé en 2013 pour l'exploitation de ses données d'historiques d'opérations clients pour la création de nouveaux services aux usagers (Karoui 2014).

Le domaine médical est également un domaine où les données sont devenues pléthoriques sur Internet (Raghupathi 2014). Le secteur sanitaire est particulièrement touché, surtout en raison de l'émergence croissante des données

épidémiques et des données issues du séquençage des génomes. Dans ce secteur, les données accumulées sont chiffrées, analysées statistiquement (Alles 2014), mais elles sont aussi non chiffrées, non quantifiables, non structurées, de type socioculturel, par exemple.

Les organisations de santé publique portent une attention particulière à ces différentes formes de données, entre autres pour orienter les politiques décisionnelles. Ces organisations recherchent aujourd'hui des méthodes et des stratégies afin de les «faire parler» (Hamel 2013), notamment pour apporter une valeur ajoutée en matière de performance organisationnelle, de prise de décision, de stratégies (Brasseur 2013 ; Vayre 2014) ou d'amélioration de l'action publique (Hoppe 2015).

Le présent article a pour objectif de présenter le dispositif de veille sanitaire développé par l'organisation militaire française, considéré comme un dispositif infocommunicationnel. Selon Couzinet (2011), un dispositif infocommunicationnel est un lieu où humains, objets matériels et liens s'organisent pour mettre en œuvre des interactions. Un tel dispositif, dédié à la «mise en commun d'informations et au partage des savoirs», se compose «d'acteurs, de techniques et d'objets matériels en interaction permanente et dans un contexte défini, tous liés entre eux» (Couzinet 2011).

Dans le dispositif étudié, des mégadonnées de l'Internet sanitaire sont exploitées dans le domaine de la santé publique militaire (Boutin 2004). Dans ce cadre, notre article présente et interroge les technologies mises en œuvre pour collecter, analyser et communiquer ces données massives.

À partir de cette recherche-action, les enjeux organisationnels soulevés par l'exploitation de tels volumes d'informations sont également contestés, notamment les transformations organisationnelles induites (Parks 2014 ; Béranger 2016). L'article a également pour objectif de réfléchir aux apports en matière de construction des connaissances et de création de valeurs.

Cadre théorique

Big data

Définition

Selon Babinet (2015), le *big data* (ou mégadonnées ou données massives) désigne l'ensemble des données numériques produites par l'utilisation des nouvelles technologies à des fins personnelles ou professionnelles. Cela recoupe les données d'entreprise (courriels, documents, bases de données, historiques de processeurs métiers...) aussi bien que des données issues de capteurs, des contenus publiés sur le

Web (images, vidéos, sons, textes), des transactions de commerce électronique, des échanges sur les réseaux sociaux, des données transmises par les objets connectés (étiquettes électroniques, compteurs intelligents, téléphones intelligents...), des données géolocalisées.

Le *big data* correspond à une évolution de la *business intelligence* qui repose sur des entrepôts de données limités en taille (quelques téraoctets) et gérant difficilement des données non structurées et des analyses en temps réel. L'avènement du *big data* ouvre une nouvelle ère technologique qui offre des architectures et des infrastructures évoluées qui permettent en particulier des analyses sophistiquées, prenant en compte ces nouvelles données intégrées à l'écosystème de l'entreprise (Chen 2014).

Le *big data* peut se définir par trois composantes (Delort 2015) :

- Son volume : on le définit généralement à partir de cinq téraoctets de données à traiter, ce que les logiciels ou les bases de données classiques ne peuvent faire ;
- Sa variété : les données acquises sont brutes ou structurées, au format texte ou image, avec des propriétaires et des droits d'utilisation aussi différents que leurs sources ;
- Sa vitesse : il faut être capable d'intégrer en temps réel les dernières données disponibles et les lier aux autres jeux de données sans recommencer une analyse complète à chaque cycle.

Un quatrième «V» pour véracité peut être rajouté à cette définition pour évoquer la nécessité de vérifier la crédibilité des sources et la qualité des contenus afin de permettre une exploitation des données (Marr 2015).

Ces différentes composantes du *big data* nécessitent donc de nouvelles formes d'outils et de traitement de l'information (Pouyllau 2013).

Technologies

Selon Cointot, le *big data* recouvre l'ensemble des technologies, métiers, approches conceptuelles permettant d'exploiter l'ensemble des données générées par les hommes de façon consciente ou non et par tous les objets connectés ou non (Cointot 2014).

Les technologies qui permettent l'exploitation de ces données au volume important présentent quelques points particuliers, notamment pour permettre la collecte et la «fouille» dans ces données de grand volume. Ainsi, pour arriver à retrouver une épingle dans une meule de foin multidimensionnelle, sans savoir à quoi ressemble l'épingle ni si la meule de foin en contient une, des outils statistiques, de

L'avènement du big data ouvre une nouvelle ère technologique qui offre des architectures et des infrastructures évoluées qui permettent en particulier des analyses sophistiquées, prenant en compte ces nouvelles données intégrées à l'écosystème de l'entreprise.

classification et de visualisation des données sont nécessaires (Weinstein 2013).

Dans ce cadre, un certain nombre de solutions de visualisation de données du *big data* et plus largement de *data-scapes* du Web (paysages de données) sont aujourd'hui utilisables, notamment par des non-experts. Les topologies du Web peuvent ainsi être générées par des logiciels de calculs de graphes à partir de moissonnage de données et de liens utilisables dans l'exploitation des données massives (Boullier 2016). C'est par exemple le cas de Gephi¹, dont nous reparlerons plus tard dans cet article.

Des capacités matérielles conséquentes pour le stockage comme pour les ressources processeurs sont également indispensables au traitement de ces données massives. Dans ce contexte, Le «*cloud*», le nuage, symbolise la multitude d'unités informatiques et accessibles depuis l'Internet pour le traitement et le stockage des données du *big data* (Delort 2015).

Enfin, dans les technologies, les Environnements Numériques de Travail (ENT) ont aussi leur importance. Ainsi, lorsque les usagers se connectent, quel que soit le lieu, ils peuvent retrouver immédiatement leur poste de travail, dossiers, documents, courriels... De la même façon, on peut préciser que l'infonuagique est aussi centrée sur la tâche à accomplir. L'utilisateur ne se préoccupe ainsi plus des types d'applications à utiliser, mais bien de la tâche à réaliser (Delort 2015).

En conclusion, la technologie occupe une place centrale dans le développement et la transformation des organisations. Les relations entre les deux (technologie/organisation) ont d'ailleurs fait l'objet de nombreux travaux (Markus 1988; Orlikowski 1992; Orlikowski 2000; Kefi 2004). Cette approche qualifiée d'ailleurs de «*déterminisme technologique*» repose sur de nombreuses disciplines telles que la sociologie, l'économie et la gestion. L'hétérogénéité de ces approches permet, selon MacKenzie (1999), d'extraire plusieurs caractéristiques. Notamment, la technologie est considérée comme un artefact doté d'un ensemble de caractéristiques matérielles dont le développement est structuré par des lois scientifiques extérieures à la sphère organisationnelle, mais dont les effets sont appréhendés en termes d'impacts sur les usages et l'organisation (MacKenzie 1999).

Transformations organisationnelles

Le paradigme du *big data* n'est donc pas que technologique. Il a ainsi imposé à l'entreprise des transformations organisationnelles pour canaliser ce flux incessant de données et pour en tirer l'information nécessaire à la prise de décision (Cointot 2014; Delort 2015).

Selon Ferguson, le vrai problème est la conduite du changement dans la mesure où le vrai défi pour les organisations est d'arriver à faire les changements nécessaires au niveau des processus et des ressources humaines pour pouvoir mettre en place les initiatives analytiques que requiert un projet d'exploitation des données du *big data* (Ferguson 2013).

Selon une étude du CIGREF: «*Le big data est une composante de la transformation de l'entreprise, il est transversal et touche tous les métiers. En ce sens, le big data n'est pas un projet SI: c'est une manière nouvelle de penser et d'appréhender l'information. Il s'agit donc davantage d'une (r)évolution culturelle et technologique que d'un nouveau sujet SI.*» (CIGREF 2013)

Le phénomène du *big data* pour les entreprises recouvre ainsi deux réalités: d'une part cette explosion des données de façon continue, d'autre part la capacité technologique de traiter et d'analyser cette grande masse de données pour en tirer un profit (CIGREF 2013). Grâce au *big data*, les entreprises peuvent désormais gérer et traiter des données massives pour en extraire de la valeur, décider et agir en temps réel (CIGREF 2013). L'enjeu central pour elles est bien «*d'améliorer l'efficacité des prises de décision par l'exploitation d'informations protéiformes*» (CIGREF 2013). Elles deviennent ainsi plus réactives et plus concurrentielles. De fait, la capacité à traiter de grandes masses de données, à tisser des liens et corrélations entre des informations hétérogènes et à faire communiquer entre elles des données structurées ou non, ouvre la voie à des traitements probabilistes qui permettent l'amélioration des opérations managériales et la conception de produits et services innovants pour les clients (CIGREF 2013). Aussi, aujourd'hui, le *big data* est corrélé à l'avantage concurrentiel des entreprises et représente pour elles un atout considérable (CIGREF 2013).

Cette appropriation d'outils, de technologies devenues incontournables pour conserver un avantage concurrentiel impose ainsi à l'entreprise de nouvelles exigences en matière de compétences, d'autonomie et de capacité d'adaptation des travailleurs dont l'efficacité est censée croître (Benraiss 2005). Cette appropriation d'outils agit sur l'organisation du travail, impose de nouvelles méthodes, de nouvelles pratiques et affecte les attitudes et les comportements (Benraiss 2005).

Ces bouleversements liés à ces nouveaux usages touchent la structure même de l'entreprise qui se modifie, «*s'horizontalise*» (Lejeune 2010). La hiérarchie devient ainsi plus informelle. La communication se fait plus librement, tant sur le plan horizontal que vertical. L'information circule plus rapidement et plus efficacement. Le dialogue social se modifie. L'organisation économique subit également des modifications. Le changement ne se situe donc plus uniquement sur le plan professionnel ou organisationnel. Il se manifeste dans de nouveaux rapports sociaux et

1. <gephi.org/>.

économiques qui découlent de l'appropriation sociale des technologies (Lejeune 2010). La culture du travail et la relation au travail ne sont pas non plus épargnées et subissent des bouleversements profonds (Lejeune 2010).

Silva a réalisé une étude sur les nouvelles pratiques sociales liées à ces technologies dans les grandes entreprises (Silva 2009). Il a notamment analysé les changements de valeurs à l'origine de transformations sociales (Silva 2009). Selon cet auteur :

[...] les TIC favorisent les fragmentations des temporalités, l'éclatement des espaces et les différenciations de personnes [...]. Les rythmes et lieux de travail ne sont plus partagés par l'ensemble de l'entreprise. Il se développe, de fait, des territorialisations, c'est-à-dire des espaces temps propres à telle équipe ou à tel projet. La cohésion de l'entreprise peut être contestée [...]. Les TIC ont catalysé différents facteurs d'éclatement des groupes de salariés. En effet, elles favorisent la déconstruction des chaînes de responsabilités, dans la mesure où les messageries, par exemple, peuvent permettre de contourner certains processus hiérarchiques ou techniques, pour favoriser des liens informels qui s'étoffent [...]. L'informatisation va accompagner et augmenter cet effet de fragmentation du travail et des relations sociales dans l'entreprise. (Silva 2009)

Les TIC modifient de multiples fonctions de l'entreprise : informations décisionnelles, informations de communication, de création et de données sociales (Silva 2009 ; Silva 2010). C'est ainsi qu'émerge une nouvelle forme d'organisation du travail, centrée notamment sur les activités collaboratives qui constituent une rupture culturelle importante, notamment lorsqu'elles impliquent des personnes situées dans des lieux, des situations hiérarchiques et des temporalités différentes (Silva 2009 ; Silva 2010). Cette forme d'organisation est non plus seulement fondée sur l'individu, mais sur la capacité qu'il a de travailler avec d'autres, base de la valeur et de l'innovation pour les entreprises (Silva 2009 ; Silva 2010).

La multiplication des flux de données impacte donc directement la performance individuelle des personnes.

Construction de connaissances

Dans leur ouvrage *big data*, Mayer-Schoenberger & Cukier introduisent leur livre de cette manière :

Le phénomène des mégadonnées désigne tout ce qui peut être fait à une large échelle et non à une échelle plus réduite, afin d'extraire de nouvelles connaissances ou de créer de nouvelles formes de valeur, bouleversant ainsi les marchés,

les organismes, les relations entre citoyens et gouvernements, et bien plus. (Mayer-Schoenberger et al. 2014)

Le *big data* n'est pas en soi une discipline dédiée à la prédiction, mais plutôt à l'analyse brute de grande quantité de données (Cointot 2014). Dans bien des cas, ces analyses sont en elles-mêmes suffisantes pour permettre d'en extraire des informations de qualité dont il pourra être fait immédiatement usage (Cointot 2014). Par exemple, en médecine, des recherches de marqueurs de pathogènes dans les analyses médicales d'un grand nombre de patients peuvent permettre de savoir quelle population exposée à un environnement particulier peut être affectée d'une pathologie particulière. Cette construction de savoir peut ainsi constituer une valeur permettant la mise en place d'actions spécifiques auprès de cette population (Cointot 2014).

L'objectif est ainsi de créer, à partir de ces données quelque chose d'utile, de pertinent permettant d'accroître la connaissance (Hyeans 2016).

L'objectif, à partir de l'exploitation de ces données massives, est de créer quelque chose d'utile, de pertinent et permettant d'accroître la connaissance pour la prise de décision et l'action.

Pour Le Coadic (2004) : « La connaissance est le résultat de l'acte de connaître, acte par lequel l'esprit humain saisit un objet. Connaître, c'est être capable de former l'idée de quelque chose, c'est avoir présent à l'esprit. »

Selon cet auteur, « la connaissance peut aller jusqu'à la compréhension exacte et complète des objets [...] dans le but de permettre une adaptation à l'environnement (naturel et humain) » (Le Coadic 2004).

La connaissance possède quelque chose de supérieur. En effet, selon Foray (2009), elle donne à son détenteur une capacité d'action intellectuelle et physique.

Création de valeurs

Outre les quatre composantes décrites précédemment – volume, variété, vitesse, véracité (Delort 2015), une cinquième composante peut également définir le *big data* : la création de valeurs (Marr 2015 ; Babinet 2015). En effet, l'un des aspects notoires des *data*, c'est que l'on ne connaît pas nécessairement a priori le trésor qui se cache en son sein. L'objectif, à partir de l'exploitation de ces données massives, est de créer quelque chose d'utile, de pertinent et permettant d'accroître la connaissance pour la prise de décision et l'action (Babinet 2015).

La création de valeurs est un concept suscitant aujourd'hui un intérêt croissant dans différents domaines des sciences de gestion : management stratégique, finance d'entreprise, comptabilité, contrôle de gestion, organisation, marketing.

Bourguignon distingue trois acceptions de la valeur : la valeur au sens de mesure (en particulier dans les disciplines scientifiques que sont les mathématiques et la physique), la valeur au sens économique et la valeur au sens philosophique (Bourguignon 1998). Le terme de valeur est synonyme de celui de richesse. Le thème de la valeur fait ainsi l'objet de regards multiples ou de paradigmes, c'est-à-dire de visions communes aux membres d'un groupe donné (Kuhn 1983). La problématique de la valeur renvoie donc à la question des destinataires de la valeur créée : pour qui créer de la valeur ? Dans le cadre de la finance d'entreprise, la valeur est souvent une valeur financière pour l'actionnaire.

Créer de la valeur en économie, c'est ainsi la faire varier dans le sens de la hausse. A contrario, détruire de la valeur, c'est la faire baisser au cours du temps. La création de la valeur économique est au cœur de l'activité des organisations et au centre de leur vocation, de leur raison d'être et de leur stratégie (Savall 2008).

La notion de création de valeurs est elle-même polymorphe, car elle génère de nombreux indicateurs qui prennent essentiellement appui sur des bases comptables, boursières ou économiques. De ce fait, dans l'entreprise, la création de la valeur s'apprécie au niveau de l'actionnaire (Elidrissi 2010). Une entreprise crée de la valeur si le résultat dégagé par l'exploitation est supérieur au coût des capitaux investis. Elle s'apprécie aussi au travers les gains réalisés au-delà du coût du capital investi.

Un investissement en systèmes d'information (SI) peut générer trois types de gains pour l'entreprise. Elle peut générer des gains mesurables en unités monétaires. Elle peut générer des gains estimables en unités non monétaires (gain de volume, temps, espace, performance, stratégique, etc.) (Elidrissi 2010). Elle peut aussi générer des gains peu tangibles (autonomie de l'utilisateur de l'information, satisfaction des clients, meilleure collaboration en interne, accumulation de l'expertise technologique, etc.) (Elidrissi 2010).

Dans le cas des organisations publiques, comme dans notre étude, la création de la valeur ne peut a priori s'estimer qu'en unités non monétaires et qu'en gains peu tangibles, notamment qu'en termes de gain stratégique et décisionnel, car elles n'ont pas vocation à faire de profits.

La performance est aussi très liée au concept de création de valeurs afin de permettre la prise de décisions managériales, stratégiques idoines et adéquates d'une part, et de rendre l'organisation concurrentielle. Certains auteurs l'assimilent à l'efficacité, à la capacité ou à la compétitivité, d'autres à l'efficience, au rendement, à la productivité, et d'autres, enfin lui associent des notions telles que la santé, la réussite, le succès et l'excellence (Little 1973 ; Bourguignon 1995 ; Tchankam 2000). Par exemple, Little (1973) l'assimile comme la mise en œuvre et sa réalisation d'une action. Tchankam (2000) définit l'entreprise performante

comme celle qui fait mieux que ses concurrents sur le court, le moyen et le long terme, dans l'idéal d'un ensemble de paramètres définissant la performance, au minimum sur ceux jugés être les plus significatifs. Bourguignon (1995) pense que la définition de la performance est corrélée à l'appréciation du marché : part de marché, chiffre d'affaires, nombre de créances douteuses ou de plaintes de la clientèle dues à la qualité des produits.

La contribution originale des pionniers à cette dimension de la performance réside dans sa relation au concept d'innovation, processus managérial consistant à innover, c'est-à-dire à chercher à améliorer constamment l'existant, un concept souvent au cœur des préoccupations stratégiques des entreprises (Drucker 1957).

Méthode

L'analyse du dispositif s'est d'abord effectuée par une observation non participante au sein du service chargé de l'exploitation des données du *big data* sanitaire dans les armées (Tanti 2010). Dans le cadre de cette approche globale qualitative, l'observation non participante a permis, par le fait que l'observateur n'est pas membre du dispositif et par le fait qu'il reste en retrait durant les observations, de garder une part d'objectivité dans les analyses, constituant ainsi un outil de valeur dans l'analyse de l'objet de recherche.

Cette observation non participante s'est déroulée durant trois mois en 2014. Elle a ainsi permis une meilleure compréhension des acteurs du dispositif, de leurs rôles dans l'organisation, de leurs modes de travail, de leurs modes de communications, ainsi que des objets informationnels et techniques produits et mis en jeu dans l'organisation, notamment des produits documentaires et des systèmes d'information développés.

L'étude s'est poursuivie par des entretiens de ces acteurs rendus possibles par l'analyse des données récoltées antérieurement. Cette enquête, également réalisée en 2014 auprès des acteurs du dispositif (avec un total de six participants interrogés), a consisté en des entretiens semi-directifs, avec grilles. Ces entretiens ont ainsi permis une connaissance plus précise des interlocuteurs, l'expression de leurs opinions, des expériences vécues, une meilleure appréhension des outils de traitements de données, de leurs modes de partage et de production de la connaissance à l'échelle individuelle et collective, des valeurs créées et des modalités de fonctionnement des équipes, des technologies et de leurs usages, de l'organisation et des collaborations.

Enfin, une analyse du dispositif sous le prisme des concepts mobilisés dans le cadre théorique de cet article a été effectuée, notamment l'analyse des aspects construction de connaissances et création de valeur du dispositif.

Résultats

Technologies d'exploitation des données massives

Sur le Web, il existe encore plus de données que d'étoiles dans l'univers. Face à cette croissance exponentielle et afin d'exploiter ces données pour en retirer l'information sanitaire utile à la prise de décision, le dispositif utilise des outils de collecte automatisée et semi-automatisée et des outils d'aide à l'analyse et de classification/visualisation de données.

Outils de collecte automatisée ou semi-automatisée

Le dispositif collecte ainsi de manière automatique les données bibliographiques du *big data* médical à partir du service MyNCBI du logiciel PubMed (Bussi res 2003). Ce service permet l'interrogation de la base de données bibliographique en sciences de la sant  Medline (*MEDical Literature Analysis and Retrieval System on LINE*) produite par la National Library of Medicine (NLM) aux  .U. Il permet la collecte automatis e de donn es massives de cette base   partir d'un profil et d' quations de recherches d finies. Les donn es sont extraites selon ce profil et sont adress es automatiquement par courriel   l'organisation soit imm diatement d s leur parution, soit selon une p riodicite  d finie en fonction des th matiques (Eveillard 2012). Par exemple, l'organisation suit avec attention, via cet outil, les derni res donn es de recherches sur les virus Chikungunya, Ebola et Zika.

Le syst me collecte  galement de mani re automatique et en temps r el les donn es  pid miques   partir du programme international de veille  pid miologique sur les maladies  mergentes PromMED-mail (Cowen 2006). Ce service permet notamment la r ception automatique d'alertes  pid miques dans la bo te aux lettres  lectronique. Dans ce cadre, lors des derni res  pid mies de grippe A (H1N1), cet outil a permis un suivi quotidien des donn es relatives   l' volution du nombre de cas   l' chelle mondiale.

  partir d'Europresse.com² et de Factiva³ et avec diff rentes  quations de recherche, le dispositif collecte automatiquement les donn es sanitaires et  pid miques issues de plus 80 m dias fran ais nationaux et internationaux. Par exemple, il extrait les donn es des m dias fran ais LeMonde.fr ou Lefigaro.fr et des m dias internationaux comme elwatan.com, theguardian.com ou nytimes.com.   titre d'exemple, le dispositif a pu effectuer un suivi des rumeurs  pid miques sur ces m dias lors de les derni res  pid mies d'Ebola (2014-2015) et de Zika (2016). Une extraction des donn es

informelles et de la rumeur est  galement effectu e   partir du Web social, notamment de Twitter⁴ et des forums de discussions comme Doctissimo.fr⁵.

Gr ce au moteur de recherche Pickanews⁶ et via le moteur de recherche d'Europresse, l'organisation effectue un suivi quantitatif en rapport avec ses mots-cl s d'int r ts (Chikungunya, Ebola, Zika...) sur les diff rents m dias et sur Twitter.

Le dispositif utilise  galement des outils d'alertes⁷ pour surveiller les donn es sanitaires en rapport avec les th matiques d'int r ts sur le Web.

Outils d'aide   l'analyse, de visualisation et de classification des donn es

Le dispositif fait usage d'un certain nombre d'outils d'aide   l'analyse permettant une visualisation des contenus et la classification de donn es.

Ainsi, le moteur de recherche Pickanews, au-del  de permettre un suivi quantitatif de l' volution des donn es d'int r ts pour la veille sur les m dias et Twitter, permet aussi de rassembler ces donn es, de les classer et de les analyser par type de presse, m dias ou date. Ce moteur permet de surveiller les mots-cl s d finis et d'acc der aux articles identifi s.

Le logiciel bluenod⁸ permet une analyse des influenceurs et de leurs r seaux sur Twitter. Il permet en particulier de g rer directement les listes de ce r seau social, d'exporter des donn es et de cibler les communaut s li es aux sujets. Enfin, cet outil construit des repr sentations cartographiques de l'information et classe les donn es. Il a permis par exemple de d terminer quels  taient les protagonistes intervenant dans le champ de l' pid mie d'Ebola 2014-2015 et l' volution de leurs discours. Le logiciel MapDTweetmap⁹ est  galement utilis  dans le dispositif pour g olocaliser sur Twitter les  changes de *tweets* et d terminer les meilleurs *hashtags*.

L'outil Gephi¹⁰, dont nous avons parl  plus haut, permet la visualisation et la cartographie de donn es sur Twitter, et l'analyse de r seaux, notamment de l'activit  en r action   des  v nements. Par exemple, dans le cadre de l' pid mie d'Ebola 2014-2015, il a permis un suivi des r seaux constitu s en r action aux pics  pid miques et l'importation de cas.

4. <twitter.com>.

5. <www.doctissimo.fr>.

6. <www.pickanews.com>.

7. Google Alertes : <www.google.fr/alerts>.

8. <bluenod.com>.

9. <www.mapd.com/demos/tweetmap/>.

10. <gephi.org/>.

2. <www.europresse.com/fr>.

3. <www.dowjones.com/products/factiva>.

Enfin, l'outil AISummarizer¹¹ est un outil d'analyse linguistique utilisé dans le dispositif pour faire des résumés multilingues concis de pages Web et de documents. Dans ce cadre, il trouve son intérêt dans l'aide à l'analyse de données épidémiques de l'OMS ou de Santé publique France.

Outils de diffusion et de communication des données

Les données collectées, classées et synthétisées sont finalement mises en forme et diffusées en temps réel ou quasi réel sur une plateforme intranet sécurisée, dans un espace de travail collaboratif avec un Environnement Numérique de Travail (ENT) dédié au dispositif (Boutin 2004; Tanti 2010).

Il faut noter les capacités conséquentes de stockage sur des serveurs sécurisés pour le traitement de ces données massives, ainsi que des ressources processeurs élevées.

Les données captées sont diffusées dans cette plateforme principalement sous forme hypertextuelle, permettant de naviguer d'une unité d'information à une autre. La navigation peut aussi se faire via un moteur de recherche (Tanti 2010). Il existe différents profils d'utilisateurs, principalement des décideurs du Service de santé des armées (SSA)

et de l'État-major des armées (EMA), des experts militaires en santé et des médecins d'unités. Trois systèmes d'information (SI) hébergés sur cette plateforme ont ainsi été développés pour répondre aux besoins informationnels de ces différents usagers.

Le premier SI a été dénommé BEDOUIN, pour Banque Épidémiologique de Données sur l'OUTre-mer et la zone INter-tropicale (Figure 1). Il diffuse les données filtrées et analysées sur les risques sanitaires pour les forces hors de France issues du *big data*. Le contenu est architecturé par pays.

Le deuxième SI s'appelle REDUVES pour base de données en REcherches et Développements de l'Unité de VEille Sanitaire (Figure 2). Il diffuse les données scientifiques du *big data* sur les agents du risque biologique (en matière de nouveaux traitements, diagnostics...). Dans ce cadre, les contenus captés et filtrés sont architecturés par risques et par agents.

Le dernier SI s'appelle INTRACESPA, pour Intranet du CESPA (Figure 3). Il diffuse différentes données internes au SSA, principalement relatives aux investigations d'épidémies et aux programmes militaires de santé publique. Les données sont consultables par auteur de publication, thème, en mode multicritère, en recherche simple/avancée et par navigation arborescente.

11. <aisummarizer.com/summarizer/index.jsp?ui.lang=fr>.

FIGURE 1

Page d'accueil du SI BEDOUIN

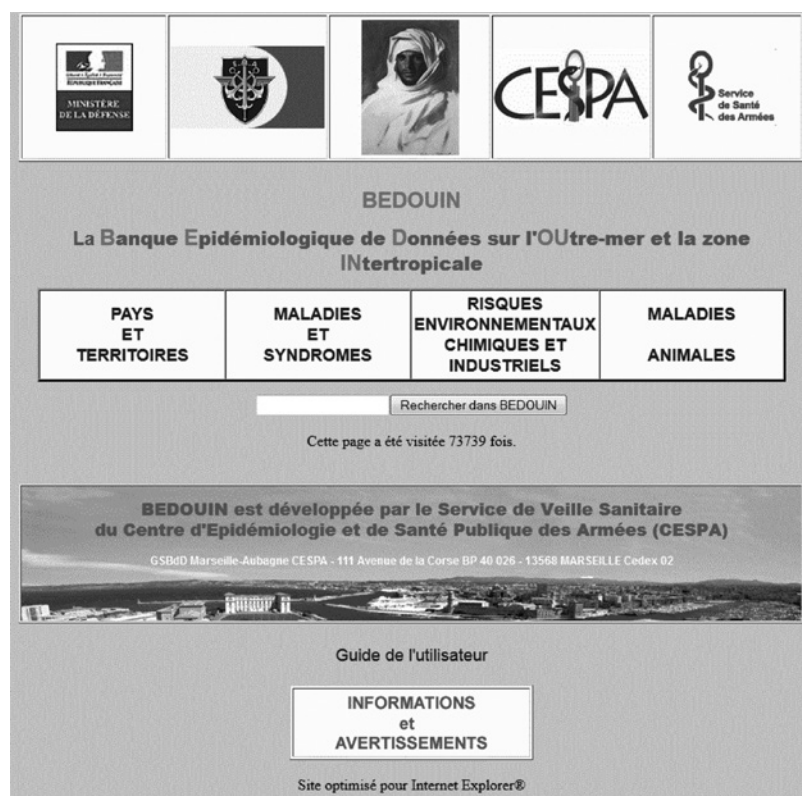


FIGURE 2

Page d'accueil du SI REDUVES

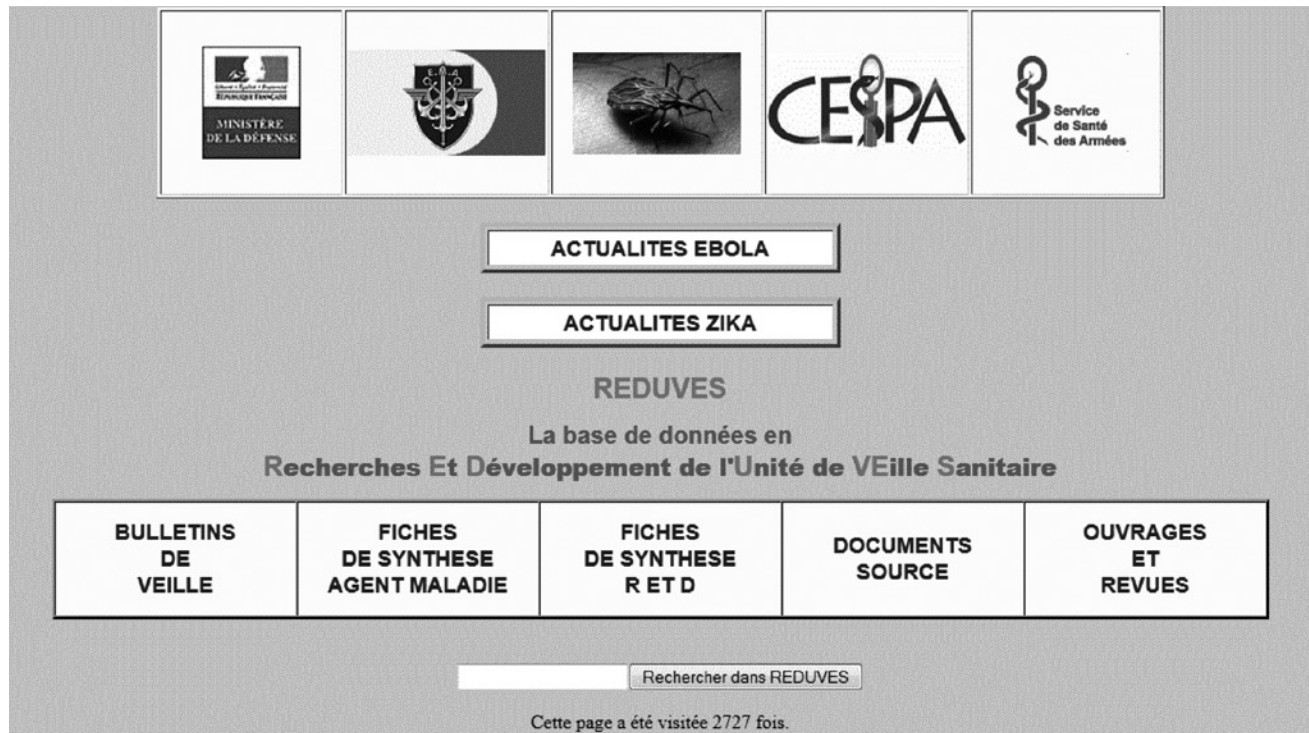
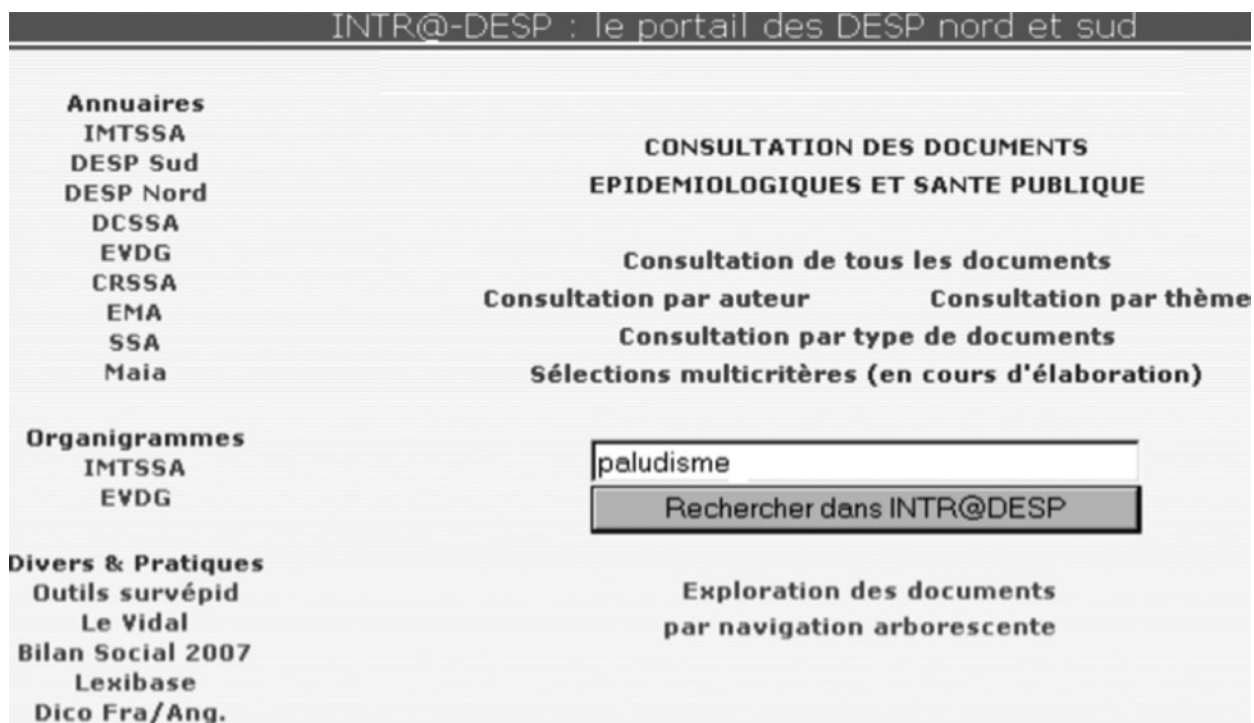


FIGURE 3

Page d'accueil du SI intraCespa



Dans cette plateforme intranet, outre les SI diffusant et stockant les données, les usagers peuvent communiquer entre eux, partager les informations, les commenter, se les réapproprier via des outils de travail collaboratif et un Environnement Numérique de Travail (ENT). La communication dans cet ENT se fait de manière asynchrone par messagerie électronique Outlook, qui tend à devenir un moyen de communication instantanée. À cette messagerie, un annuaire des experts, un agenda électronique et un gestionnaire de communautés de « projets » sont associés pour permettre une réappropriation et un partage collectif des données et la création d'intelligence, de nouvelles connaissances et de la valeur.

Transformations organisationnelles

Les entretiens avec les acteurs du dispositif ont révélé que les technologies de collecte, d'aide à l'analyse et de diffusion/communication des données utilisées dans l'exploitation du *big data* sanitaire ont induit des transformations dans l'organisation étudiée.

Transformations induites des outils de collecte et d'analyse

Les outils de collecte automatisée, de classification de données et d'aide à l'analyse ont permis à l'organisation un gain considérable de temps dans la gestion et la réalisation des tâches pour atteindre les objectifs définis. Les outils se sont ainsi mis au service des humains.

Un questionnaire sur l'impact des usages des outils de collecte a été mené, notamment pour déterminer les nouveaux échanges et les nouvelles sociabilités induites. Il en ressort que les abonnements aux listes de diffusion professionnelles (PromMED-mail, MyNCBI) ont généré de nouvelles formes de sociabilités dans la mission de veille sanitaire de défense en créant de nouvelles communautés professionnelles virtuelles au service du partage et de la gestion de la même connaissance (Soulier 2004). De nouvelles activités cognitives ont été envisagées : la création, l'échange coopératif volontaire et la réutilisation des connaissances entre les professionnels du réseau (Lefebvre 2004). La communauté de pratiques qui est née a fait émerger une nouvelle unité de coordination qui a permis un traitement plus efficace des connaissances au profit de la mission de veille (Michaux 2004).

La mutualisation de ces outils et cette réappropriation collective des données ont ainsi permis de renforcer le « *coworking* » et le travail collaboratif. Le travail individuel et donc

collectif est devenu plus efficace. L'acteur de l'exploitation des données à son échelle a ainsi pu percevoir de manière plus fine son utilité à l'échelle collective.

Les outils, par leur mutualisation et leur réappropriation collective, ont également permis l'extraction de données « cachées » menant à une amélioration de la créativité, de l'innovation, de l'activité intellectuelle, de l'intérêt au travail et de ses conditions d'exercice, raison d'être du travail humain. Une transformation organisationnelle en a été induite, notamment en gain d'efficacité, de créativité, d'innovation, de productivité et de compétences individuelles et collectives.

Dans les transformations, il est important de noter que les outils de collecte et d'aide à l'analyse soutiennent les acteurs de l'exploitation et le dispositif face à un flux de mégadonnées de qualité variable, mais de quantité pléthorique. Ils permettent au dispositif, comme nous l'avons dit, un gain de temps dans la collecte, la sélection, la classification, la mise en forme, la visualisation des données. Mais ils ne remplacent pas l'analyse humaine et l'expertise qui demeurent indispensables ! Dans ce dispositif, il existe un réseau d'experts (médecins, scientifiques, vétérinaires...) qui exploite, interprète de manière intellectuelle la masse de données collectées. Malgré l'intérêt de l'ensemble des technologies, seul l'humain a la capacité de recouper les données collectées et de vérifier les erreurs de doublonage. Seul l'humain a la faculté de repérer dans les discours ce qui se fait au-delà de ce qui se dit (Lupton 1992) et d'apporter une expertise sur les échanges concernant les problématiques questionnées. Seul l'humain a le pouvoir d'effectuer une sélection et une validation des contenus, en fonction des événements traités, et par comparaison aux événements

[Les outils de collecte et d'aide à l'analyse] permettent au dispositif [...] un gain de temps dans la collecte, la sélection, la classification, la mise en forme, la visualisation des données. Mais ils ne remplacent pas l'analyse humaine et l'expertise qui demeurent indispensables !

antérieurs et connus. Seul l'humain peut analyser le caractère novateur d'un contenu, son intérêt, sa crédibilité, sa fiabilité, en accord avec une question déterminée au préalable par l'autorité de tutelle ou dans une approche inductive, en cherchant à contester les objets. Enfin, c'est le seul à pouvoir synthétiser et interpréter les données collectées en fonction de son expérience, selon les objectifs définis, à les mettre en forme et à les transformer en véritable information utile pour

l'utilisateur du dispositif (Leray 2008).

Transformations induites des outils de diffusion et de communication

Les outils de diffusion et de communication des données déployés sur les plateformes permettent des échanges de données en temps quasi réel entre les acteurs de

l'exploitation du *big data* et leurs usagers. Cela permet également une réappropriation collective des données entre les différents étages de l'organisation et un décroisement spatial, pour créer une intelligence collective. Les données permettent elles-mêmes de soutenir la communication dans l'organisation, car elles sont réappropriées collectivement pour créer de la valeur stratégique.

L'Environnement Numérique de Travail (ENT), de stockage de données sur l'intranet et la messagerie Outlook ont favorisé, dans l'organisation, en tant que technologie cognitive, l'émergence d'une culture numérique. De par leur efficacité technique, elles ont valorisé culturellement la pratique de la messagerie électronique et de l'ENT par l'organisation. Elles ont défini, en tant que dispositif matériel, de nouvelles formes d'expression fondées sur le réseau. À travers cette manière spécifique d'agir et de penser en réseau, elles ont instauré un nouveau rapport au savoir et à la connaissance. Elles ont valorisé une communication rapide moins formelle et ont promu les échanges immédiats, s'apparentant à de la communication synchrone.

La mise en place de la plateforme intranet qui héberge les différents SI a facilité l'accès aux flux d'informations, a accéléré l'acheminement des données, a créé de nouvelles formes de partage de la connaissance et a promu le *knowledge management*. Cette plateforme à guichet unique a modifié les relations sociales, la communication et les échanges entre les différents acteurs de l'organisation. L'autonomie des usagers en a été renforcée et les structures ont été décroisées. Des relations nouvelles en réseau entre les acteurs de l'exploitation et les utilisateurs ont été développées et redistribuées avec un effet très structurant sur l'organisation et la coordination des décisions. Le travail est devenu de plus en plus coopératif ou « collaboratif » entre tous les acteurs, entraînant notamment des effets « feedback ». Cette ouverture a représenté un facteur important de création de valeur ajoutée à tous les échelons et un phénomène qui a transformé le cadre spatiotemporel de l'activité de veille. Les notions traditionnelles de localisation, de distance et de territoire ont été modifiées. Les différentes équipes et services répartis en des lieux différents du territoire peuvent ainsi consulter les données collectées, classées et synthétisées sans contrainte de distance. Le temps et les coûts des recherches documentaires pour l'ensemble des acteurs de l'organisation en ont été réduits.

En conclusion, ces transformations organisationnelles ont elles-mêmes constitué un enjeu managérial, dans la mesure où elles ont été portées par l'équipe de cadres militaires, dans un contexte global de conduite du changement.

Construction de connaissances

Ces mégadonnées « filtrées », hiérarchisées, « clustérisées », purgées de l'infobésité et mises à disposition des usagers,

sous la forme de cette plateforme intranet dédiée, a permis des corrélations inédites de données et la construction de nouvelles connaissances.

Un certain nombre d'exemples de construction de savoirs ont ainsi été donnés en entretiens.

Par exemple, en 2014-2015, lors de l'épidémie d'Ebola, l'exploitation des données de PromMED-mail a permis le suivi de la dynamique épidémique en temps quasi réel. Cet outil de collecte automatisée a ainsi permis de révéler plus rapidement l'émergence et l'expansion de la maladie que les méthodes de collectes traditionnelles sur le terrain. La diffusion des données filtrées, débarrassées du « bruit informationnel » et classées, sur la plateforme Web, a permis aux usagers décideurs de l'EMA un suivi épidémique en quasi instantané, l'anticipation d'éventuelles menaces pour les forces et la mise en œuvre de contre-mesures médicales rapides. De plus, dans ce cadre, l'exploitation des données de génomique et de protéomique de PubMed sur la maladie d'Ebola a permis pour les usagers du système d'ouvrir la voie vers la construction de nouveaux savoirs médicaux et d'explorer des innovations pharmaceutiques, notamment vaccinales et thérapeutiques. Il a permis par exemple de détecter les essais vaccinaux en cours au profit de la population militaire. Le croisement de ces différentes données a ainsi conduit à des corrélations inédites qui ont permis à l'étage décisionnel de développer des modélisations dans une optique d'aide à la décision. Des connaissances ont ainsi été construites et ont permis d'estimer les paramètres fondamentaux de la transmission de la maladie et d'évaluer a priori l'impact des stratégies de contrôle et d'interventions militaires, en y intégrant les aspects médico-économiques.

Création de valeurs

Dans cette étude, la valeur créée est principalement d'ordre stratégique. C'est-à-dire une valeur qui contribue à diminuer l'incertitude du décideur militaire, du médecin militaire de terrain, dans le choix de ses décisions notamment en situation d'urgence sur le théâtre ou avant déploiement ou encore en cas de crises sanitaires. Cette création de valeur permet d'anticiper sur les décisions militaires et prévenir les risques pour le soldat, notamment les risques sanitaires en opération extérieure. Elle permet aussi une meilleure orientation de l'action publique.

Dans notre étude, la valeur stratégique est une valeur plus vaste que la valeur économique ou financière, puisqu'elle intègre aussi des aspects médicaux, réglementaires et décisionnels. La création de la valeur demeure au centre des préoccupations des organisations militaires. Nous la définissons comme l'augmentation de la valeur stratégique pour le décideur militaire en santé. C'est celle qui va procurer des avantages pour anticiper sur les décisions militaires, pour prévenir les risques pour le soldat, notamment les

risques sanitaires, les risques épidémiques qui ont par définition un impact sur les opérations militaires.

Un certain nombre d'exemples de création de valeur ont été donnés en entretiens.

Par exemple en 2005, des forces maritimes françaises sont envoyées en Asie du Sud-Est en soutien humanitaire après le tsunami. Un certain nombre d'épidémies touche la région, notamment une épidémie d'une maladie peu connue à l'époque, l'encéphalite japonaise, qui fait énormément de décès dans le territoire. Il y a beaucoup d'incertitudes sur cette maladie, notamment son risque de transmission aux militaires déployés. L'ensemble des données collectées, notamment à partir de ProMED-mail et de MyNCBI, ont permis de révéler des corrélations inédites. Des données à forte valeur ajoutée mettant en évidence une létalité de 5 à 40 %, pouvant atteindre les 60 %, ont ainsi pu être détectées. Des informations ont également souligné que dans les formes graves, l'infection pouvait entraîner des séquelles neurologiques sévères à des fréquences non négligeables. Il a également été décelé, dans cette masse d'informations, l'existence d'un vaccin efficace. L'ensemble de ces données visualisées par l'étage décisionnel dans la plateforme dédiée ont été réappropriées collectivement et partagées. Elles ont conduit à une décision stratégique importante : l'achat de doses vaccinales au laboratoire le fabricant et l'administration à titre préventif aux forces militaires dans la zone pour prévenir tout risque pour la santé.

Nous pouvons également reprendre l'exemple précédent de l'épidémie d'Ebola 2014-2015 pour illustrer notre propos.

Dès l'émergence des premiers cas, l'exploitation des données massives collectées, analysées, filtrées et diffusées dans le dispositif a permis aux décideurs militaires et aux experts du Service de santé des armées (SSA) de suivre en temps quasi réel la dynamique épidémique de la maladie. Ces données ont été partagées dans la plateforme par les usagers et ont été réappropriées collectivement. Elles ont permis aux utilisateurs d'avoir une connaissance permanente et fine de la menace liée à cette maladie, en particulier de mortalité élevée, d'absence de traitements et de vaccins.

Cette connaissance a été particulièrement prégnante stratégiquement, lorsque pour faire face à cette épidémie d'ampleur sans précédent, un centre de traitement dédié exclusivement au personnel des soignants d'Ebola (CTS) est créé en Guinée en 2015. En effet, dans ce contexte, des experts militaires français ont été envoyés dans ce centre pour apporter leur soutien. Les données du dispositif ont

été partagées par ces experts avant le départ et pendant toute la durée du déploiement. Par des corrélations inédites et le croisement de données venant notamment du terrain, de nouveaux savoirs ont été construits. Ces savoirs ont particulièrement été réutilisés dans le cadre de communautés et de groupes de travail « Ebola ». Ainsi, les experts ont pu détecter des innovations thérapeutiques et vaccinales qui ont fait l'objet d'un traitement et d'un partage particulier étant donné les impacts en termes de mortalité et l'absence de moyens de prévention. L'identification précoce de ces innovations a ainsi permis de créer de la valeur en anticipant et prévenant les risques pour les personnels militaires déployés dans le centre. Par le partage de ces savoirs, la performance dans la prise de décisions en a été améliorée et a ainsi permis de préserver l'état de santé des militaires projetés dans les zones à risque.

Conclusions

L'exploitation des données sanitaires du *big data* via des technologies de collecte, de visualisation et de communication a permis au SSA la création de valeurs, en particulier stratégiques. Par exemple, cette exploitation a permis, après le tsunami en Asie du Sud-Est en 2004, d'anticiper sur les épidémies pouvant toucher les forces envoyées en soutien humanitaire. Plus récemment encore, dans le cadre de l'épidémie d'Ebola 2014 en Afrique, des vies humaines ont pu être préservées. Même si le dispositif d'exploitation des données sanitaires du *big data* décrit dans cet article a été développé dans une organisation publique qui ne fait pas de profits financiers, il serait intéressant d'évaluer, de manière quantitative et économique, la performance sur les plans coût/efficacité, productivité et rentabilité liée à cette exploitation. Cela nous amènerait également à réfléchir à

L'identification précoce de ces innovations a ainsi permis de créer de la valeur en anticipant et prévenant les risques pour les personnels militaires déployés dans le centre.

ce qu'est la performance dans les organisations publiques, notamment celles du secteur de la santé publique. Enfin, il a été mis en évidence que l'appropriation des technologies a entraîné des bouleversements de la culture organisationnelle du service en matière de découplage spatio-temporel des activités, de partage des connais-

sances et de coordination du travail. Cependant, les technologies n'ont pas remplacé le traitement humain. On peut même s'interroger si l'outil remplacera un jour l'homme dans l'analyse et l'interprétation des données. Enfin, le *big data* est aujourd'hui une innovation en matière de modèle économique et social. Mais, s'agit-il tout simplement d'une évolution de la performance des outils existants ou bien d'un simple effet de mode ? Des questions qui restent ouvertes et qui méritent d'être débattues.

SOURCES CONSULTÉES

- Alles, Michael & Miklos A Vasarhelyi. 2014. Thick data : Adding context to big data to enhance auditability. *International Journal of Auditing Technology* 2 (2) : 95-108.
- Babinet, Gilles. 2015. *Big data, penser l'homme et le monde autrement*. Paris : Le Passeur Éditeur.
- Benraiss, Laila *et al.* 2005. TIC et performance des salariés : quel rôle pour la responsabilité sociale de l'entreprise ? *Revue internationale sur le travail et la société* 3(2) : 915-933.
- Béranger, Jérôme. 2016. La valeur éthique des big data en santé. *Les Cahiers du numérique* 12 : 109-132.
- Boullier, Dominique. 2016. *Sociologie du numérique*. Paris : Edition Armand Colin.
- Bourguignon, Annick. 1995. La performance, essais de définition. *Revue Française de Comptabilité* 26 : 61-66.
- Bourguignon, Annick. 1998. *Management Accounting and Value Creation : Value Yes but what Value ? Working Paper*. Paris : ESSEC.
- Boutin, Jean-Paul *et al.* 2004. Pour une veille sanitaire de défense. *Medecine et Armées* 32 (4) : 66-372.
- Brasseur, Christophe. 2013. *Enjeux et usages du big data : technologies, méthodes et mise en œuvre*. Éditions Hermès.
- Bussi res, Jean-Fran ois *et al.* 2003. Utilisation pratique et efficace de PubMed. *Qu bec pharmacie* 50 (1) : 20-22.
- Chen, Min *et al.* 2014. Big data : A survey. *Mobile Networks and Applications* 19 (2) : 171-209.
- Cigref. 2013. *Big data : la vision des entreprises : opportunit s et enjeux*. Rapport d'octobre 2013.
- Cohen, Corine. 2013. *Business intelligence : The effectiveness of strategic intelligence and its impact on the performance of organizations*. Hoboken, NJ : Wiley-ISTE.
- Cointot, Jean-Charles & Yves Eychenne. 2014. *La R volution Big data : les donn es au c ur de la transformation de l'entreprise*. Paris : Edition Dunod.
- Couzinet, Viviane. 2011. Les dispositifs : question documentaire. In *Approche de l'information-documentation : concepts fondateurs* sous la direction de C cile Gardi s, Caroline Courbi res, Viviane Couzinet *et al.* Toulouse : C padu s, 120-121.
- Cowen, Peter *et al.* 2006.  valuation of ProMED-mail as an electronic early warning system for emerging animal diseases : 1996 to 2004. *Journal of the American Veterinary Medical Association* 229 (7) : 1090-1099.
- Delort, Pierre. 2015. *Le big data*. Paris : Presses Universitaires de France.
- Drucker, Peter. 1957. *La pratique de la direction des entreprises*. Paris : Les  ditions d'Organisation.
- Elidrissi, Djamila & Ali Elidrissi. 2010. Contribution des syst mes d'information   la performance des organisations : le cas des banques. *La Revue des Sciences de Gestion* 241 : 55   61.
- Eveillard, Philippe. 2012. PubMed en 10 le ons. *La Revue du Praticien* 62 : 1428.
- Ferguson, Ren e. 2013. Big data and Big Change Management : A Path Forward. *MIT Sloan Management Review*.
- Foray, Dominique. 2009. *L' conomie de la connaissance*. Paris : La D couverte.
- Hamel, Marie-Pierre & David Marguerit. 2013. Quelles possibilit s offertes par l'analyse des big data pour am liorer les t l services publics ? *Revue fran aise d'administration publique* 146 (2) : 437-448.
- Hoppe, Magnus. 2015. Intelligence as a discipline, not just as a practice. *Journal of Intelligence Studies in Business* 5 (3) : 47-56.
- Hyeans, Andy. 2016. *Sport Data Revolution : l'analyse des donn es au service de la performance sportive*. Paris : Edition Dunod.
- Karoui, Myriam, Gr goire Devauchelle & Aur lie D dezert. 2014. Big data : mise en perspective et enjeux pour les entreprises. *Ing nierie des syst mes d'information* 19 (3) : 73-92.
- Kefi, Hajer & Michel Kalika. 2004. * valuation des syst mes d'information : une perspective organisationnelle*. Paris : Economica.
- Kuhn, Thomas. 1983. *La structure des r volutions scientifiques*. Paris : Flammarion.
- Le Coadic, Yves-Fran ois. 2004. *La science de l'information*. Paris : Presses Universitaires de France.
- Lefebvre, Philippe *et al.* 2004. Les th ories des Communaut s de Pratique   l' preuve : conditions d' mergence et organisation des communaut s. *Syst mes d'Information et management* 9 (1) : 25-48.
- Lejeune, Yannick. 2010. *TIC 2025 les grandes mutations. Comment Internet et les technologies de l'information et de la communication vont dessiner les prochaines ann es*. Paris : FYP  ditions.
- Leray, Christian & Lise Chartier. 2008. *L'Analyse de contenu : de la th orie   la pratique, la m thode Morin-Chartier*. Qu bec : Presses de l'Universit  du Qu bec.
- Little, William *et al.* 1973. *The Shorter Oxford English Dictionary*. Oxford, England : Oxford University Press.
- Lupton, Deborah. 1992. Discourse analysis : A new methodology for understanding the ideologies of health and illness. *Australian Journal of Public Health* 16 (2) : 145-150.
- MacKenzie, Donald & Jude Wajcman. 1999. *The social Shaping of Technology*. Second edition. Buckingham, UK : Open University Press.
- Markus M-Lynne & Daniel Robey. 1988. Information technology and organizational change : Causal structure in theory and research. *Management Science* 583-598.
- Marr, Bernard. 2015. *Big data : Using SMART Big Data, Analytics and Metrics to Make Better Decisions and Improve Performance*. USA : John Wiley & Sons Ed.
- Mayer-Sch nberger, Viktor, Kenneth Cukier *et al.* 2014. *Big data : la r volution des donn es est en marche*. Paris : Robert Laffont.
- Michaux, Val ry & Frantz Rowe. 2004. Compl mentarit  entre syst me d'information informatis , communaut  de pratiques et vigilance dans la haute fiabilit  : le cas d'une compagnie d'assistance. *Syst mes d'Information et management* 9 (1) : 49-79.
- Orlikowski, Wanda. 1992. The duality of technology : Rethinking the concept of technology and organizations. *Organization Science* 3 : 397-427.
- Orlikowski, Wanda. 2000. Using technology and constituting structures : A practice lens for studying technology in organizations. *Organization Science* 4 : 404-428.
- Parks, Malcolm R. 2014. Big data in communication research : Its contents and discontents. *Journal of Communication* 64 : 355-360.
- Pouyllau, St phane. 2013. Web de donn es, big data, open data, quels r les pour les documentalistes. *Documentaliste - Sciences de l'Information* 50 : 32-33.
- Raghupathi, Wullianallur & Viju Raghupathi. 2014. Big data analytics in healthcare : Promise and potential. *Health Information Science and Systems* 2 : 3.

- Savall, Henri *et al.* 2008. *Mastering Hidden Costs and Socio-Economic Performance*. Charlotte, USA: Information Age Publishing.
- Silva, François & Stéphane Hugon. 2009. *Émergence des nouveaux usages liés à l'introduction des nouveaux outils dans les grandes entreprises*. Rapport CIGREF – ORSE.
- Silva, François & Anis Ben Ali. 2010. Émergence du travail collaboratif: nouvelles formes d'organisation du travail. *Management & Avenir* 36 (6): 340-365.
- Soulier, Eddy. 2004. Les communautés de pratique au cœur de l'organisation réelle des entreprises. *Systèmes d'information et management* 9 (1): 3-24.
- Tanti, Marc *et al.* 2010. Computerized support of decision-making in medical intelligence. In *Information systems and economic intelligence* sous la direction de Malek Ghenima, Kamel Smaili & Sahbi Sidhom S. Tunis: IHE Edition, 57-65.
- Tchankam, Jean-Paul. 2000. *Performance comparée des entreprises publiques et privées: une étude empirique dans un PVD*. Cahier de recherche, LAREMA.
- Vayre, Jean-Sébastien. 2014. Manipuler les données. Documenter le marché. Les implications organisationnelles du mouvement big data. *Les Cahiers du Numérique* 10 (1): 95-125.
- Weinstein, Marvin *et al.* 2013. *Analyzing big data with Dynamic Quantum Clustering*. SLAC National Accelerator Laboratory.

RÉSULTATS D'UNE ENQUÊTE SUR LES PRATIQUES ET ATTITUDES DES CHERCHEURS DE L'UNIVERSITÉ CONCORDIA EN MATIÈRE DE GESTION DES DONNÉES DE RECHERCHE

Danielle Dennie

Bibliothécaire responsable des communications savantes et de la réserve électronique
Bibliothèque de l'Université Concordia
danielle.dennie@concordia.ca

Alex Guindon

Bibliothécaire responsable des services de données géospatiales et numériques
Bibliothèque de l'Université Concordia
alex.guindon@concordia.ca

RÉSUMÉ | ABSTRACT

Selon les résultats d'un sondage et d'une série d'entrevues réalisés auprès de professeurs à l'Université Concordia à Montréal en 2015-2016 pour comprendre les besoins et les attitudes des chercheurs canadiens en matière de gestion des données, la majorité des chercheurs souhaitent mieux gérer, préserver et partager leurs données de recherche et reconnaissent les avantages de ce faire. En revanche, certaines contraintes les en empêchent, tels le manque de mesures incitatives ou de ressources humaines et technologiques, les problèmes associés à la confidentialité ou encore la volonté de garder un certain niveau de contrôle sur l'utilisation des données par autrui.

The Practices and Attitudes of Researchers at Concordia University with Regard to Research Data Management: Results of an Investigation

A survey and a series of interviews of Concordia University (Montréal) professors were undertaken in 2015-2016 to better understand the needs and attitudes of Canadian researchers in the area of data management. The majority of researchers want to better manage, preserve and share their research data. They also recognise the merits of doing so. However, certain barriers prevent them from better managing their data, namely the lack of incentives, the shortage of human and technological resources, confidentiality issues and the ability to exercise a certain level of control on the use of the data by others.

Introduction

Depuis maintenant plusieurs années, la question de la gestion des données de recherche (GDR) a été l'objet de beaucoup de discussions dans le milieu universitaire. On a d'abord déploré le peu d'efforts consacrés à la description, à l'archivage et au partage de ces données. La célèbre revue *Nature* a même publié un éditorial à ce sujet en 2009 (Data 2009). Puis, les choses se sont mises à évoluer peu à peu. Des états de la question ont été publiés et des groupes et des associations ont été formés, par exemple le Research Data Alliance, CODATA ou Datacite. Par la suite, des centres

d'expertise ont émergé tels que le Digital Curation Centre en Grande-Bretagne et le Australian National Data Service (ANDS) (aux États-Unis, l'Inter-University Consortium for Political and Social Research [ICPSR] existait déjà depuis plusieurs années). Finalement, les organismes de financement de la recherche ont commencé à exiger des actions de la part des chercheurs et des institutions universitaires afin d'assurer la pérennité et le partage des données issues de la recherche scientifique. En 2011, le Research Councils UK (RCUK) qui regroupe les principaux organismes subventionnaires du Royaume-Uni a publié une série de principes

relatifs à la GDR. La même année, la National Science Foundation (NSF) aux États-Unis a commencé à exiger, comme condition de remise des subventions de recherche, que les chercheurs soumettent des plans de gestion des données. De façon générale, dans la plupart des pays anglo-saxons (États-Unis, Royaume-Uni et Australie) et, dans une certaine mesure, en Europe, on a donc assisté à un véritable démarrage des initiatives visant l'adoption de meilleures pratiques de GDR.

Pendant ce temps au Canada les développements se sont fait attendre. Ce n'est qu'en 2016 que les grandes agences subventionnaires fédérales (*tri-agency*) ont émis une *Déclaration de principes des trois organismes sur la gestion des données numériques*¹. Nous attendons d'ailleurs, au cours de l'année qui vient, le dépôt par ces agences de politiques officielles accompagnées de directives à l'attention des universités et des chercheurs eux-mêmes. Si l'on se fie à la *Déclaration de principes* et à ce qui s'est passé chez nos voisins du Sud, on peut raisonnablement s'attendre à ce que ces exigences comprennent la rédaction d'un plan de gestion des données ainsi que l'archivage et le partage (sauf exception) des données de recherche.

À la veille de la mise en place de ces exigences, il importe de mieux comprendre les besoins et les attitudes des chercheurs canadiens quant à la gestion des données qu'ils produisent dans le cadre de leur recherche. Jusqu'à tout récemment, il existait peu d'études canadiennes à ce sujet, particulièrement pour les institutions de taille moyenne comme l'Université Concordia.

Cet article présente donc les principaux résultats d'une enquête auprès des chercheurs de l'Université Concordia à Montréal en matière de gestion des données de recherche. L'étude, qui s'est déroulée de l'automne 2015 à l'hiver 2016, consistait en un sondage en ligne suivi d'une série d'entrevues. Le sondage a été envoyé à l'ensemble des professeurs de l'Université à l'exception de trois départements (Géographie, Science politique, Sociologie) puisque ceux-ci avaient déjà été consultés lors d'une première étude effectuée en 2013. La première partie de l'article porte sur les résultats du sondage, tandis que la seconde présente un résumé des informations obtenues grâce à des entrevues avec les chercheurs.

Revue de la littérature

Bien que nous ayons répertorié un nombre relativement modeste d'études empiriques sur l'attitude et les pratiques des chercheurs en matière de GDR au Canada, plusieurs enquêtes de ce genre ont été menées aux États-Unis et en Grande-Bretagne. En général, ces études portent sur une seule université, mais il existe aussi quelques exemples

d'enquêtes multi-institutionnelles. La méthodologie employée consiste le plus souvent en un sondage Web envoyé aux chercheurs ou, à quelques occasions, à la combinaison d'un sondage et d'une série d'entrevues dans une perspective similaire à notre propre recherche. Cette revue de littérature ne se veut pas exhaustive, mais présente simplement quelques études récentes, en Amérique du Nord, qui pourront servir de points de comparaison avec notre enquête à l'Université Concordia. Aux lecteurs qui souhaiteraient consulter davantage d'études empiriques, nous recommandons d'examiner le tableau comparatif qui apparaît dans le récent article d'Elizabeth Berman (2017, 3).

À notre connaissance, une des plus grandes enquêtes quantitatives est celle effectuée par Carol Tenopir et ses collègues (Tenopir *et al.* 2011) d'octobre 2009 à juillet 2010. Ce sondage a rejoint 1 329 répondants (on estime que 15 000 personnes ont reçu le sondage, donc le taux de participation s'établirait autour de 9 %) provenant majoritairement de l'Amérique du Nord (75 %) ainsi que, dans une moindre mesure, des pays européens. Parmi les résultats les plus intéressants de cette étude, on remarque qu'environ 75 % de chercheurs disent partager leurs données. Par contre, seulement 36 % de répondants indiquent que leurs données sont « facilement accessibles ». En pratique, cela signifie que dans la plupart des cas les chercheurs partagent leurs données uniquement sur demande et qu'il existe des barrières importantes au partage plus systématique par voie électronique. Les principaux obstacles cités à ce propos sont le manque de temps pour préparer les données (54 %) et l'absence de fonds pour faire ce travail (40 %). Un autre aspect intéressant des résultats se situe au niveau du soutien institutionnel fourni par les universités. Ici, les auteurs séparent les résultats en deux catégories : les répondants provenant d'institutions axées sur la recherche et ceux qui sont affiliés à des institutions se concentrant sur l'enseignement. Comme on pourrait s'y attendre, le niveau de soutien semble plus fort dans les universités axées sur la recherche. Par exemple, 19 % des répondants des institutions de recherche sont fortement d'accord (*agree strongly*) avec l'énoncé « *My organization or project has a formal established process for managing data during the life of the project* », contre seulement 8,9 % pour les répondants venant des institutions d'enseignement. Pour l'énoncé « *My organization or project provides the necessary funds to support data management beyond the life of the project* », ces chiffres sont respectivement de 18,6 % (recherche) et de 9,4 % (enseignement). Finalement, les répondants qui se considèrent comme très satisfaits (*strongly agree*) pour ce qui est de la formation offerte par les organisations en matière de GDR sont de 24,9 % (recherche) contre 18,3 % dans les institutions d'enseignement.

Plus récemment, quelques études portant sur une seule université ont retenu notre attention. En 2014, Buys & Shaw (2015) ont mené une enquête à l'Université Northwestern.

1. <www.science.gc.ca/eic/site/063.nsf/fra/h_83F7624E.html>.

Des 12 940 personnes contactées — la population cible était constituée de tous les chercheurs, étudiants aux cycles supérieurs et candidats au postdoctorat —, 833 ont répondu à certaines questions et 788 ont complété le sondage, pour un taux de participation d'environ 6,4 %. Parmi les résultats intéressants, on remarque une grande diversité de pratique en matière de stockage des données :

Sixty-six percent use computer hard drives, 47 % use external hard drives, 50 % use departmental or school servers, 38 % store data on the instrument that generated the data, and 27 % use flash drives. Additionally, 31 % use cloud-based storage services. When asked to name their cloud-based storage (180 written responses) Dropbox (Dropbox, Inc., 2007) was the most popular choice (63 %). Only 6 % of the respondents use external data repositories. (Buys & Shaw 2015, 12)

Notons que les résultats ne font pas de distinction entre stockage durant le projet et archivage à la conclusion de celui-ci. Par ailleurs, pour ce qui est de la période de conservation, une pluralité de répondants veut préserver indéfiniment ses données brutes (181 répondants sur 644) et ses données publiées (230 sur 631). Les autres choix de réponses offerts pour la durée de préservation étaient : moins d'un an ; de 1 à 5 ans ; de 5 à 10 ans ; et plus de 10 ans.

La volonté de partager ses données varie selon les personnes avec qui on partage les données et selon que le partage se fait avant ou après la publication d'un article : avec les membres du groupe de recherche (avant publication 47 %, après 14 %) ; avec les collègues de Northwestern (avant publication 15 %, après 4 %) ; avec les chercheurs du même domaine (avant publication 35 %, après 47 %) ; avec le public général (avant publication 3 %, après 35 %). Ce dernier chiffre montre que le partage libre des données avec le grand public (via un dépôt de données par exemple) reste problématique tant que les chercheurs n'ont pas publié leurs résultats.

Il nous paraît intéressant de signaler une récente étude menée à l'Université du Vermont (Berman 2017) qui, comme notre propre enquête, repose sur la combinaison d'une approche qualitative — dans ce cas, une analyse du contenu de 35 plans de gestion des données (PGD) et une série d'entrevues — et d'une méthode quantitative, soit un sondage auprès des chercheurs ([N=319], taux de participation 26,8 %). L'auteur utilise une méthode exploratoire séquentielle (*exploratory sequential mixed method*) qui débute par une analyse des données qualitatives (entrevues et analyse de contenu des PGD) afin d'en dégager les thèmes principaux. Ceux-ci servent ensuite de base au développement du questionnaire utilisé pour le sondage. Dans notre étude à Concordia, nous avons procédé à l'inverse, c'est-à-dire en débutant par le sondage et en explorant certaines questions plus en profondeur lors des entrevues en personne.

Voici un aperçu des résultats les plus intéressants issus de l'étude à l'Université du Vermont. On remarque, aussi bien au niveau des résultats qualitatifs que quantitatifs, le manque de métadonnées créées pour décrire les données de recherche. L'auteur souligne que cette observation se retrouve dans de nombreuses études comme celle de Tenopir (2011). Pour ce qui est du partage des données, comme la plupart des études sur la GDR le rapportent, il y a bel et bien un désir de dissémination, mais celui-ci se heurte à des obstacles :

For researchers who have submitted DMPs, the issues focused around fear of misinterpretation, intellectual property concerns, and a variety of legal issues, including confidential, proprietary, or classified information. The major limitations from the quantitative phase of the study focused on the ability to maintain confidentiality, the lack of time, personnel, and tools/infrastructure to make data available, and intellectual property concerns. (Berman 2017, 12)

Comme dans les autres études mentionnées plus haut, l'auteur remarque que l'utilisation des dépôts de données est encore le fait d'une minorité : dans la phase qualitative de l'étude, 17 des 35 PGD mentionnent l'utilisation de dépôts ; pour ce qui est des répondants au sondage, seuls sept chercheurs sur 218 ont recours à ce genre de plateforme pour partager leurs données.

Nous terminerons cette revue de littérature en examinant ce qui s'est fait sur la scène canadienne au cours des dernières années. Les enquêtes portant sur la GDR sont relativement rares au pays, ce qui pourrait s'expliquer par le nombre modeste d'universités axées sur la recherche (comparativement aux États-Unis) et surtout par le retard qu'a pris le Canada sur les États-Unis et l'Europe au niveau des politiques et du soutien institutionnel pour la gestion des données de recherche. À notre connaissance, une des premières études est celle de Carol Perry (2008) qui portait spécifiquement sur l'archivage et le partage des données par les récipiends de bourses de recherche du CRSH. En 2013, Mowers, Humphrey & Perry (2013) ont publié les résultats sommaires d'une étude auprès de plus de 300 chercheurs canadiens. À ces études, s'ajoute la première mouture de notre enquête à l'Université Concordia qui a été menée en 2013 (Guindon 2014).

En 2015, un groupe de cinq universités canadiennes (l'Université de Toronto, l'Université de la Colombie-Britannique, l'Université de Waterloo, l'Université de l'Alberta et l'Université Queen's) a adopté un questionnaire commun (créé par l'Université de Toronto) pour sonder les chercheurs des facultés de science et de génie (Sewerin *et al.* 2016). Depuis, quatre autres universités ont annoncé qu'elles se joindraient à ce groupe. L'objectif de ce projet était de créer un ensemble de données sur la GDR qui permettrait de comparer les universités canadiennes. Le sondage s'est déroulé d'avril à décembre 2015 et a permis de rejoindre 780 participants. En matière de documentation des données, 38,5 % des répondants indiquent qu'ils croient avoir fourni assez

d'information pour que d'autres chercheurs, à l'extérieur de leur laboratoire ou équipe de recherche, puissent réutiliser les données contre 61,5 % qui répondent par la négative ou qui ne savent pas. Ces résultats vont donc dans la même direction que ceux de l'Université du Vermont et de plusieurs autres études. À la question qui portait sur la méthode de stockage des données (notons que les réponses multiples étaient acceptées), 60,9 % des chercheurs indiquent utiliser le disque dur d'un ordinateur, mais une proportion étonnamment élevée (59,6 %) mentionne utiliser un dépôt de données tel que Protein Data Bank, Cambridge Structural Database, GitHub, Dryad ou Figshare. Il est fort probable que ce résultat élevé soit lié au fait que les répondants provenaient uniquement des disciplines scientifiques. L'offre de dépôts spécialisés en science est assez importante et il semble que la pratique d'archivage formel des données de recherche y ait évolué plus rapidement que dans d'autres disciplines. En matière de partage des données, 57,6 % des répondants indiquent que la méthode de dissémination la plus fréquente est « *by personal request only* ». Une proportion significative (mais tout de même minoritaire) de chercheurs indique ne pas partager ses données, mais on note des variations assez importantes selon les domaines : « *Disciplines with the highest number of respondents that indicated they were not currently sharing data were computer science (44.9 %, n=22) and the engineering areas (civil/mineral/environmental, 34.2 %, n=25; biological/chemical/materials/mechanical, 24.1 %, n=20; electrical/computer, 25.0 %, n=3).* » Sur une note encourageante, quand on leur demande d'évaluer leur volonté de partager leurs données dans le futur, 30,3 % des répondants mentionnent qu'ils vont considérer l'utilisation de dépôts de données et seuls 11,5 % répondent qu'ils ne prévoient pas partager leurs données. Au chapitre des raisons invoquées pour justifier la décision de ne pas partager leurs données, les chercheurs indiquent à 48,1 % le désir de publier avant de disséminer les données ; on indique aussi des problèmes de propriété intellectuelle (23 %), des obligations contractuelles envers une tierce partie (19,7 %) et des questions de confidentialité dans 17 % des cas.

Méthodologie

L'équipe de recherche était composée d'Alex Guindon, bibliothécaire responsable des services de données statistiques et géospatiales, de Danielle Dennie, bibliothécaire de référence, ainsi que de Jennifer McGrath, professeure associée au Département de psychologie. Nous voudrions également remercier Dubravka Kapa, directrice associée à la Bibliothèque de Concordia et Muhammed Idris étudiant au postdoctorat pour leur aide à la création du questionnaire et pour la planification générale du projet de recherche.

Bien que l'élaboration du questionnaire ait été le fruit du travail de tous les collaborateurs mentionnés ci-dessus, la partie qualitative de la recherche — les entrevues effectuées

auprès des chercheurs ainsi que leur analyse — relevait exclusivement de Danielle Dennie et d'Alex Guindon.

Le sondage en ligne

L'objectif de notre projet de recherche était double : d'abord d'obtenir un portrait général des pratiques courantes et de l'attitude des chercheurs de l'Université Concordia dans la gestion de leurs données ; ensuite, d'identifier les services de GDR qui répondraient aux besoins de ces mêmes chercheurs. Le questionnaire² est une adaptation libre du Data Asset Framework (DAF) créé par le Humanities Advanced Technology & Information Institute (HATII) de l'Université de Glasgow avec l'aide du Digital Curation Centre³. Nous y avons ajouté plusieurs questions afin de mieux évaluer les pratiques et besoins des chercheurs en ce qui a trait à la manipulation des données, à leur traitement statistique et à leur visualisation. Puisque nous souhaitons obtenir des résultats quantitatifs, le sondage est composé de questions fermées avec, dans plusieurs cas, la possibilité de choisir la catégorie « autre » et d'ajouter des précisions. La partie qualitative de l'enquête repose sur les entrevues effectuées par la suite.

L'Université Concordia compte environ 28 000 étudiants (équivalent temps plein), dont 4 000 aux études supérieures. On y compte approximativement 1 000 professeurs à temps plein. La population du sondage, comme mentionné plus haut, était constituée de tous les professeurs à temps plein sauf ceux des départements de Géographie, Science politique et Sociologie qui avaient déjà été consultés lors d'un précédent sondage sur la GDR. Le sondage a été diffusé de décembre 2015 à janvier 2016 via la plateforme Survey Monkey. Le taux de participation s'est élevé à 19 %, soit 132 participants sur une liste de distribution de 696. Cependant, les participants n'étaient pas tenus de répondre à chacune des questions donc le taux de participation réel varie d'une question à l'autre. Ce taux de réponse est satisfaisant pour une enquête en ligne de ce type, mais il faut tout de même souligner qu'il est possible qu'un « biais d'autosélection » influence les résultats, c'est-à-dire que les chercheurs ayant choisi de répondre au sondage sont davantage intéressés par les questions de GDR que l'ensemble de la population.

La répartition de l'échantillon selon les grandes disciplines reflète assez bien la population de chercheurs à Concordia, sauf pour ce qui est des Arts où l'échantillon sous-représente clairement la population de chercheurs dans ce domaine. Il est probable que, malgré la définition très large du terme « données de recherche » que nous avons adoptée (qui pourrait être interprétée comme inclusives des productions artistiques), les professeurs de ce domaine ne se soient pas sentis concernés par cette étude.

2. <[dx.doi.org/10.11573/spectrum.library.concordia.ca.00982722](https://doi.org/10.11573/spectrum.library.concordia.ca.00982722)>.

3. <www.dcc.ac.uk/resources/tools/data-asset-framework>.

TABLEAU 1

Répartition des répondants par discipline

Discipline		
Science & Engineering	44	34 %
Social sciences	32	25 %
Humanities	19	15 %
Business	21	16 %
Arts	12	9 %

Les entrevues

Dix entrevues ont été effectuées auprès de chercheurs. Ceux-ci avaient indiqué dans le sondage qu'ils accepteraient d'être interviewés. Les chercheurs ont été sélectionnés d'après leur niveau d'intérêt relatif au partage et à l'archivage de leurs données, puisque nous considérons que les chercheurs ayant réfléchi à la question de la GDR peuvent apporter des réponses plus riches lors des entrevues. Un chercheur de chacun des départements suivants a été interviewé : Génie du bâtiment, civil et environnemental, Design et arts numériques, Informatique, Physique, Sciences de l'exercice, Gestion, Gestion de la chaîne d'approvisionnement et des technologies d'entreprises, Communications, Arts plastiques, Études françaises.

Les entrevues étaient d'une durée d'une heure, de format semi-structuré, et comme le sondage, basées sur une adaptation du Data Asset Framework du Humanities Advanced Technology & Information Institute (HATII) à l'Université de Glasgow. L'analyse des entrevues s'est faite en deux étapes : un codage descriptif était appliqué lors d'un premier cycle de codage (Saldana 2009, 70), suivi d'un codage sélectif (*focused coding*), lors d'un deuxième cycle (Saldana 2009, 155).

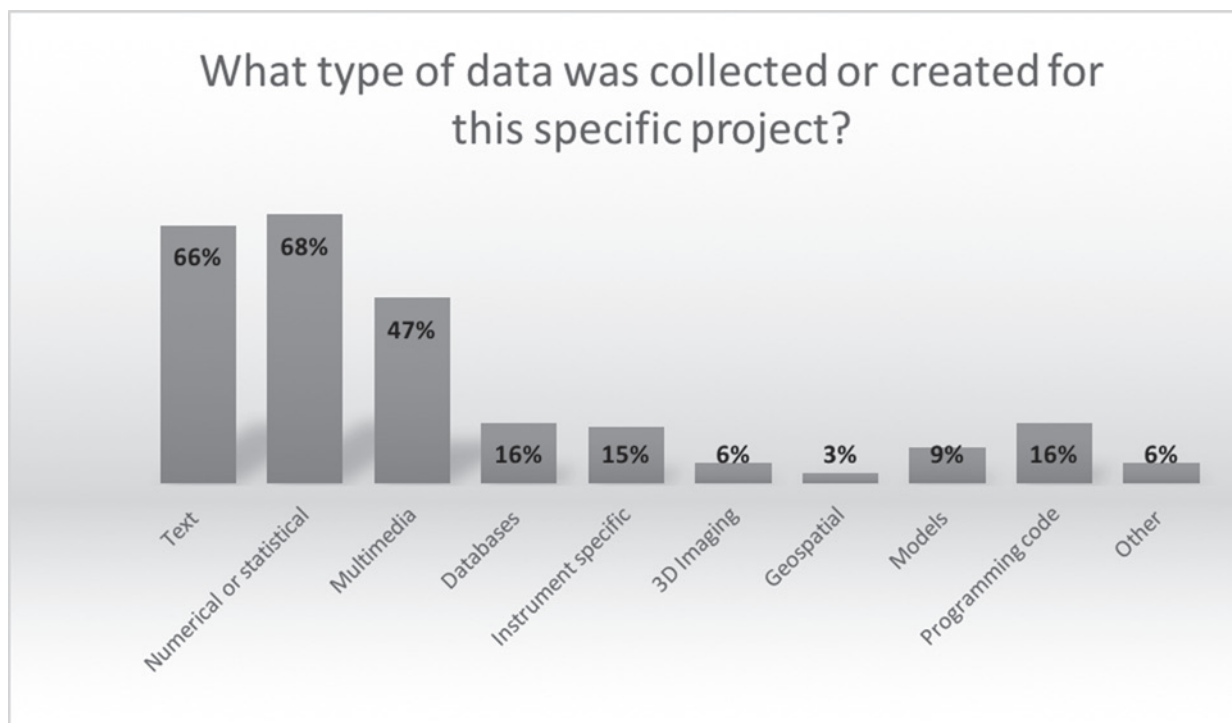
Résultats du sondage

Nous ne présentons ici que quelques-uns des résultats qui nous apparaissent les plus importants. L'ensemble complet de nos données agrégées est disponible en ligne⁴.

Pour bien comprendre nos résultats, une précision méthodologique s'impose : sauf avis contraire, les pourcentages rapportés reflètent le nombre de participants ayant donné une réponse X. Comme pour la plupart des questions, les répondants pouvaient choisir plusieurs réponses, le total des pourcentages pour une question donnée est supérieur à 100 %. En d'autres termes les différentes catégories (réponses) ne sont pas mutuellement exclusives.

FIGURE 1

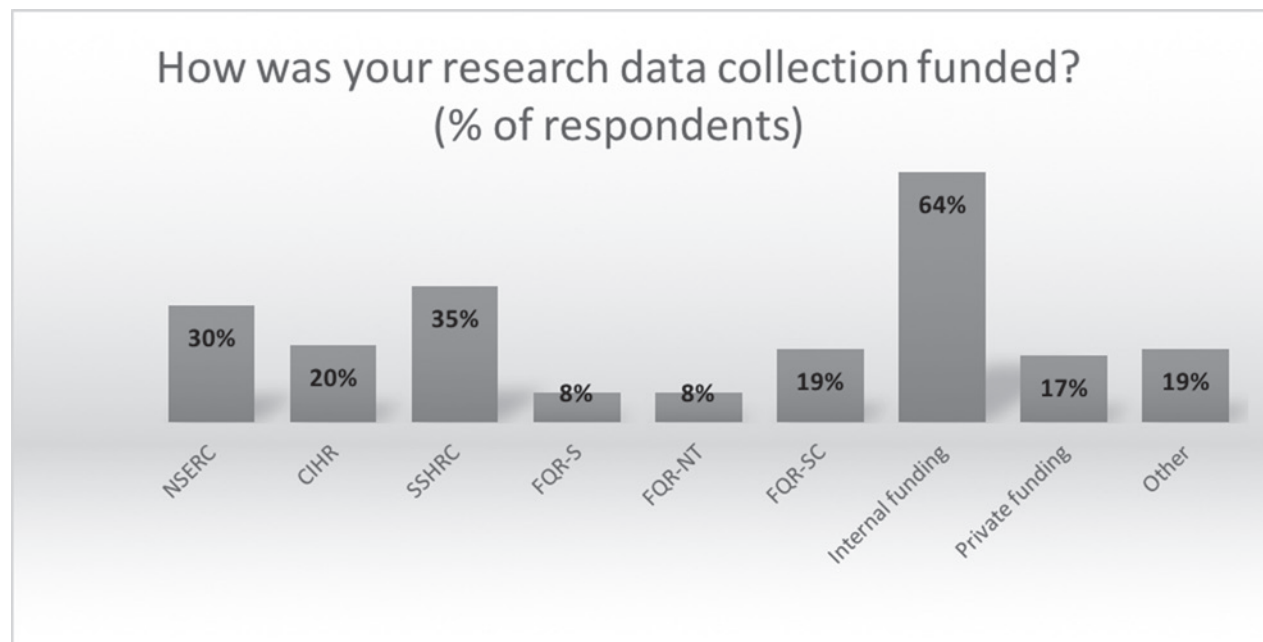
Quel type de données avez-vous collectées ou créées pour ce projet?



4. <[dx.doi.org/10.11573/spectrum.library.concordia.ca.00982722](https://doi.org/10.11573/spectrum.library.concordia.ca.00982722)>.

FIGURE 2

Comment votre collecte de données a-t-elle été financée?



Les graphiques et tableaux sont présentés en anglais puisque c'est dans cette langue que le sondage a été proposé. Par contre, nous avons ajouté un titre en français qui apparaît au-dessus de chaque figure.

Type de données

La Figure 2 démontre la grande diversité des données utilisées par les chercheurs. Les plus communes étant les données numériques ou statistiques (68 %), les données textuelles (66 %) et les données multimédias (images, audio, vidéo) (47 %). Sur cette base, on peut conclure que les services et outils offerts pour soutenir la GDR devraient être très flexibles et pouvoir s'adapter à divers types de données. Néanmoins, les services les plus développés ainsi que la plupart des fonctionnalités des outils devraient probablement se concentrer sur les statistiques et textuelles. Les objets multimédias sont d'une nature très différente et nécessiteront sans doute des outils particuliers.

Financement des projets de recherche

Le financement de la recherche à Concordia repose sur une multitude de sources. On remarque un fort taux de chercheurs (64 %) ayant obtenu un financement interne et un pourcentage significatif (17 %) ayant eu recours à des fonds privés.

Comparons maintenant le financement des trois grandes agences fédérales (*tri-agency*) avec celui obtenu d'autres sources. Pour ce faire, il nous semble utile de présenter les

résultats en fonction du nombre de réponses obtenues pour chaque catégorie plutôt que selon le nombre de participants ayant choisi chacune des catégories. Le nombre de réponses est une meilleure approximation de la distribution des subventions puisque le même chercheur peut avoir indiqué plusieurs sources de financement (les réponses multiples étant acceptées) pour des projets de recherche différents voire même pour un seul projet.

On remarque qu'uniquement 39 % des réponses indiquent un financement provenant des trois grands organismes (CRSH, CRSNG, IRSC) contre 32 % pour les autres agences et 29 % de financement interne (subventions provenant de l'Université Concordia) pour un total de 61 % pour les sources autres que le *tri-agency*. Comme les premières politiques de GDR et les exigences qui s'y rapportent proviennent du *tri-agency*, seule une minorité de projets à Concordia y serait soumise. Si cela peut donner davantage de marge de manœuvre et de temps à l'Université pour se préparer, il faudra par contre se poser la question du financement de la GDR pour les projets qui ne relèvent pas du *tri-agency*. En effet, ceux-ci ne pourront pas compter sur d'éventuelles subventions spécifiques à la gestion des données. Par ailleurs, le nombre important de projets financés par des fonds privés pose un problème particulier, car des ententes de confidentialité interviennent fréquemment entre chercheurs et entreprises qui financent leur recherche. Dans ces conditions il est souvent impossible de partager les données publiquement, voire même avec d'autres chercheurs de la même institution.

FIGURE 3

Financement en fonction du nombre total de réponses

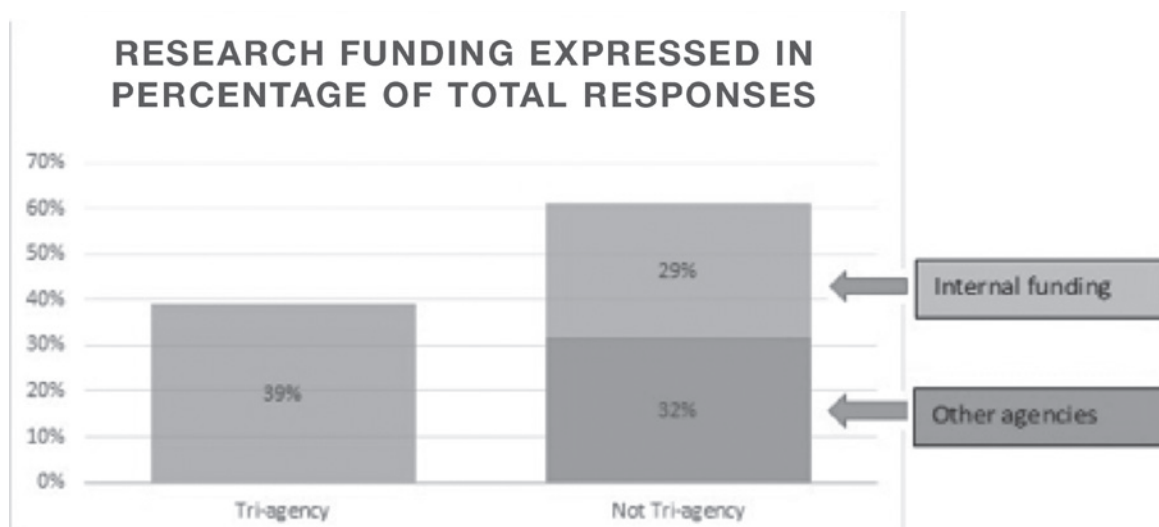
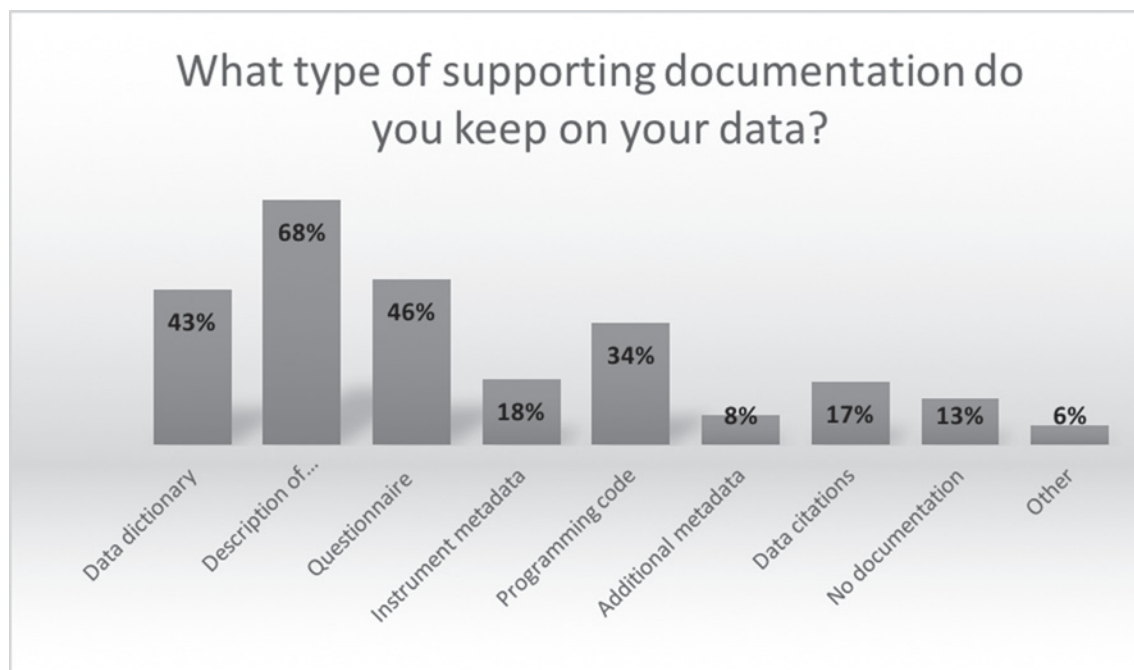


FIGURE 4

Quel type de documentation avez-vous créée pour décrire vos données ?



Documentation des données

Les résultats montrent que les chercheurs développent une large gamme de documents pour décrire leurs données. Seuls 13 % d'entre eux indiquent ne pas garder de documentation. Cette grande variété de documentation rappelle

encore une fois l'importance pour les universités de disposer de dépôts de données flexibles qui peuvent accueillir différents types de fichiers. Cela étant, lorsque ceux-ci sont disponibles dans un domaine scientifique, les dépôts disciplinaires spécialisés (GenBANK, ICPSR, etc.) seront souvent

plus appropriés que des dépôts multidisciplinaires puisqu'ils sont développés afin de répondre aux types de données et de fichiers de documentation ou de métadonnées existant dans cette discipline.

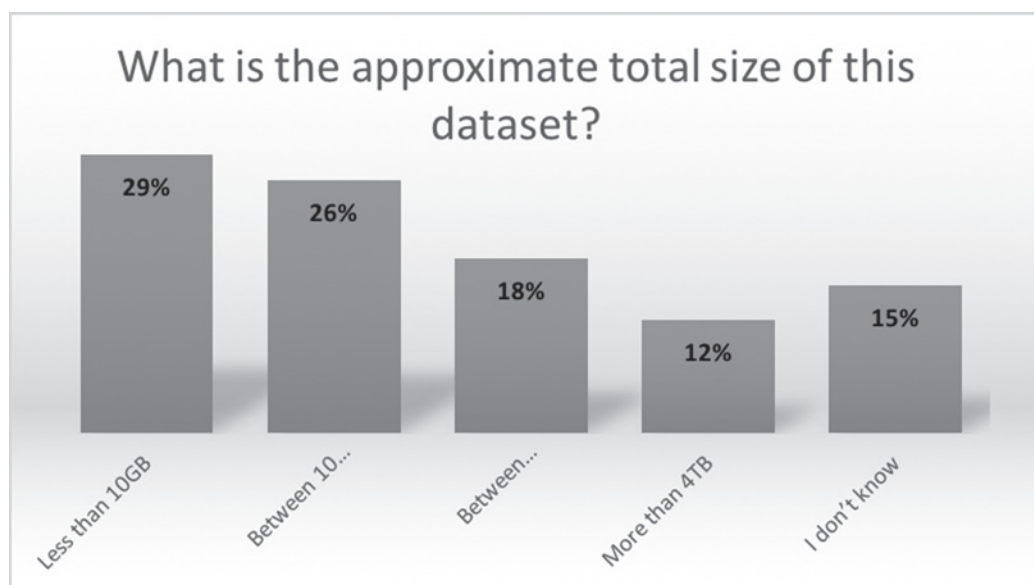
Taille des ensembles de données

Nous avons demandé aux chercheurs d'évaluer la taille de l'ensemble de données qu'ils considèrent comme étant leur plus important. Cinquante-cinq pour cent (55 %) des répondants rapportent que leur ensemble de données fait moins de 100 Go. Par contre, 30 % des chercheurs possèdent un gros

ensemble de données (plus de 100 Go) et 15 % indiquent ne pas savoir quelle est la taille de leur ensemble de données. Ces résultats sont plutôt encourageants, car ils indiquent que la majorité des ensembles de données pourraient sans doute être archivés par des dépôts de données reposant sur une infrastructure technologique relativement modeste. Quant aux données massives (plus de 100 Go), nous croyons qu'elles proviennent davantage des disciplines scientifiques (des analyses supplémentaires seraient nécessaires pour confirmer cette impression) dans lesquelles il existe souvent des dépôts spécialisés qui peuvent les accueillir.

FIGURE 5

Quelle est la taille approximative de votre ensemble de données?



Stockage des données

Les réponses à cette question reflètent la très grande diversité d'options disponibles pour les chercheurs. On peut toutefois conclure, puisque 91 % des répondants utilisent le disque dur de leur ordinateur et que seulement 20 % mentionnent utiliser un serveur départemental et 1 % celui des TI (IITS est le service des technologies de l'information à Concordia), qu'ils n'utilisent pas des moyens de stockage idéaux. En effet, les disques durs employés seuls ne sont pas une façon sécuritaire de stocker les données ou de les partager avec les collaborateurs. Par ailleurs, l'utilisation de solutions infonuagiques commerciales (41 %) peut faciliter le partage avec les membres de l'équipe de recherche, mais peut poser problème en termes de sécurité et de confidentialité. Ces résultats démontrent le besoin d'une solution de stockage ou d'un *environnement de recherche*

virtuel basé à Concordia et qui répondrait aux besoins de sécurité et de partage des données⁵.

Archivage des données

Presque un tiers (30 %) des chercheurs déclarent ne pas vouloir archiver leurs données à la conclusion du projet de recherche. Cette attitude peut s'expliquer de plusieurs façons, notamment par une interprétation restrictive des obligations en matière d'éthique ou une méconnaissance des avantages liés à l'archivage et au partage des données. Il est aussi possible que, dans certains cas, les données ne présentent effectivement que très peu de valeur à long terme. Cependant, il faut noter que, si l'on additionne les résultats pour Spectrum (dépôt institutionnel de Concordia) (21 %) et ceux pour la catégorie « dépôt généraliste ou disciplinaire » (10 %), un nombre légèrement plus élevé

5. Ce besoin a aussi été identifié lors d'un sondage interne effectué par IITS auprès des grands laboratoires de recherche de Concordia.

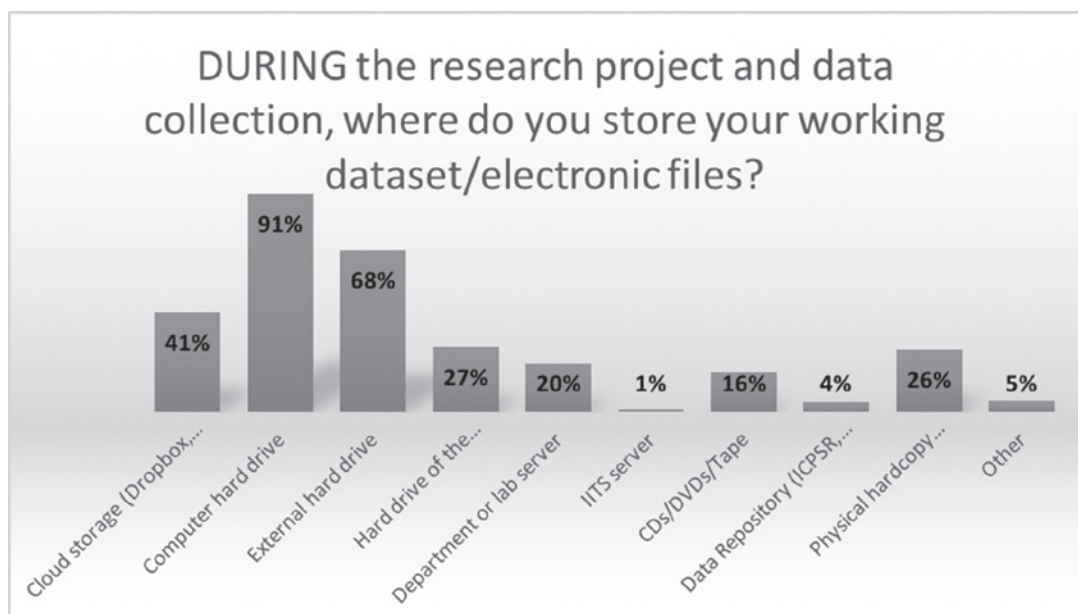
(31 %) est intéressé par l'archivage des données. Il est intéressant de noter cette préférence pour une option de stockage locale plutôt que l'utilisation d'un dépôt externe. On peut penser que cela reflète une confiance en l'habileté de Concordia à fournir un service sécuritaire ou alors un attachement des chercheurs à leur institution. Par contre, une fraction encore plus importante des répondants (42 %) ne connaît pas les dépôts de données, mais souhaiterait en apprendre davantage. Cela indique sans doute une méconnaissance des dépôts spécialisés comme ICPSR, GenBANK, PsychData, etc. De toute évidence, il y a beaucoup de travail

à effectuer pour faire connaître les différents dépôts disponibles pour l'archivage et le partage des données.

Finalement, notons que certains résultats sont très semblables à ceux obtenus lors du sondage de 2013 (Guindon 2014) dans lequel 32 % des répondants indiquaient qu'ils n'étaient pas intéressés à archiver leurs données. Par contre, le choix de réponses pour les options d'archivage était différent lors du premier sondage ce qui rend la comparaison un peu boiteuse, mais mentionnons tout de même que 39 % des répondants avaient exprimé un intérêt pour Spectrum et 29 % pour des dépôts externes. Encore ici, on remarque donc une préférence pour la solution locale.

FIGURE 6

Durant votre projet de recherche et pendant la collecte de données, où gardiez-vous vos jeux de données/fichiers de travail ?



Période de préservation des données

Nous notons une forte préférence (41 %) pour un archivage de durée indéfinie. Bien que cela puisse être souhaitable pour plusieurs ensembles de données, ce n'est pas toujours nécessaire ou même techniquement réalisable. En effet, certains jeux de données peuvent perdre leur pertinence après un certain nombre d'années comme l'indiquent aussi les réponses à cette question. Nous croyons qu'il faudra établir des politiques de rétention claires et alignées sur les éventuelles exigences des agences de financement.

Méthodes de partage des données

Il est très intéressant de voir que seuls 15 % des répondants indiquent qu'ils préfèrent ne pas partager leurs données. Vingt-trois pour cent (23 %) des chercheurs disent quant à eux ne pas partager leurs données jusqu'à maintenant, mais voudraient en apprendre davantage sur les options de

partage qui s'offrent à eux. Par contre, lorsque l'on s'attarde aux méthodes de partage, on voit que seule une minorité des chercheurs utilisent des dépôts de données pour la dissémination (7 % en accès libre et 16 % avec un contrôle sur l'accès). À ceux-ci, on peut ajouter les répondants qui partagent leurs jeux de données via des sites Web spécialisés. Les autres préfèrent des méthodes plus « manuelles » de partage, telles que le transfert physique (46 %) (via un disque dur ou une clé USB par exemple) ou par courriel (51 %). S'il est évident que ces méthodes de partage ne sont pas idéales au niveau de la sécurité en plus de ne pas être très pratiques, elles reflètent le désir des chercheurs de garder un contrôle direct sur la dissémination de leurs données. Enfin, si l'on compare ces résultats avec ceux obtenus en 2013 (Guindon 2014), on remarque que le taux de répondants qui ne sont pas intéressés à partager leurs données est assez constant : 15 % en 2013 contre 13 % en 2016.

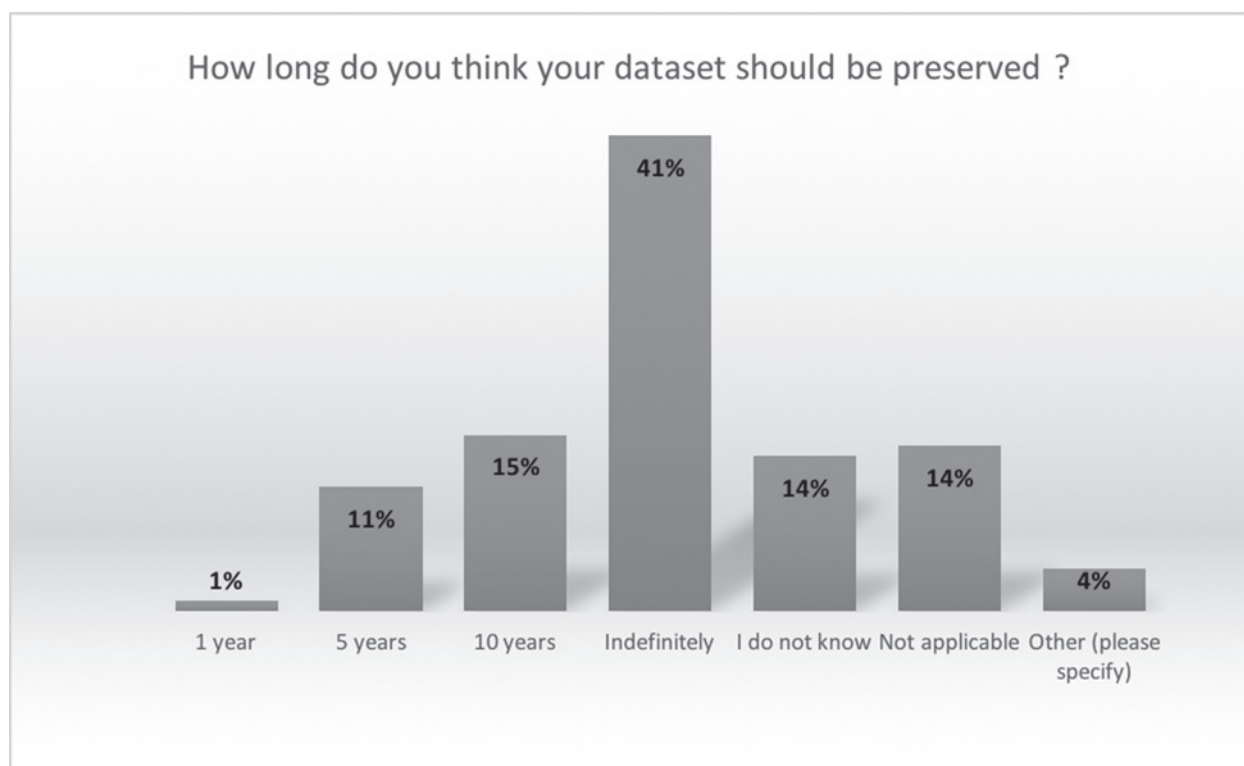
FIGURE 7

À la conclusion de votre projet de recherche, dans quel dépôt de données considérez-vous archiver votre ensemble de données?



FIGURE 8

Combien de temps vos données devraient-elles être préservées?



Expérience avec les périodiques qui exigent le dépôt des données

Le nombre de périodiques qui exigent le dépôt des données soutenant les articles scientifiques est en pleine croissance. Rappelons que ces revues ne demandent pas nécessairement que les données soient offertes en libre-accès. En effet, dans certains cas, on accepte le dépôt sur une plateforme qui permet de contrôler l'accès aux données. Nous observons que 28 % des répondants ont en effet soumis des manuscrits à de telles revues. Cependant, un plus grand nombre de chercheurs (34 %) ont jusqu'à maintenant évité de publier dans ces périodiques. Il sera intéressant de voir comment l'attitude des chercheurs évoluera au fur et à mesure que les politiques de dépôt de données deviendront de plus en plus fréquentes.

Offre de services de GDR

Un des objectifs importants de notre étude était de cerner les besoins des chercheurs en matière de soutien aux activités de GDR. Nous avons demandé aux chercheurs d'indiquer

leur niveau d'intérêt pour une série d'ateliers ou de services d'aide en GDR. Les résultats du tableau 2 ci-dessous sont présentés en ordre décroissant d'intérêt pour les répondants (selon la colonne « *very interested* »). L'élément qui suscite, de loin, le plus d'intérêt est « *data management activities that will be required by grant funding agencies* ». Cela n'est pas une surprise, mais confirme que la GDR prendra beaucoup d'importance aux yeux des chercheurs une fois que des exigences concrètes seront mises de l'avant par les agences de financement. De façon générale, il semble que les répondants soient intéressés par une vaste gamme d'activités. Ainsi, même l'activité la moins populaire selon la colonne « *total interested* », « *help in ensuring the confidentiality/anonymity of data on human participants* », obtient tout de même un score de 49 %. Il sera difficile de développer des services répondant à tous ces besoins et il faudra sans aucun doute compter sur plusieurs partenaires au sein des universités (bibliothèque, services des TI, bureau de la recherche, services juridiques).

FIGURE 9

Comment partagez-vous vos données avec d'autres utilisateurs?

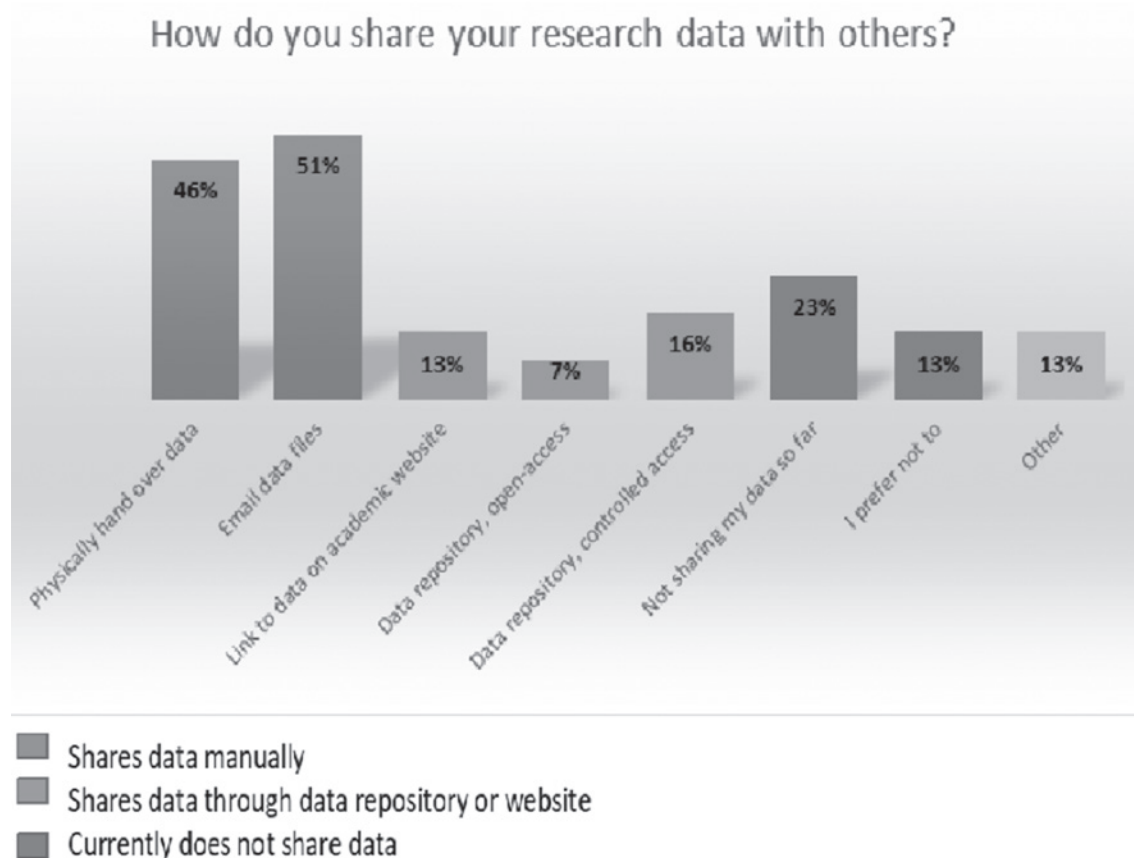


FIGURE 10

Quelle est votre expérience avec les périodiques qui exigent le dépôt des données?

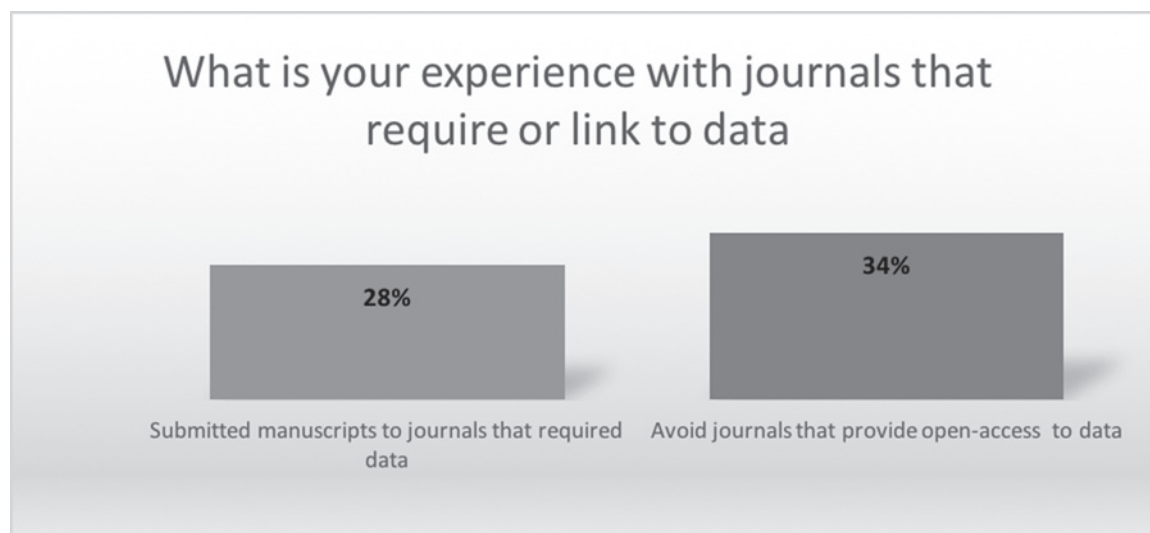


TABLEAU 2

Niveau d'intérêt pour des services de soutien en GDR

Workshop/Resource	Very interested	Somewhat interested	Total interested	Not interested
Data management activities that will be required by grant funding agencies	43 %	34 %	77 %	15 %
Data management plan	30 %	42 %	72 %	21 %
Research collaboration software tools	29 %	39 %	69 %	25 %
Finding and accessing existing data sources	29 %	34 %	63 %	23 %
Data archiving	27 %	43 %	69 %	22 %
Data visualization	27 %	44 %	71 %	22 %
Digitization of paper records	27 %	39 %	65 %	26 %
Data sharing and access	23 %	40 %	63 %	30 %
Help in ensuring the confidentiality/anonymity of data on human participants	21 %	28 %	49 %	29 %
Data management workshop	19 %	49 %	68 %	25 %
Data documentation and metadata	18 %	38 %	56 %	33 %

Sommaire des entrevues

Documentation décrivant les données

Quoique la majorité des chercheurs aient indiqué dans le sondage qu'ils produisaient de la documentation décrivant leurs données, plusieurs participants aux entrevues affirment pour leur part qu'ils n'en ont pas. Les seuls à détenir des documents sont deux chercheurs en sciences pures. Pour les autres, les raisons évoquées pour expliquer ce manque de documentation sont multiples: certains croient que leurs données sont suffisamment explicites par elles-mêmes ou que la section méthodologie de leurs publications décrit

suffisamment bien leurs données; d'autres chercheurs, en sciences humaines, croient qu'il n'y avait pas d'incitatifs ou de récompenses liés à la documentation ou la gestion des données. D'après une chercheuse: «[I]f there was acknowledgement of sustainability and there were rewards for that, that would be fantastic. There's more rewards for bringing new stuff in then there is rewards for careful preservation. And that's my observation from 25 years here.» Finalement, certains chercheurs ont indiqué qu'ils n'avaient pas assez de temps pour bien documenter leurs données. Un ingénieur a dit: «[...] documentation is gonna be [...] extra work for me and my team. And so, [...] I'm not really sure if you want to go that way

because I prefer students to do scientific work and publish papers then, you know, be data managers.»

Stockage des données

La majorité des chercheurs interviewés éprouvent de la difficulté à trouver de l'espace pour le stockage de leurs données de recherche en cours ou complétée. Plusieurs d'entre eux aimeraient avoir un serveur institutionnel, géré par l'université, avec des politiques et des directives d'utilisation claires, qui pourrait servir à stocker les données de recherche et qui serait doté d'un logiciel pour gérer ces données. Les chercheurs en sciences et en génie sont de ceux pour qui ce besoin était le plus criant. Un des scientifiques a bien résumé ce besoin :

We need an institution server [...]. [W]e should have secured access with secured account for each research team, for each lab. [...] [T]here should be a flexibility [to accommodate] people with space needs. [We] need [...] a technical team supporting this server [...] someone to [...] check that there is enough available space, see if there's problems, to fix the problem, to manage the access, the accounts, the codes, and make sure [it is] regularly backed up and whether there should be some [...] duplicate somewhere, to make sure everything's safe.

Les chercheurs ayant des données confidentielles provenant de sujets humains ou de sources de données secondaires (données obtenues de tierces parties) insistent sur le fait que ce serveur se devrait d'être sécurisé et que l'université devrait créer des politiques et des directives indiquant où et comment les données de recherches devraient être stockées. Quoique 41 % des répondants du sondage utilisent un serveur commercial infonuagique, quelques-uns des chercheurs interviewés ont exprimé leurs inquiétudes vis-à-vis la sécurité des données sur ce type de serveurs basés aux États-Unis. D'après un chercheur en sciences humaines : *«I think we need our own indigenous resources, so that [...] we're under the injunctions of the agencies that are federally funded and mandated, yet the problem of data leakage or data surveillance or data theft, once it is outside of our borders, it's in some ways beyond our control [...].»*

Une petite minorité des interviewés ne veut pas stocker ses données sur un serveur institutionnel afin de pouvoir en garder un contrôle absolu et de pouvoir répondre rapidement aux problèmes techniques potentiels.

Partage des données

Les résultats du sondage démontrent que la majorité des chercheurs partagent leurs données, mais ne nous permettent pas de savoir avec qui ces données sont partagées. Lors des entrevues, la majorité des chercheurs, peu importe leur discipline, ont indiqué qu'ils partageaient leurs données de recherche seulement avec des collaborateurs ou des étudiants directement engagés dans leur projet de recherche. Certains

ne partagent pas leurs données à l'extérieur de ce groupe malgré le fait qu'ils aient reçu des fonds d'organismes subventionnaires ayant des politiques de partage de données. De tous les chercheurs interviewés, seulement trois partagent leurs données avec des personnes autres que leurs collaborateurs immédiats ou leurs étudiants. Ils le font soit parce qu'ils croient aux bienfaits du partage ou parce qu'ils doivent se conformer aux politiques des organismes subventionnaires.

Les règles de confidentialité dans les formulaires de consentement relatif à l'éthique de la recherche sont la raison principale évoquée par les chercheurs pour expliquer la raison du non-partage. Une chercheuse en gestion craint que des politiques de partage des données imposées par les organismes subventionnaires peuvent faire entrave à la recherche puisque certains participants refuseraient alors de se faire interviewer. Finalement, quelques chercheurs indiquent aussi qu'ils ne croient pas que le partage des données brutes soit nécessaire dans leur communauté de recherche.

Cela étant, la majorité des chercheurs sont disposés à partager leurs données puisqu'ils y voient des avantages, tels que la possibilité de vérifier ou de reproduire la recherche et la possibilité pour l'ensemble de données d'être cité. Un chercheur en gestion a dit : *«[...] [G]et[ting] citations [...] is [...] how my university recognizes the value of what I've done. [...] I realize that if your article is valuable, people might cite it, but a lot of the time, the most valuable thing about an article is the data that was collected.»*

Par contre, certaines conditions de partage sont vues comme nécessaires, telles que l'anonymisation, un accès contrôlé, des embargos, ou même des systèmes de gestion des droits numériques (GDN). Quelques chercheurs soulèvent aussi la complexité liée au partage des données secondaires (données obtenues de tierces parties).

Archivage des données

De tous les chercheurs interviewés, un seul, en sciences, archive ses données dans un dépôt de données. Les chercheurs ont tout de même une opinion sur les caractéristiques jugées essentielles pour ce genre de dépôt.

La première est la capacité de stocker des données pour de très longues durées. Cette caractéristique est considérée comme particulièrement importante pour des données qui ont été difficiles à récolter, qui pourraient être utilisées pour des études longitudinales, qui pourraient être liées avec des données d'une autre recherche, ou pour l'exploration de données (*data mining*). La deuxième caractéristique essentielle est la présence d'un financement stable — préférablement en provenance des organismes subventionnaires — afin d'assurer la pérennité du dépôt de données. Malheureusement, d'après un des chercheurs, *«[...] the funding doesn't work that way. [They] pay you to do the project, and once the project's over, [y]ou [get] zero support. And if the*

data is still there when someone starts up the machine the next time, good [...]». Des politiques claires de préservation et d'accès sont d'autres caractéristiques importantes des archives. Certains chercheurs ayant des données confidentielles prônent des mécanismes qui permettraient de sécuriser les données. Une petite minorité de participants ne croient pas à la nécessité de l'archivage des données puisque leur communauté de recherche n'aurait pas, selon eux, le besoin de réutiliser ou d'analyser à nouveau les données.

Conclusion

Les résultats du sondage et des entrevues démontrent que la majorité des chercheurs voudraient adopter une meilleure approche pour gérer, préserver et partager leurs données de recherche. La majorité reconnaît les avantages de la GDR en général et du partage des données en particulier, à la fois pour les chercheurs qui produisent les données et pour la communauté scientifique en général.

Par contre, à l'heure actuelle, peu de chercheurs ont adopté de bonnes pratiques de gestion des données. Lorsqu'il y a de la documentation, elle respecte rarement les normes acceptées; lorsqu'il y a des procédures de sauvegarde, elles sont rarement systématiques ou n'adhèrent pas à des mesures de sécurité appropriées; lorsque les données sont préservées à la fin d'un projet, c'est souvent sur un disque dur ou, au mieux, sur le serveur du laboratoire ou du département. Et lorsque les données sont partagées, cela est fait de façon informelle, soit en personne (transfert sur disque dur ou clé USB) ou par courriel, ce qui corrobore ce que d'autres études ont observé (Tenopir *et al.* 2011; Sewerin *et al.* 2016).

Les obstacles à l'adoption des meilleures pratiques de GDR sont multiples. Les participants à notre étude ont quant à eux mentionné, entre autres:

- le manque d'incitatifs dans le monde universitaire pour la gestion et le partage des données;

- le manque de ressources — à la fois humaines et technologiques — nécessaires à la conservation des données;
- les problèmes associés à la confidentialité et de l'éthique;
- la volonté de garder un certain niveau de contrôle sur l'utilisation de leurs données.

Ces éléments concordent bien avec ceux relevés dans la littérature (Berman 2017; Sewerin *et al.* 2016).

Nous croyons que l'annonce prochaine des politiques des trois organismes subventionnaires ainsi que des exigences qui y seront associées constituera un moment charnière dans l'évolution des attitudes et des pratiques des chercheurs en matière de GDR. Ainsi, plus de 75 % des répondants à notre sondage ont indiqué qu'ils souhaitent suivre des formations sur la gestion des données reliées aux exigences des organismes subventionnaires. En conséquence, nous pouvons nous attendre à une plus grande demande d'aide pour la GDR dans les universités canadiennes lorsque ces nouvelles exigences seront adoptées. Les récentes études sur les pratiques et attitudes des chercheurs en matière de GDR entreprises par plusieurs universités canadiennes peuvent constituer un point de départ pour évaluer les besoins les plus immédiats de nos chercheurs. Cependant, il reste beaucoup de travail à faire et le succès de l'entreprise dépendra de la collaboration de plusieurs acteurs au sein des universités: les chercheurs eux-mêmes, bien sûr, mais aussi les bibliothèques, les services de TI, les bureaux de la recherche, les comités d'éthique et les services juridiques. De plus, il est illusoire de croire que le développement de services de GDR pouvant répondre aux exigences des agences de financement de la recherche puisse se faire à coût nul. Il faudra donc compter sur du financement et des infrastructures qui proviendront des universités et des différents paliers de gouvernement, mais aussi, espérons-le, des agences subventionnaires elles-mêmes.

SOURCES CONSULTÉES

- Berman, E. A. 2017. An exploratory sequential mixed methods approach to understanding researchers' data management practices at UVM: Integrated findings to develop research data services. *Journal of eScience Librarianship* 6 (1): e1104. doi: 10.7191/jeslib.2017.1104
- Buys, C., & P. Shaw. 2015. Data management practices across an institution: Survey and report. *Journal of Librarianship and Scholarly Communication* 3 (2): p.eP1225. doi: 10.7710/2162-3309.1225
- Data's shameful neglect. 2009. *Nature* 461 (7261). doi: 10.1038/461145a
- Guindon, A. 2014. Research data management at Concordia University: A survey of current practices. *Felicitier* 60 (2): 15-17. <search.proquest.com/docview/1526973990>.
- Hickson, S., K. A. Poulton, M. Connor, J. Richardson & M. Wolski. 2016. Modifying researchers' data management practices: A behavioural framework for library practitioners. *IFLA Journal* 42 (4): 253-265.

- Mowers, S., C. Humphrey & C. Perry. 2013. *Summary report: Survey of researchers' needs and practices regarding research data management in Canada*. Consulté le 1^{er} août 2017. <gsg.uottawa.ca/data/open/aa-interim-survey-report/20130801-en.pdf>.
- Saldana, Johnny. 2009. *Coding Manual for qualitative researchers*. Los Angeles: Sage.
- Sewerin, C. *et al.* 2016. From coast to coast: Canadian collaboration in a changing RDM seascape. Consulté le 1^{er} août 2017. <hdl.handle.net/1807/72802>.
- Tenopir, C. *et al.* 2011. Data sharing by scientists: Practices and perceptions. *PloS One* 6 (6): e21101. doi:10.1371/journal.pone.0021101
- Van Tuyl, S. & G. Michalek. 2015. Assessing research data management practices of faculty at Carnegie Mellon University. *Journal of Librarianship and Scholarly Communication* 3 (3). doi: 10.1184/RDL/01092015

VALORISER LES DONNÉES D'ENQUÊTES QUALITATIVES EN SCIENCES SOCIALES : LE CAS FRANÇAIS DE LA BANQUE D'ENQUÊTE BEQUALI

Selma Bendjaballah

Ingénieure de recherche, CDSP
selma.bendjaballah@sciencespo.fr

Sarah Cadorel

Archiviste, CDSP
sarah.cadorel@sciencespo.fr

Emilie Fromont

Archiviste, CDSP
emilie.fromont@sciencespo.fr

Guillaume Garcia

Ingénieur de recherche coordinateur
du projet beQuali, CDSP
guillaume.garcia@sciencespo.fr

Emilie Groshens

Documentaliste, CDSP
emilie.groshens@sciencespo.fr

Emeline Juillard

Documentaliste, CDSP
emeline.juillard@sciencespo.fr

RÉSUMÉ | ABSTRACT

Réutiliser des matériaux d'enquêtes qualitatives en sciences sociales pour produire de nouvelles recherches et enseigner les méthodes : tel est le questionnement scientifique ayant conduit en 2011 à la création de la banque d'enquêtes qualitatives au Centre de données socio-politiques (CDSP).

Le présent article expose les réflexions menées au sein de beQuali autour des enjeux de la réutilisation des données qualitatives. Il détaille pour chaque étape du processus — de la collecte des archives à l'exploration des corpus sur le site Web — les différentes problématiques qui ont préfiguré la mise en place du dispositif beQuali et les moyens mis en œuvre pour construire et faire fonctionner l'équipement tel qu'il existe aujourd'hui.

Highlighting the Data of Qualitative Studies in the Social Sciences : The Case of beQuali, the French Investigation Bank

The scientific goal that led to the creation of the qualitative investigation bank, beQuali, at the *Centre de données socio-politiques* (CDSP) in 2011 was the reuse of qualitative research materials in the social sciences to produce new research and to teach the methods.

This paper outlines how the beQuali team addressed this issue. Each stage of the work is described in detail—from the collection of archives to the exploration of text corpora on the website. The paper also explains the issues that preceded the implementation of beQuali and the different methods used to construct and operate the equipment as it exists today.

Financée depuis 2011 par un programme national d'investissement¹ en faveur d'équipements scientifiques innovants en France, la banque d'enquêtes qualitatives (beQuali) est l'un des instruments de l'Equipex DIME-SHS². Ce projet développé au Centre de données sociopolitiques³ (CDSP) vise à doter la France d'outils en matière de collecte,

d'enrichissement et de diffusion des données pour la recherche en sciences humaines et sociales. À l'instar d'autres projets européens tels que ESDS Qualidata désormais rattaché au UK Data Service, beQuali s'inscrit dans le mouvement d'ouverture des données de la recherche et propose gratuitement à la communauté scientifique en SHS un service de numérisation, de documentation, de mise à disposition et d'archivage numérique d'enquêtes qualitatives. La banque d'enquêtes répond principalement à trois besoins : préserver les données de recherche en SHS, rendre accessible des enquêtes qualitatives pour la production de nouvelles recherches et proposer des ressources pour l'enseignement des méthodes.

1. Programme d'Investissements d'avenirR : <www.agence-nationale-recherche.fr/investissements-d-avenir>.

2. Équipement d'excellence « Données Infrastructures et Méthodes d'Enquête en Sciences Humaines et Sociales » : <www.sciencespo.fr/dime-shs/>.

3. Unité mixte de service Sciences Po Paris/CNRS : <cdsp.sciences-po.fr/>.

Dès sa création, beQuali a été confrontée à plusieurs risques inhérents à ce type de démarche, notamment à celui de voir mettre à mal l'anonymat des enquêtés ainsi qu'un risque d'erreurs d'interprétation lié à la décontextualisation des données (Bishop 2006, 2014). Les principes qui animent son développement visent à réduire ces risques en tenant compte des expériences similaires ayant eu lieu en Europe ou dans le monde. Son ambition est d'essayer de dépasser leurs limites, notamment en ce qui concerne les conditions dans lesquelles les corpus d'enquête sont reconstitués, de manière à garantir une compréhension suffisante du processus de recherche originel et de l'organisation du corpus de documents rassemblés. Pour répondre à ces différents enjeux, l'équipe beQuali a dû mettre en œuvre un processus de traitement original, à la croisée de plusieurs pratiques documentaires, mêlant archivistique traditionnelle, bibliothéconomie, humanités numériques et documentation de données.

Cet article présentera d'abord le fonctionnement de ce service et notamment le nécessaire travail collaboratif associant le producteur et plusieurs compétences métiers. Puis, il abordera les spécificités de l'unité documentaire telle que définie et traitée dans la banque d'enquêtes, avant de décliner les choix de traitement et de documentation opérés par beQuali pour transformer ces données originales en corpus numériques enrichis et exploitables pour des usages scientifiques. La question des usages sera ensuite traitée à partir d'un retour d'expérience d'enseignement des méthodes qualitatives et à travers une présentation du catalogue et des outils d'exploration de beQuali.

Le fonctionnement du dispositif

Partager ses enquêtes sur beQuali : comment ça marche ?

Le traitement des enquêtes, de la collecte à la publication, est pris en charge par l'équipe de beQuali. Si les plateformes d'archives ouvertes (comme HAL ou Spire⁴) permettent le dépôt de publications par leurs auteurs, ce modèle ne paraissait pas adapté pour le partage des enquêtes qualitatives. Tandis que sur les premières, un dépôt correspond généralement à un fichier numérique (un article, un chapitre de livre, une thèse), les enquêtes contiennent plusieurs centaines de documents, soit autant de fichiers à documenter (décrire, nommer, dater) lors du téléversement. De plus, le fait de considérer les données de recherche comme un matériau communicable est un paramètre nouveau de la recherche. Contrairement à la littérature scientifique, les données peuvent ne pas être prêtes pour une communication en l'état, et exiger au préalable une série d'interventions. Parmi les documents

collectés, on relève par exemple des transcriptions d'entretiens imprimées et non anonymisées, des fichiers nativement numériques dont les noms ou l'arborescence de classement révèlent de l'information sur les enquêtés, des cartons dans lesquels les archives d'une enquête sont mélangées avec d'autres archives sans qu'il y ait d'unité sémantique, etc. Avant la mise en ligne, beQuali prend donc en charge le classement, la numérisation, l'anonymisation, la sélection, et la documentation de ces données.

Le dépôt d'enquêtes sur le site beQuali passe aujourd'hui par un appel à propositions qui a lieu une à deux fois par an⁵. L'appel s'adresse aux chercheurs ou aux institutions ayant produit ou conservant des enquêtes qualitatives en sciences sociales. Le formulaire de proposition permet aux déposants de renseigner différentes informations décrivant l'enquête et les matériaux produits et conservés (la méthodologie employée, le type et le volume des matériaux, leur état, le caractère sensible des données, les publications). Les candidatures sont examinées et sélectionnées par le Comité scientifique et technique (CST) de beQuali, composé d'une dizaine d'experts : chercheurs et enseignants de différentes disciplines (sociologie, anthropologie, sciences politiques), spécialistes de l'IST, des archives et des humanités numériques.

Associer les producteurs dès la collecte

Une fois l'enquête sélectionnée par le CST, la collecte est formalisée par la contractualisation avec le déposant et l'institution de rattachement. Le service d'archives ou de documentation compétent est également intégré si les matériaux y sont conservés. Les documents sont transférés au CDSP le temps de leur traitement et de leur mise en ligne. Un premier rendez-vous avec les déposants permet de faire le point sur les documents conservés, en s'assurant notamment de leur état, formats et de la faisabilité technique du traitement. Cette réunion est également l'occasion de collecter un maximum d'information pour comprendre les matériaux et plus généralement la réalisation de la recherche.

Les producteurs des enquêtes sont associés aux diverses étapes de leur traitement : le protocole d'anonymisation est validé sinon coconstruit avec le chercheur, qui participe aussi à la recontextualisation de l'enquête lors d'un entretien accordé à beQuali, et peut être sollicité ponctuellement pour identifier un document.

Valeur ajoutée d'une équipe plurimétiers

Le dispositif beQuali s'est construit, dès l'origine, sur la base d'une interaction de métiers — archivistique, sciences sociales et design Web — dans le but de pouvoir appréhender et maîtriser l'ensemble des logiques et des contraintes

4. HAL est l'archive ouverte pluridisciplinaire en France : <hal.archives-ouvertes.fr/>.

SPIRE est l'archive ouverte institutionnelle de Sciences Po : <spire.sciencespo.fr/>.

5. Le prochain appel à propositions aura lieu à l'automne-hiver 2017.

nécessaires à la réalisation. Ce parti-pris s'est traduit par la mise en œuvre d'une « polyphonie » indispensable à la bonne marche du projet (Both & Garcia 2014). Aujourd'hui, le dispositif repose sur une équipe projet de six personnes (documentalistes, archivistes, ingénieurs de recherche en sciences sociales), les opérations de design Web étant assurées par l'équipe informatique du CDSP. Une des spécificités de l'activité de traitement des enquêtes est que celle-ci est systématiquement réalisée par un binôme composé d'un profil archiviste/documentaliste et d'un profil sciences sociales. Traiter une enquête selon le modèle mis en place à beQuali suppose en effet de remplir un ensemble de tâches faisant appel à des expertises multiples, qui nécessitent de travailler de manière collaborative. Il faut par exemple pouvoir classer et préparer des documents papier pour la numérisation tout en organisant intellectuellement des corpus de documents souvent complexes et volumineux, en prenant bien soin en parallèle d'anticiper des problématiques comme celles de l'anonymisation ou de la contextualisation du processus de recherche. Derrière ces opérations, ce sont des savoirs et des savoir-faire très divers et étendus qui sont mobilisés dans les domaines des techniques de classement d'archives ou de contrôle qualité de numérisation, des standards internationaux de documentation des données, des théories et des méthodes des sciences sociales, etc. La stratégie de développement de beQuali s'est ainsi fortement appuyée sur la valeur ajoutée d'une équipe plurimétiers, gage de vitalité.

L'enquête qualitative : un corpus de données complexe

Les données collectées sont des corpus d'enquêtes qualitatives en sciences sociales (sociologie, science politique, ethnologie, notamment). Il s'agit de données hétérogènes et complexes à plusieurs titres.

Des données d'observation humaine, à caractère sensible

Tout d'abord, ces matériaux sont délicats de par leur nature. En effet, les enquêtes qualitatives renvoient à des données de terrain, c'est-à-dire recueillies selon des méthodes d'entretien approfondi, en face à face ou en groupe, ou d'immersion et d'observation ethnographique au plus près des enquêtés, souvent complétées par de la documentation. Ces données ainsi collectées, observées et enregistrées par un chercheur (et non par un instrument) le sont à partir d'une interaction humaine et sociale. Elles sont donc le plus souvent coproduites ou coconstruites par cette relation entre le chercheur et l'enquêté, ce qui induit un rapport intime et de confiance entre le chercheur et « ses » données, le chercheur et « ses » enquêtés. En outre, ces données interrogent des pratiques, des opinions et des situations sociales, touchant par exemple

à la politique, à la religion, à la sexualité, à la santé ou à l'ethnicité. S'attachant à décrire des personnes et résultant de récits de vie et de jugements, elles sont donc souvent de nature personnelles et sensibles. Leur mise à disposition doit répondre à des contraintes juridiques fortes (droit d'auteur, réglementation informatique et libertés⁶) visant à protéger les chercheurs et les enquêtés (ou les personnes citées).

Des archives scientifiques

Ensuite, ces données sont spécifiques en raison de leur origine et de leur usage. Elles sont en effet collectées dans les archives de chercheurs ou dans les fonds d'archives de leurs laboratoires, mais seulement une fois passée « la durée d'utilité scientifique » (Both & Cadorel 2014). Ce délai renvoie au temps pendant lequel le chercheur voudra exploiter ses données pour ses recherches présentes ou futures. Ce passage au statut « d'archives » peut correspondre au moment de la retraite ou bien au passage de la publication de la thèse. Il varie d'un chercheur à l'autre et dépend toujours d'une volonté personnelle de léguer et partager son travail à d'autres, avec l'idée d'en faire profiter des collègues étudiant des problématiques connexes.

Une unité archivistique originale qui prend forme grâce au numérique

L'unité documentaire traitée par beQuali se positionne à la croisée de plusieurs pratiques documentaires. Ce sont des archives au sens du Code du patrimoine qui les définit comme « l'ensemble des documents, quels que soient leur date, leur lieu de conservation, leur forme et leur support, produits ou reçus par toute personne physique ou morale et par tout service ou organisme public ou privé dans l'exercice de leur activité » (article L211 du Livre 2 du Code du patrimoine), mais le traitement opéré par beQuali sur ces archives déroge à certaines règles de la pratique archivistique. En effet, la constitution même des corpus d'enquêtes mêle plusieurs fonds d'archives : archives de chercheurs, de laboratoires, etc. Par la multiplicité de ses sources et l'originalité de son unité, le corpus « enquête » dans beQuali s'éloigne de la logique de producteur propre à l'archivistique et se rapproche davantage de celle de la constitution de collections documentaires regroupant des documents sur la base de certaines caractéristiques (formes, support, sujet, etc.). Pour autant, le respect des fonds n'est pas mis à mal, car les outils informatiques rendent possible la création d'un artefact numérique (Both & Cadorel 2014) guidé par une logique scientifique, celle de l'enquête de terrain, sans altérer le caractère organique des fonds d'archives conservés dans les institutions.

6. En France, la Commission nationale de l'informatique et des libertés (CNIL) veille au respect des droits de l'homme, à la vie privée et aux libertés individuelles ou publiques en matière de données informatiques : <www.cnil.fr/>.

Des corpus de recherche multiformes

Enfin, ces jeux de données sont particuliers par leur unité (de sens) et leur diversité (de formes). Dans beQuali, l'unité documentaire (ou archivistique) correspond à l'unité scientifique « enquête », c'est-à-dire qu'elle renvoie à une étude dont les bornes temporelles et le périmètre matériel sont parfois difficiles à restituer, et que seul le chercheur est en mesure de distinguer. Il peut en effet étudier un terrain géographique ou une population de manière espacée dans le temps (c'est le cas de l'enquête *Les réseaux économiques souterrains* de N. Murard et J. F. Laé, portant sur la revisite de familles à 30 ans d'intervalle), ou bien mener une recherche au long cours, dans plusieurs terrains complémentaires. Une enquête dans beQuali peut dès lors constituer un agrégat d'enquêtes et de terrains (comme l'enquête *Choisir son école* d'A. van Zanten, composée de cinq enquêtes menées de façon progressive et au gré des financements, ou celle *La profession d'architecte : nouvelle théorie sociologique des professions* de F. Champy dans lequel il réinvestit des matériaux d'enquêtes antérieures, depuis sa thèse.)

FIGURE 1

« Fouille » de documents au domicile du chercheur



Collecte par beQuali chez un chercheur déposant.

Le corpus de matériaux sélectionnés et mis à disposition correspond également à une masse de documents allant au-delà des données de terrain *stricto sensu*. Cette

conception large et exhaustive de la matérialité de l'enquête englobe des documents depuis la préparation du projet de recherche jusqu'à l'analyse des données et vise à résoudre la difficulté inhérente à la mise à disposition de données brutes dans une optique de réutilisation scientifique. La collecte du corpus (papier ou numérique) doit être la plus large possible afin de renseigner sur les différentes étapes du processus de recherche et de rendre possible la compréhension et l'exploitation des matériaux par d'autres chercheurs. Elle reflète ainsi les trois grandes étapes de la recherche : la préparation, la collecte et l'analyse des données. Parmi la diversité des documents collectés on peut citer par exemple des dossiers de financements, des devis, des comptes-rendus de réunion, des grilles méthodologiques, des transcriptions d'entretiens, des notes d'observation, des fiches de synthèse, des brouillons de rédactions intermédiaires.

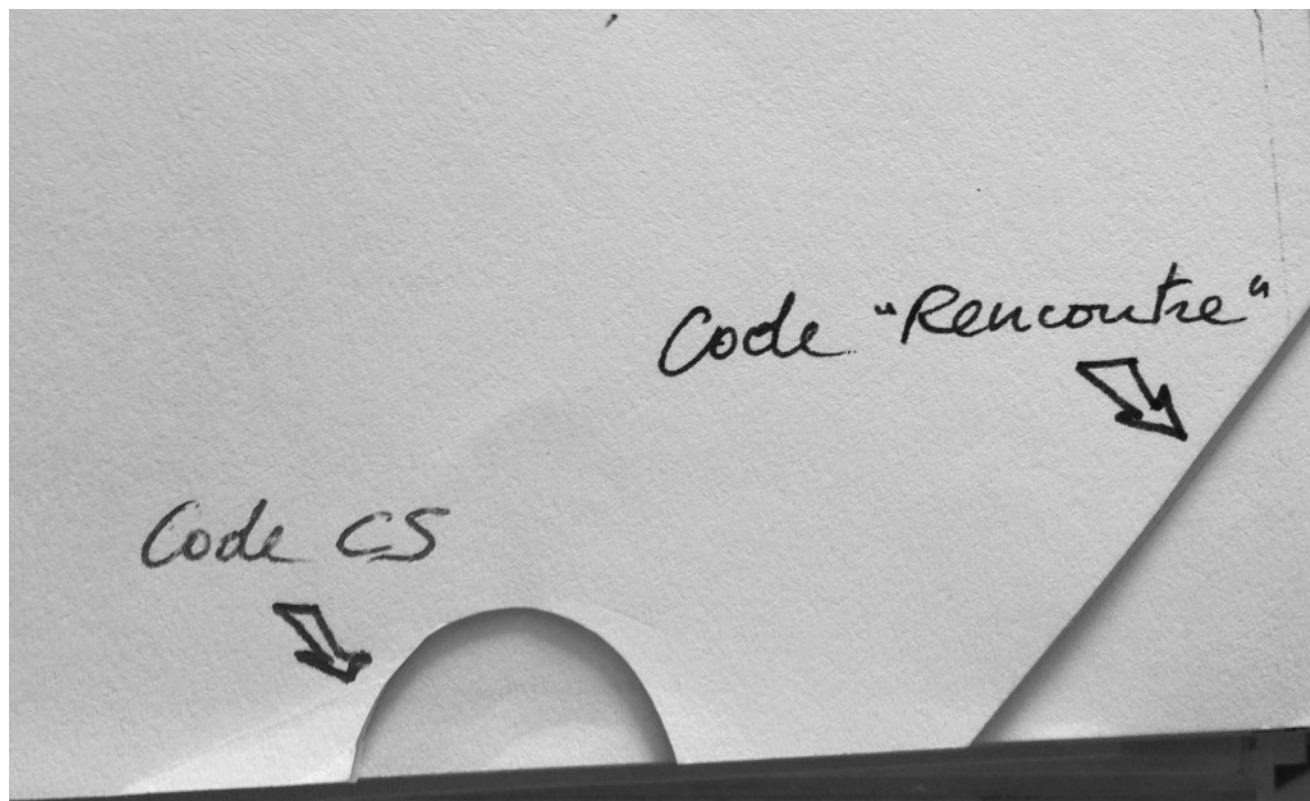
Au-delà des contenus, les documents d'une enquête qualitative sont enfin de volume et de support divers. Variant d'un à dix cartons, une enquête peut comporter des cahiers et feuilles volantes, des cassettes audio, des vidéos, des photographies, des schémas, des magazines, des brochures, des tracts, des fichiers numériques ou encore des assemblages et des collages de tout format. Ce corpus composite, de documents en plus ou moins bon état, implique une série de traitements archivistiques, documentaires et numériques afin d'en constituer une unité numérique signifiante, mise en ligne sur un site permettant leur consultation et leur exploitation scientifique.

Quels traitements documentaires ?

Pour répondre au double objectif patrimonial et scientifique, différentes étapes jalonnent le travail de documentation des données, depuis la collecte jusqu'à l'archivage des jeux de données mis en ligne. Ce travail est réalisé en collaboration avec les producteurs des données. Il vise à restituer le plus fidèlement possible la forme et le sens des données collectées et à contextualiser leur essence et leur genèse. Les enquêtes de beQuali sont documentées « sur mesure », empruntant autant aux outils maison conçus par l'équipe et adaptés aux spécificités des données traitées, qu'aux standards de documentation et d'archivistique dans un souci de pérennité et d'interopérabilité.

Décrire finement des corpus volumineux

Une fois l'enquête collectée, un inventaire pièce à pièce (parfois jusqu'à 800 documents) décrivant l'ensemble des documents est réalisé. Description, date, nom du fichier, état du document, nombre de pages, présence de données personnelles, mise en ligne sont autant de métadonnées qui y sont répertoriées. Cet inventaire est mis à disposition sur le site sous forme de tableur et d'un instrument de recherche arborescent (plan de classement) selon le standard EAD.



Extrait de l'enquête *La Formation des couples*, de M. Bozon et F. Héran.

C'est également à ce moment que la sélection des documents à mettre en ligne s'effectue. Ainsi, certains documents peuvent être écartés de la numérisation et de la mise en ligne pour des raisons de droit de communicabilité, d'édition ou de protection des données personnelles. Il s'agit par exemple des articles consultables en ligne sur des portails de revues scientifiques (comme Cairn.info, Persée, Revues.org) ou des archives de presse. Cela peut être le cas d'une liste de contacts d'enquêtés, contenant généralement des noms et coordonnées. Dans ce cas, afin de ne pas perdre trop d'information, les données non personnelles peuvent être reprises sous la forme d'un tableau créé par beQuali et mis à disposition sur le site. Ce travail de sélection du corpus est réalisé en collaboration avec le chercheur : pour repérer les données sensibles, pour trier différentes versions d'un même document et pour choisir les matériaux les plus pertinents en vue d'une réutilisation.

Classer en respectant l'intégrité du fonds

En parallèle de ce travail, les documents sont organisés selon un plan de classement défini par l'équipe beQuali, à partir des pratiques de plusieurs chercheurs. Cela permet d'homogénéiser le classement de chaque enquête selon

quatre niveaux d'arborescence : sous-fonds (préparation, collecte, analyse, compléments), dossier, sous-dossier puis article. Nous avons défini une trame commune de classement selon ces quatre niveaux avec un système de nommage unique que nous adaptons ensuite à chaque enquête. Chaque enquête étant un cas particulier, le plan de classement lui est spécifique. Ce principe nous permet de respecter l'intelligibilité de l'enquête, de garder la trace de l'organisation du chercheur tout en permettant le fonctionnement d'outils d'exploration sur un site web. Par exemple l'enquête *La Formation des couples* de M. Bozon et F. Héran a été menée en trois étapes : une enquête préliminaire par entretiens, une vague d'enquête par questionnaire puis une enquête post-quantitative par entretiens menés auprès de répondants aux questionnaires. Le plan de classement à facettes nous a permis de décrire et d'inscrire ces différentes étapes dans le nommage des documents.

Le plan de classement permet également à l'utilisateur de se repérer dans la masse de documents constituant une enquête. L'enquête *Choisir son école* réalisée par Agnès van Zanten comporte 476 documents. Il est donc indispensable de classer les documents et de leur attribuer un nommage compréhensible par tous. Un exemple de nommage peut

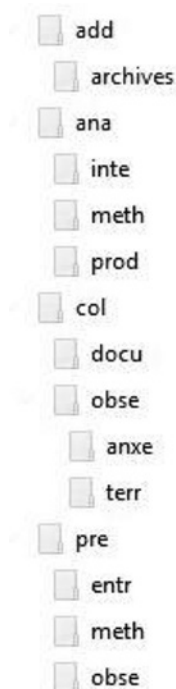
être «pre_meth_indv_notes» qui correspond aux notes, réalisées lors de la préparation de l'enquête, et relatives à un entretien individuel. Cela renvoie donc à l'élaboration de la méthodologie utilisée.

Préparer les documents à la numérisation

Les trombones, agrafes et *post-it* sont retirés afin de préparer les documents à la numérisation et également pour les conserver dans de bonnes conditions d'archivage. Chaque document est conservé dans une pochette en papier neutre. Les documents papier nécessitant d'être anonymisés le sont également à ce moment. Des *post-its* sont utilisés afin de cacher les informations personnelles ou confidentielles. L'anonymisation s'effectue également sur les documents nativement numériques. La plupart du temps, un hyperonyme du type «([anonym: nom enquêté])» ou «([anonym: nom lieu])» est utilisé pour remplacer l'information. L'anonymisation se fait au cas par cas après discussion avec le chercheur (Bendjaballah *et al.* 2018). Avant la numérisation, des réparations peuvent également être effectuées sur les documents si nécessaire.

FIGURE 3

Arborescence des fichiers d'enquête selon le nommage de beQuali



Collecte par beQuali chez un chercheur déposant.

Les documents constituant l'enquête et conservés sous format papier sont envoyés à un prestataire externe afin d'être numérisés aux formats ODT (pour l'édition des transcriptions), JP2 (pour l'archivage pérenne) et PDF (pour la

diffusion). Les cassettes audio sont également numérisées afin d'être conservées dans un format pérenne (WAV). Les enregistrements audio, qui posent des questions spécifiques d'anonymisation, ne sont actuellement pas mis en ligne sur le site beQuali. Une fois les documents numérisés, un contrôle qualité est effectué par l'équipe beQuali.

Documenter selon des standards internationaux etinteropérables

Les transcriptions d'entretiens sont ensuite éditées sous le format XML-TEI pour la mise en ligne. Cela permet de décrire la structuration du texte (tours de paroles entre les interviewés et les enquêteurs, éléments de paraverbal, etc.). L'enquête est également documentée selon les standards de métadonnées DDI⁷ et *Dublin Core*. L'affichage des métadonnées générales prend la forme d'une notice intitulée «l'enquête en bref» sur le site beQuali. Elle reprend les principales informations descriptives telles que l'auteur, les dates, les publications associées et le type et nombre de matériaux disponibles. Des informations de contexte de l'enquête (couverture géographique, population enquêtée, méthode de collecte, etc.) sont également renseignées au format DDI.

Contextualiser le travail de recherche : l'enquête sur l'enquête

En parallèle des différentes actions décrites ci-dessus, des entretiens sont réalisés avec le chercheur pour recontextualiser l'enquête dans une perspective de réutilisation scientifique. Ce travail donne lieu à la rédaction d'un rapport écrit intitulé «l'enquête sur l'enquête». Il vise à apporter un éclairage scientifique et une mise en perspective historique sur le travail mené par le chercheur. L'enquête sur l'enquête retrace d'abord la carrière de ce dernier, pour se concentrer ensuite sur la genèse de l'enquête, ses modalités de réalisation (orientations théoriques, démarche choisie, options méthodologiques et péripéties du terrain), le corpus conservé, l'analyse menée sur les matériaux, ainsi que sur sa postérité et ses possibles réutilisations. L'enquête sur l'enquête constitue un outil d'aide précieux pour comprendre et s'appropriier les documents d'une enquête. Elle est librement consultable sur le site beQuali sous deux formats : la lecture d'un rapport téléchargeable en format PDF ainsi que l'écoute en ligne de l'entretien oral avec le chercheur, monté et chapitré.

Une fois ces différentes étapes réalisées, les documents sont mis en ligne à la fois sur le site de beQuali (pour l'exploration des documents) et sur le portail français Quetelet⁸ d'accès aux

7. *Data Documentation Initiative* : standard utilisé pour les enquêtes par questionnaire, adapté par UK Data Service et beQuali aux enquêtes qualitatives.

8. <www.reseau-quetelet.cnrs.fr/spip/>.

données pour les sciences humaines et sociales (pour le téléchargement de tous les fichiers d'une enquête).

Archiver les matériaux originaux et les données enrichies

Lorsque le traitement de l'enquête est terminé et que celle-ci a été mise en ligne sur le site, les documents originaux sont renvoyés aux chercheurs ou aux institutions responsables des archives pour archivage physique. L'enquête est également archivée dans sa version numérique sous des formats pérennes. Cette opération fait l'objet d'un partenariat⁹ avec le Centre informatique national de l'enseignement supérieur¹⁰ (CINES) qui est un centre de calcul et d'archivage numérique ayant reçu l'agrément du Service interministériel des archives de France¹¹ (SIAF) pour prendre en charge l'archivage numérique des établissements de l'enseignement supérieur et de la recherche.

En assurant la conservation pérenne des enquêtes mises à disposition, beQuali entend couvrir l'ensemble de la chaîne documentaire liée à la mise à disposition des données. En effet, sont déposés au CINES les matériaux numériques ou numérisés des enquêtes ainsi que toute la documentation créée par beQuali. Les originaux des documents sont, quant à eux, retournés auprès des producteurs (chercheurs ou services d'archives des établissements) et conservés selon les règles du réseau des archives en France¹². Ce dispositif de mise à disposition fait l'originalité de beQuali. La banque d'enquête est actuellement la seule à proposer un archivage à proprement parler, et non le seul stockage sécurisé des données enrichies.

Le projet d'archivage pérenne est un processus assez long à mettre en place, car il requiert de l'information technique très précise afin d'assurer sur le long terme l'intégrité et l'intelligibilité des documents ainsi que la maintenance des formats. La première étape a été de définir l'unité documentaire à verser : un « paquet d'archives » correspondra à une enquête numérisée et documentée. Il a ensuite fallu déterminer — avec l'aide des équipes du CINES et en se basant sur leur outil de validation FACILE¹³ — la liste des formats de fichiers qui seront versés : PDF-A, ODS, XML, XML-TEI, ODT, CSV, JPEG-2000, PNG,

MPEG-4 et WAVE. Enfin, un profil de description des paquets d'archives a été défini. Cette trame, destinée à contenir les métadonnées spécifiques pour l'archivage au CINES, a été créée selon le Standard d'échange des données électroniques (SEDA) utilisé pour le versement d'archives publiques électroniques. L'information contenue dans ce « profil SEDA » est reprise dans les fichiers de métadonnées utilisés pour la mise à disposition sur le site beQuali. Le fichier de métadonnées pour l'archivage est créé automatiquement par un script informatique qui suit le *mapping* réalisé par l'équipe de beQuali entre les différentes métadonnées utilisées.

Quels usages ? Le cas de l'enseignement des méthodes en SHS

Une offre diversifiée

La banque d'enquêtes offre plusieurs perspectives d'utilisation, principalement à des fins de recherche et d'enseignement. L'objectif global est de faciliter la découverte et l'appropriation des enquêtes par les enseignants-chercheurs et les étudiants.

La première réponse à cet enjeu renvoie à la recherche de diversification du catalogue d'enquêtes proposées à la réutilisation, qui sont pour l'instant au nombre de huit (10 autres sont en cours de traitement). Comme le montre le tableau ci-dessous, beQuali propose un catalogue diversifié, que ce soit par les disciplines, les méthodes, les thèmes, les périodes ou encore les terrains. Cette diversité constitue un atout dans le cadre des enseignements de méthode en SHS en particulier.

TABEAU 1

Diversité du catalogue d'enquêtes beQuali

Période	Discipline
< 1980 : 1	sociologie : 12
1980 < 2000 : 9	science politique : 6
> 2000 : 8	
Comparaison	Méthode
internationale : 3	entretiens : 17
infranationale : 2	individuels : 15
	collectifs : 2
	générationnels : 1
	observations : 4
	documentation : 5

Répartition du nombre d'enquêtes en traitement par catégories.

9. Le CINES fonctionne comme un prestataire, ses services sont payants et sont fournis après la signature d'une convention d'archivage avec les services versants.

10. <www.cines.fr/>.

11. <francearchives.fr/article/26287441>.

12. Les archives d'universités et d'établissements de recherche publics sont des services d'archives intermédiaires ; ils sont soumis au contrôle scientifique et technique du SIAF et versent ensuite leurs archives aux Archives nationales ou aux Archives départementales selon leur localisation et leur statut.

13. Outil de validation de formats pérennes du CINES : <facile.cines.fr/>.

Par exemple, le catalogue permet de montrer la variété des méthodes de recherche, qu'il s'agisse de présenter la méthodologie de l'entretien semi-directif ou de l'observation. Il couvre aussi des thématiques recoupant les grandes disciplines enseignées en science politique ou en sociologie, que ce soit la sociologie des organisations (avec l'enquête *Les Ministères de l'Enseignement supérieur en France et en Allemagne*, de C. Musselin), la sociologie de l'immigration et de la famille (avec l'enquête *Deux générations d'immigrés africains*, de J. Barou), le genre (*Des femmes en politique*, de M. Sineau), les institutions politiques (*Le Parlement et l'Europe*, d'O. Rozenberg) ou encore le comportement électoral (*Les Français et la politique*, d'E. Schweisguth). Dans le cadre d'un enseignement, cette diversité des thèmes et des méthodes employées dans ces enquêtes permet de montrer aux étudiants la richesse des manières de faire une enquête de terrain en SHS. Elle leur permet de saisir non seulement les différentes méthodologies applicables à la recherche en SHS, mais aussi de comprendre qu'à des thèmes voisins ou similaires peuvent correspondre plusieurs démarches de recherche différentes. Par exemple, en Europe, les méthodologies employées par deux enquêtes de la banque sont l'entretien semi-directif (*Le Parlement et l'Europe*) et l'entretien collectif (*Quand des Français, des Anglais et des Belges parlent d'Europe*). Le catalogue des enquêtes permet ainsi de mener des analyses comparées et transverses, pour une meilleure compréhension de la recherche de terrain.

TABEAU 2

Diversité du catalogue d'enquêtes beQuali

	SP1	SP2	SP3	SP5	SP6	S1	S3	S5
1. Préparer son terrain								
Elaborer un guide d'entretien								
Constituer un échantillon								
2. Faire son terrain								
Evaluer l'effet enquêteur								
Mener des entretiens non directifs								
Mener des entretiens semi directifs								
Mener des entretiens avec des élites								
Mener des entretiens collectifs								
Faire de l'observation ethnographique								
3. Du terrain à l'analyse								
Comparer entretien et questionnaire								
Analyser les données								

Les références SP1, S1, etc. renvoient aux identifiants des enquêtes.

La deuxième réponse à cet enjeu pédagogique renvoie à la conception exhaustive que beQuali donne à chaque enquête, laquelle constitue un atout majeur dans l'utilisation de la banque à des fins d'enseignement. beQuali collecte en effet tous les documents de l'enquête, depuis la préparation du projet d'enquête jusqu'à l'analyse, permettant de renseigner sur les différentes étapes du processus de recherche. Comme le montre le tableau ci-dessous, chaque enquête de la banque permet ainsi de couvrir une phase de la recherche de terrain.

Des outils d'exploration pour favoriser l'appropriation des données

Afin de simplifier l'entrée dans les enquêtes, la plateforme beQuali offre plusieurs séries d'outils d'exploration en ligne.

Une première série d'outils accompagne la compréhension du corpus de documents qui forment l'unité « enquête ». Chaque enquête peut être consultée à partir d'outils d'exploration destinés à faciliter les tris dans ces ensembles de données. Actuellement, les utilisateurs peuvent filtrer ces collections de données en choisissant un ou plusieurs critères parmi trois possibilités : analytique, géographique et temporel. Pour optimiser le fonctionnement de ces filtres, une classification commune aux différentes enquêtes a été adoptée et chaque document est indexé en fonction de ces métadonnées. Le premier tri, analytique, utilise les trois principales catégories du plan de classement des documents décrit plus haut¹⁴. Ce plan de classement a été explicitement construit pour refuser la dichotomie « données/documentation¹⁵ ». Chacune de ces trois catégories permet d'explorer la composition du corpus de l'enquête selon des regards différents, qui sont suffisamment génériques et englobants pour permettre de classer tout type d'enquête, quelles que soient ses spécificités : les grandes étapes du processus de recherche (préparation, collecte, analyse), le type d'activité précisément impliqué par le processus de recherche (par exemple, l'administration du projet, sa documentation, la réflexion méthodologique, l'enregistrement d'observations sur le terrain, la production de résultats scientifiques, etc.) ou encore les types de documents, qui objectivent plus directement la matérialité de l'enquête (par exemple, les enregistrements audio, les transcriptions, les notes, les rapports, les articles, les communications, la correspondance, etc.). La répartition des documents selon ces trois

14. Précisément, les catégories sous-fonds, dossier et article vues plus haut.

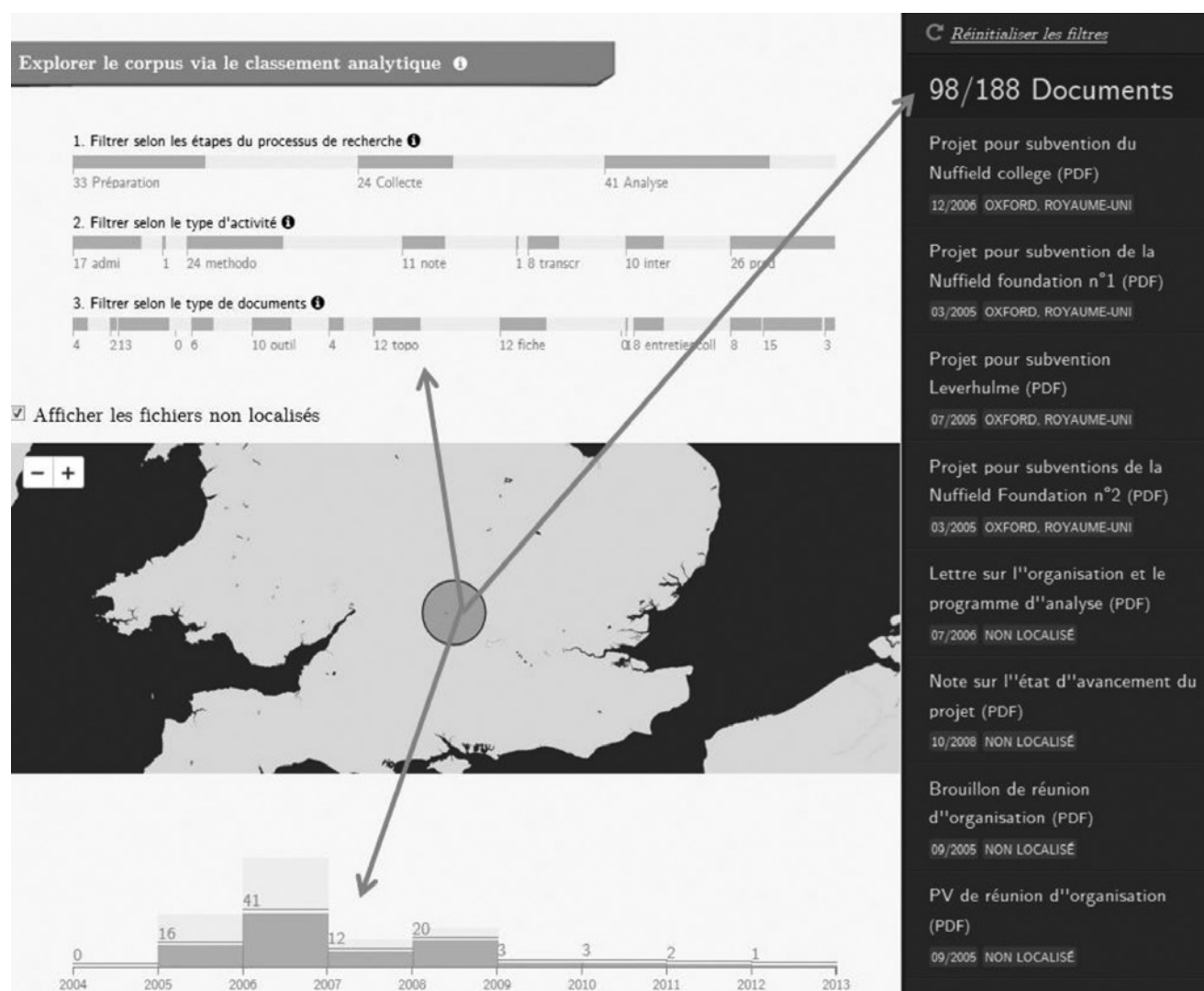
15. On ne trouve d'ailleurs nulle part le terme « données » dans le plan de classement et la « documentation » désigne uniquement la documentation compilée par le chercheur sur le terrain (par exemple, les documents fournis par les enquêtés).

niveaux du plan de classement permet de mieux appréhender la structure documentaire de l'enquête sans être enfermé dans un seul regard, celui d'une arborescence figée de catégories et de sous-catégories; un tel système a en effet l'inconvénient de prédéterminer les chemins d'exploration possibles, en canalisant l'approche du corpus par une série d'emboîtements successifs. Ici, au contraire, il est possible d'explorer l'enquête d'emblée par exemple du point de vue des types de documents et de voir comment ces types de documents se répartissent selon la phase de préparation, de collecte ou d'analyse — bref d'inverser l'approche spontanée qui consisterait à aller du plus général au plus spécifique. Cette approche a l'avantage d'être fidèle à l'enquête de terrain telle qu'elle se construit.

Par ailleurs, chaque document se voit attribuer d'autres métadonnées, telles que sa date et sa localisation géographique, quand l'information est disponible. La possibilité de répartition des documents sur une ligne de temps annuelle (les documents sont regroupés généralement par année) donne ainsi à voir la chronologie de l'enquête, et donc l'étalement dans le temps du processus de recherche. Leur répartition sur une carte donne quant à elle à voir la dimension géographique de l'enquête (les documents sont regroupés autour de l'unité géographique de taille la plus adaptée). Il est donc possible, quand cela est bien sûr pertinent, de manipuler la dimension comparative du processus de recherche. Ce prisme permet par exemple de montrer aux étudiants le caractère unique de chaque terrain.

FIGURE 4

Filtrage des documents par localisation géographique



Surtout, l'exploration du corpus est interactive : le principe de ce module est de dépasser les inconvénients d'une classification fixe en permettant à l'utilisateur de faire varier les modes de visualisation du corpus documentaire¹⁶. Fonctionnellement, les outils d'exploration permettent de filtrer dans le corpus, simultanément sur ces différents champs, analytique, temporel et géographique. Ces filtres sont interdépendants : toute action effectuée sur un des filtres se répercute automatiquement sur les deux autres. Concrètement, les trois visualisations sont affichées conjointement à l'écran. Si on filtre en sélectionnant une catégorie du plan de classement, par exemple les documents qui relèvent de la phase de « collecte », on obtiendra la répartition géographique et temporelle de ces documents sur les deux autres visualisations. Autre exemple, si on sélectionne une année donnée sur la frise temporelle, on obtiendra la répartition géographique et analytique de ces documents sur les deux autres visualisations.

Retour d'expériences

Grâce à la diversité et à la richesse aussi bien des documents et des informations disponibles sur les enquêtes que des outils d'exploration, il est possible d'engager avec les étudiants des réflexions approfondies sur la démarche d'enquête, par exemple une réflexion autour de la préparation du terrain, avec des questions telles que la construction et la sélection de l'échantillon, les modes de recueil des données ou encore le processus de production du guide d'entretien. Sur ce dernier point, les matériaux disponibles permettent souvent de montrer qu'un guide d'entretien se construit progressivement et nécessite plusieurs ajustements et corrections tout au long du terrain : par exemple, pour les enquêtes *Choisir son école* d'A. van Zanten et *Deux générations d'immigrés africains* de J. Barou, ce sont sept versions du guide d'entretien qui sont consultables. La banque permet aussi de montrer aux étudiants comment le travail d'analyse s'effectue concrètement : qu'il s'agisse de sociogrammes, de graphes, de typologies ou de fiches synthétiques, le catalogue des enquêtes en ligne permet d'expliquer comment traiter les données recueillies du terrain. Les documents d'analyse montrent aussi que ce travail se fait parfois avec tâtonnement, nécessitant un retour sur le terrain. En somme, l'ensemble des matériaux conservés permet de montrer que la recherche de terrain ne constitue pas une succession d'étapes distinctes et cloisonnées les unes des autres. Le fait de pouvoir manipuler aisément les documents selon le plan de classement, la date ou la localisation géographique vise à faciliter l'exploration de ces dimensions.

Les possibilités d'engager un travail de réflexion avec les étudiants sont renforcées par des outils permettant d'explorer les enquêtes à un niveau de granularité fin. En particulier, un

affichage dynamique des transcriptions d'entretiens est rendu possible par leur édition préalable en XML-TEI. Deux séries d'éléments sont ici exploitables : les « intervenants » (les locuteurs, autrement dit le ou les enquêteurs ainsi que les individus interrogés) et le niveau de finesse de la transcription (les conventions de transcriptions du niveau verbal et paraverbal). L'outil étant interactif, on peut choisir par exemple de n'afficher que les tours de parole de l'enquêté, ou ceux de l'enquêteur, ou encore de cacher tout ou une partie des indications du paraverbal. Cet outil permet ainsi de faire travailler les étudiants sur la relation entre enquêteur et enquêté, par exemple sur la nature des échanges ou le niveau de distance entre ces deux derniers. Enfin, l'ensemble des transcriptions étant téléchargeable dans des formats utilisables avec des logiciels CAQDAS (*Computer Aided Qualitative Data Analysis Systems*), les étudiants peuvent faire émerger des analyses à partir des données elles-mêmes.

FIGURE 5

Options d'affichage des transcriptions d'entretien

Options

Voir/cacher les intervenants

Protagoniste :

- ☒ Didier

Enquêté :

- ☒ Yannick
- ☒ Mohamed
- ☒ Malik
- ☒ Lamine

Enquêteur :

- ☒ Sabrina Aouici

Voir/cacher le paraverbal

- ☒ **Tout cocher**
 - ☒ “” Commentaire
 - ☒ ... Points de suspension
 - ☒ 📶 Ton
 - ☒ 😄 Rires
 - ☒ ⏸ Interruption
 - ☒ 🗣 Gestuelle
 - ☒ 🔒 Passage anonymisé
 - ☒ 🗨 Interaction
 - ☒ 🚫 Inaudible
 - ☒ ⌚ Durée ou heure

16. Fondé sur le modèle de l'*elastic list* : <stefaner.eu/projects/elastic-lists/>.

FIGURE 6

Affichage d'un fichier TEI sur le site

Mohamed.2184	☺ Il se croit encore ***.
Didier.2185	Tu diras ça dans dix ans ? ☺ 66.
Mohamed.2186	📶 'Nous les Noirs, nous les Noirs nou
Lamine.2187	📶 'Nous sommes forts, nous sommes :
Malik.2188	'Tu es Noir, tu resteras Noir mon fils'

Site beQuali.fr

L'équipe de beQuali a l'occasion d'expérimenter l'usage des enquêtes du catalogue dans le cadre d'enseignements de *master* de sociologie. Plusieurs thématiques ont pu être discutées avec les étudiants : par exemple, à l'aide des transcriptions d'entretiens et de la publication finale associée à l'enquête *Le Parlement et l'Europe*, les étudiants ont procédé à une analyse comparée des transcriptions d'entretiens avec la publication finale en réfléchissant à des questions telles que « Que reste-t-il de l'entretien ? Quelle forme et quel contenu ? Quelle méthode d'analyse ? ».

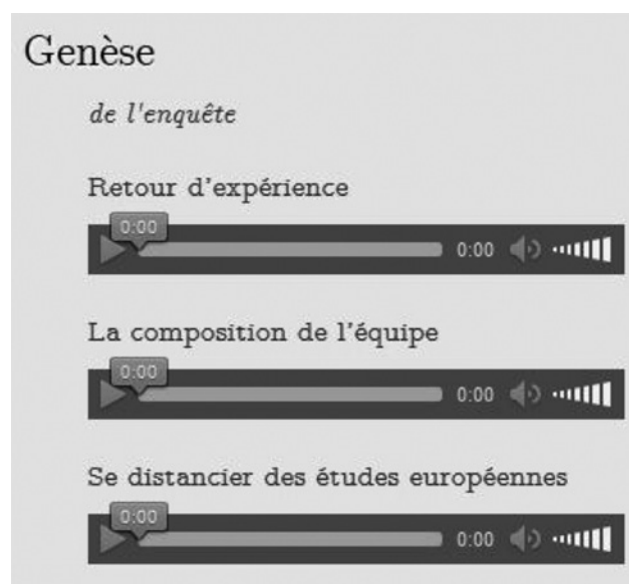
Ce travail les a conduits notamment à constater que dans les conclusions finales tirées par le chercheur, les entretiens étaient rarement mobilisés dans leur exhaustivité. Les premiers enseignements que l'équipe tire de cette expérience sont nombreux : d'abord, sans surprise, l'appropriation des matériaux pour un enseignement demande un investissement important, tant du côté des étudiants que des enseignants, qui ne seraient pas familiers des thèmes ou des méthodes développées dans les enquêtes. De ce point de vue, l'équipe a pu constater tout l'intérêt et l'importance du travail de contextualisation qu'elle opère : beQuali produit en effet des documents qui consistent en une compilation d'informations qui n'existent nulle part ailleurs dans le corpus de documents. Il peut s'agir dans certains cas de synthèses des principaux résultats de l'enquête (notamment des catégories utilisées pour réaliser des typologies ou des classifications), présentées sous forme de tableaux. Le plus souvent, il s'agit de listes récapitulant les personnes enquêtées, éventuellement caractérisées par leurs principales

propriétés sociographiques. Ces différents documents d'enrichissement des données, que ce soit l'inventaire détaillé ou l'enquête sur l'enquête, facilitent grandement l'appropriation des matériaux et la compréhension des enjeux théoriques et méthodologiques de l'enquête.

L'enquête sur l'enquête se présente comme un document particulièrement fructueux : pour saisir l'usage qui peut être fait des archives d'enquêtes pour son propre terrain, il est en effet nécessaire de comprendre comment le chercheur premier a entrepris sa propre recherche. Or, il y a généralement peu de documents (articles scientifiques notamment) dans lesquels le chercheur expose lui-même, de manière approfondie, ses propres choix, les difficultés et péripéties auxquelles il a fait face dans la constitution de son échantillon par exemple. De ce point de vue, le travail mené conjointement avec le chercheur dans le cadre de l'enquête sur l'enquête s'avère particulièrement bénéfique. Les étudiants et enseignants ont ainsi la possibilité d'écouter depuis le site de beQuali des extraits audio (cf. plus haut) dans lesquels le chercheur expose lui-même sa démarche, sans filtre d'analyse ou d'interprétation secondaire faite par un tiers.

FIGURE 7

Extraits audio de l'enquête sur l'enquête



Source : Site beQuali.fr

On comprend bien dès lors l'importance du travail de documentation réalisé à l'échelle de chaque document, en collaboration avec le ou les chercheurs premiers, qui seul permet d'aboutir à ce résultat. La définition d'un protocole de documentation commun, appliqué à chaque enquête

facilite la démarche pédagogique dans la mesure où elle permet de comparer les enquêtes du catalogue les unes avec les autres à partir de problématiques identiques. Elle permet ainsi de bien faire ressortir les différences et les similitudes au-delà de la spécificité de chaque terrain.

Ensuite, faire travailler des étudiants sur un terrain « tiers » favorise la mise à distance et le développement d'un regard critique sur la recherche, ce qui facilite une réflexion plus autonome sur l'enquête de terrain. Enfin et surtout, il ressort de ces premières expériences de très nombreux questionnements et usages possibles sur la démarche de recherche, indépendamment du nombre d'enquêtes disponibles dans le catalogue. Le séquençage de chaque enquête selon les différentes phases du terrain permet de croiser les thématiques et les « entrées » dans le catalogue : par exemple, en travaillant avec les étudiants sur la phase d'analyse, il est apparu qu'il y avait suffisamment de matériaux (qu'il s'agisse des synthèses familiales dans l'enquête *Deux générations d'immigrés africains* de J. Barou ou des graphes dans l'enquête *Représentations du champ social, attitudes politiques et changements socio-économiques* de J.M. Donégani, G. Michelat et M. Simon) avec deux ou trois enquêtes pour expliquer la variété des formes que peut prendre le processus de montée en généralité.

Conclusion

Les choix de développement de la banque d'enquêtes présentés dans cet article répondent aux trois enjeux initiaux qu'il s'agissait de concilier à l'origine du projet : sauver des enquêtes de la perte et de l'oubli à un moment où les données qualitatives sont encore très peu conservées en France ; proposer des outils dynamiques pour explorer ces jeux de données complexes et denses et pour se les approprier en vue de produire de nouvelles connaissances ; ouvrir à la communauté académique l'occasion de découvrir des matériaux inédits et intéressants d'un point de vue pédagogique. Les compétences mobilisées pour construire le dispositif, le mode de fonctionnement et les priorités mises en avant ont permis de construire un outil à la fois évolutif et pérenne, d'offrir à la communauté scientifique un catalogue d'enquêtes diversifié et riche, et de répondre au besoin de préserver et de valoriser des enquêtes qualitatives.

Comparativement à d'autres centres de données similaires et précurseurs, à l'image de ESDS Qualidata ou du Finnish Social Science Data Archive, beQuali privilégie une conception élargie de ce qui constitue une enquête, en aucun cas limitée aux « données brutes », quand bien même elles sont accompagnées de leur « documentation ».

Dans un souci constant d'accompagnement des utilisateurs, l'un des objectifs de beQuali est également d'enrichir les outils permettant d'explorer les enquêtes sous de nouveaux angles.

En complément, « l'enquête sur l'enquête » qui accompagne systématiquement chaque mise à disposition vise à restituer à l'utilisateur la dynamique de recherche ayant guidé le ou les auteurs de l'enquête ; elle a été pensée comme un outil permettant de découvrir l'enquête dans sa globalité

avant toute réutilisation de ses matériaux. En ce sens, beQuali a été globalement conçu dans un esprit attentif « à la contextualisation, la prise en main et l'exploration des corpus » (Duchesne & Garcia 2014).

Si le site est conçu pour un type de données particulier au départ (les enquêtes menées par entretiens), de

plus en plus d'enquêtes ethnographiques (menées par observation) sont entrées dans le catalogue et élargissent progressivement le plan de classement et les outils de visualisation. Ainsi, chaque nouveau type de corpus et de matériau amène son nouveau lot de questionnements méthodologiques pour ajuster au mieux les outils disponibles sur le site et présenter les documents de la manière la plus pertinente pour de nouveaux usages. De même les photos, vidéos et enregistrements audio sont voués à pouvoir être mis à disposition en ligne, dès lors que les conditions d'utilisation le permettent (accord des enquêtés pour la diffusion). En outre, si une sélection est indispensable pour pouvoir traiter des enquêtes dans une optique de réutilisation, le CST de beQuali n'a pas vocation à juger la valeur scientifique des enquêtes. L'appel à propositions est ouvert à tous types d'enquêtes entrant dans le champ des méthodes qualitatives et des sciences sociales et beQuali souhaite rester un entrepôt national, basé sur le dépôt volontaire, et un intermédiaire pour continuer à valoriser des archives sur le très long terme.

Dans un souci constant d'accompagnement des utilisateurs, l'un des objectifs de beQuali est également d'enrichir les outils permettant d'explorer les enquêtes sous de nouveaux angles. À cette fin, les premières expérimentations de la banque d'enquêtes dans un but pédagogique sont discutées dans le cadre d'un groupe de travail sur l'enseignement, mené à l'initiative de beQuali, et rassemblant des enseignants-chercheurs intéressés par la problématique de l'usage d'archives d'enquêtes pour leurs cours. Ces réflexions donnent lieu à un jeu de supports pédagogiques sur des questions pratiques d'enseignement qui seront disponibles prochainement sur le site de beQuali. Une journée d'étude organisée à l'initiative de beQuali permettra de faire la synthèse de ces premières conclusions et de dégager des pistes pour développer des corpus d'enquête simplifiés et « prêts à l'emploi » de façon à proposer des modules permettant aux enseignants et aux étudiants de se saisir de manière plus didactique des enquêtes disponibles.

Par ailleurs, beQuali entend permettre les revisites de terrain et les réanalyses de corpus constitués pour certains il y a plusieurs dizaines d'années. À l'instar d'initiatives déjà réalisées (Abrial *et al.* 2017), la démarche de beQuali favorise en effet la réappropriation des matériaux afin d'éclairer autrement les terrains et les analyses qui en sont faites. Il s'agira de développer ces premières expérimentations en sensibilisant notamment les chercheurs et doctorants à l'occasion de journées d'études ou de conférences. beQuali est déjà engagée sur ce terrain avec des communications réalisées dans le cadre des réseaux Mate-SHS (Réseau Méthodes Analyses Terrains Enquêtes en SHS) et ECPR (European Consortium for Political Research).

À ce titre, l'équipe de beQuali participe régulièrement¹⁷ à des journées d'étude thématiques dans le cadre de réseaux professionnels, notamment sur la confidentialité des données, les archives de la recherche, l'enseignement ou encore l'analyse secondaire. L'équipe intervient régulièrement dans les laboratoires de sciences sociales pour sensibiliser la communauté scientifique à la conservation des données de recherche et accompagner les chercheurs et les professionnels de l'IST qui souhaitent préparer ou anticiper le partage d'enquêtes sur un entrepôt.

17. Rubrique « Actualités » du site beQuali : <bequali.fr/fr/actualites/>.

SOURCES CONSULTÉES

- Abrial, Stéphanie, Mathieu Brugidou & Annie-Claude Salomon. 2017. Types idéologiques et classe résiduelle dans l'enquête d'Étienne Schweisguth : les Français et la politique, 1982-1988. Réanalyse à partir de deux familles de logiciels, CAQDAS et ADT. *Recherches Qualitatives, ARQ Association pour la Recherche Qualitative, La réanalyse des enquêtes qualitatives à l'épreuve de l'expérimentation*, Hors-Série « Les Actes » (21) : 29-53.
- Barou, Jacques (coord.), Sabrina Aouici, Claudine Attias-Donfut, Rémi Gallou & Nicole Lapierre. 2017. De l'Afrique à la France, d'une génération à l'autre (2006-2010). Centre de données sociopolitiques, beQuali. <bequali.fr/fr/les-enquetes/lenquete-en-bref/cdsp_bq_s5>.
- Bendjaballah, Selma, Sarah Cadorel, Émilie Fromont, Guillaume Garcia, Émilie Groshens & Émeline Juillard. 2018. Problèmes d'anonymat et de confidentialité des données : l'expérience et les solutions mises en œuvre par beQuali. In *Éthique et droit*. Aix-en-Provence : Presses universitaires de Provence.
- Bishop, Libby. 2006. A proposal for archiving context for secondary analysis. *Methodological Innovations Online* 1 (2) : 10-20. <www.ukdataservice.ac.uk/media/262210/miobishop_pp10_20.pdf>.
- Bishop, Libby. (2014). Re-using qualitative data : A little evidence, on-going issues and modest reflections. *Studia Sociologica* 3 (214) : 167-176. <www.data-archive.ac.uk/media/492811/bishop_reusingqualdata_stsoc_2014.pdf>.
- Both, Anne & Sarah Cadorel. 2015. Pour en finir avec l'original ? Des effets du numérique sur les archives scientifiques : le cas de beQuali. In *Frontières d'archives : recherches, mémoires, savoirs*, sous la direction de Jean-François Bert & Marc J. Ratcliff. Paris : Éditions des Archives Contemporaines, 157-164.
- Both, Anne, Sarah Cadorel. 2014. En quête d'enquêtes pour mise à disposition : la collecte auprès de chercheurs en SHS : le cas de beQuali. In *Conférence International Council on Archives 2014*, Section des universités et institutions de recherche, 8-10 juillet. Paris : ICA. <spire.sciencespo.fr/hdl:/2441/7n44mitjqu918a1nhp0u-l6u8bg>.
- Both, Anne & Guillaume Garcia. 2014. Le chercheur, l'archiviste et le webmaster : la polyphonie patrimoniale ? Le cas de beQuali, banque d'enquêtes qualitatives en sciences sociales. In *Heritage and Digital Humanities : how should training practices evolve ?*, sous la direction de Bernadette Dufrene. Berlin : Lit Verlag, 353-364.
- Bozon, Michel & François Heran. [à paraître]. La formation des couples (1983-2007). Centre de données sociopolitiques, beQuali. <bequali.fr/fr/les-enquetes/lenquete-en-bref/cdsp_bq_s2>.
- Champy, Florent. 2017. La profession d'architecte : nouvelle théorie sociologique des professions (2004-2010). Centre de données sociopolitiques, beQuali. <bequali.fr/fr/les-enquetes/lenquete-en-bref/cdsp_bq_s3>.
- Donégani, Jean-Marie, Guy Michelat & Michel Simon. 2015. Représentation du champ social, attitudes politiques et changements socio-économiques (1977-1981). Centre de données sociopolitiques, beQuali. <bequali.fr/fr/les-enquetes/lenquete-en-bref/cdsp_bq_sp5>.
- Duchesne, Sophie, Élisabeth Frazer, Guillaume Garcia, André-Paul Frogner, Florence Haegel & Virginie van Ingelgom. 2013. Quand des Français, des Anglais et des Belges (francophones) parlent d'Europe (2005-2012). Centre de données sociopolitiques, beQuali. <bequali.fr/fr/les-enquetes/lenquete-en-bref/cdsp_bq_sp1>.
- Duchesne, Sophie & Guillaume Garcia. 2014. beQuali : une archive qualitative au service des sciences sociales. In *Archives de la recherche. Problèmes et enjeux de la construction du savoir scientifique*, sous la direction de Marie Cornu, Jérôme Fromageau & Bertrand Müller. Paris : L'Harmattan, 35-56. Consulté le 25 juillet 2017. <halshs-00922690>.
- Murard, Numa & Jean-François Laé. [à paraître]. Les réseaux économiques souterrains (1967-2012). Centre de données sociopolitiques, beQuali. <bequali.fr/fr/les-enquetes/lenquete-en-bref/cdsp_bq_s7>.
- Rozenberg, Olivier. 2014. Le Parlement français et l'Union Européenne : l'Europe saisie par les rôles parlementaires (1993-2005). Centre de données sociopolitiques, beQuali. <bequali.fr/fr/les-enquetes/lenquete-en-bref/cdsp_bq_sp3>.
- Schweisguth, Étienne. 2013. Les Français et la politique (1982-1988). Centre de données sociopolitiques, beQuali. <bequali.fr/fr/les-enquetes/lenquete-en-bref/cdsp_bs_sp2>.
- Sineau, Mariette. 2016. Des femmes en politique (1984-1985). Centre de données sociopolitiques, beQuali. <bequali.fr/fr/les-enquetes/lenquete-en-bref/cdsp_bq_sp6>.
- van Zanten, Agnès. 2016. Choisir son école (1999-2007). Centre de données sociopolitiques, beQuali. <bequali.fr/fr/les-enquetes/lenquete-en-bref/cdsp_bq_s1>.

Équipe beQuali (France) L'équipe de la banque d'enquêtes qualitatives beQuali est composée de Selma Bendjaballah et Guillaume Garcia, ingénieurs de recherche et docteurs en sciences politiques, de Sarah Cadorel et Emilie Fromont, archivistes, et d'Emilie Groshens et Emeline Juillard, documentalistes. L'équipe collecte, enrichit et valorise des archives d'enquêtes qualitatives pour leur mise à disposition de la communauté scientifique sur le site beQuali.fr. En lien avec ce projet, l'équipe participe au consortium ArchiPolis, qui œuvre pour le repérage et la sauvegarde des archives d'enquêtes en sciences sociales du politique. Les membres de l'équipe mènent également des activités de formation auprès d'enseignants chercheurs et de professionnels de l'information scientifique, et participent aux réflexions et activités de réseaux professionnels sur les archives scientifiques (AAF) ou les méthodes en sciences sociales (Mate SHS, ECPR).

Lyne DA SYLVA (Canada) a une formation pluridisciplinaire en mathématiques, informatique, linguistique et linguistique informatique. Après des années de travaux de recherche et de développement en traitement automatique de la langue, elle a été nommée professeure à l'École de bibliothéconomie et des sciences de l'Université de Montréal, où elle est actuellement professeure agrégée et responsable du nouveau diplôme d'études supérieures spécialisées (D.E.S.S.) en édition numérique offert par la Faculté des arts et des sciences.

Ses activités d'enseignement comprennent l'indexation et la condensation (manuelle et automatisée) et la sémiotique des systèmes d'information. Ses recherches couvrent divers aspects de la gestion de l'information numérique, y compris l'automatisation de l'indexation et de la condensation, les bibliothèques numériques et la gestion de différents types de données (surtout, données liées et données de recherche).

À titre de chercheuse, elle est membre de l'OLST (Observatoire de linguistique Sens-Texte), du Département de linguistique et de traduction et du Centre de recherche interdisciplinaire sur les humanités numériques (CRIHN) de l'Université de Montréal. Elle siège sur divers comités dédiés aux données : le groupe francophone de l'Initiative canadienne sur les données liées, le Groupe d'experts en Web sémantique (données liées) dirigé par BAnQ et le Groupe d'experts sur la recherche et l'intelligence (RIEG) de Portage (sur les données de recherche); elle est aussi impliquée dans le groupe d'intérêt SIG-DL (*Digital Libraries*) de l'ASIS&T.

Danielle DENNIE (Canada) est bibliothécaire responsable des communications savantes et de la réserve électronique ainsi que bibliothécaire en mathématiques et statistiques à l'Université Concordia. Elle détient une maîtrise en microbiologie appliquée de l'Institut Armand-Frappier et une maîtrise en bibliothéconomie et sciences de l'information de l'Université McGill.

Marielle St-Germain (Canada) est étudiante au doctorat à l'École de bibliothéconomie et des sciences de l'information (EBSI) de l'Université de Montréal.

Ses intérêts de recherche généraux sont essentiellement l'appropriation des technologies du Web sémantique et du Web de données en milieux documentaires, le développement de plateformes facilitant la découvrabilité des contenus et la découverte fortuite ainsi que la valeur pratique des données ouvertes. Elle occupe aussi un poste à la Direction de l'architecture et de la conception à Bibliothèque et Archives nationales du Québec (BAnQ) où elle a l'occasion d'explorer ces questions dans un cadre pratique.

Marc TANTI (France) est chef de projets de recherches en sciences de l'information et de la communication au sein de l'UMR 912-SESSTIM (Marseille-France). Il est responsable adjoint du service de veille sanitaire du Centre d'Epidémiologie et de Santé Publique des Armées Françaises. Il enseigne à Aix-Marseille Université et à l'École du Val-de-Grâce. Il est chercheur associé à l'IRSIC- EA-4262 (Marseille-France). Auteur de plus d'une centaine de publications, ses travaux de recherches portent sur la veille sanitaire de défense et l'exploitation des données massives dans le domaine de la santé publique.

Réjean SAVARD (Canada) détient un doctorat en bibliothéconomie de l'Université de Toronto et une maîtrise en bibliothéconomie de l'Université de Montréal. Il est associé à l'École de bibliothéconomie et des sciences de l'information (EBSI) de l'Université de Montréal depuis de nombreuses années en tant que professeur et chercheur. Il est l'auteur de plusieurs communications scientifiques et professionnelles, tant sur le plan national que sur le plan international. Très impliqué dans les milieux professionnels (Fédération internationale des associations et institutions de bibliothèques [IFLA], Corporation des bibliothécaires professionnels du Québec [CBPQ], Association pour l'avancement des sciences et des techniques de la documentation [ASTED]), il a été vice-président du conseil d'administration de la Grande Bibliothèque du Québec à sa création. Enfin, il est le président fondateur de l'Association internationale francophone des bibliothécaires et documentalistes (AIFBD).

INFONUAGIOUE



001

002

003

004

004.1

004.2

004.3

004.5

Disponible à 004.6782

La Classification décimale Dewey^{MD} : toujours à jour, pertinente et indispensable.

De nouveaux concepts apparaissent constamment – comme *le régime à base d'aliments crus*, *l'infonuagique*, *la psychologie positive* et *l'hameçonnage* – et les éditeurs de la Classification décimale Dewey^{MD} demeurent à l'affût. Ils s'assurent que les révisions reflètent les besoins actuels des bibliothèques et de leurs usagers.

WebDewey[®] en français

Obtenez en ligne l'index complet de la 23^e édition de la Classification décimale Dewey (CDD23), des indices construits supplémentaires et la possibilité d'ajouter aux indices de classification vos propres notes ajustées aux pratiques de votre institution.

La WebDewey[®] est un outil de classification à valeur ajoutée. Elle permet une navigation plus rapide, mais aussi, grâce à ses fonctionnalités automatisées, d'éliminer certaines tâches répétitives et ainsi d'améliorer l'efficacité au travail.

Abonnez-vous !

Pour informations : communiquez avec Sylvie Langlois
514-281-5012 | info@asted.org | www.asted.org



RÉPERTOIRE DE VEDETTES-MATIÈRE

Les thésaurus du RVM

Pour vos besoins d'indexation

- Le **RVM (Sujets)** : plus de 295 000 vedettes-matière et subdivisions pour indexer les ressources par sujet
- Le **RVMGF** : plus de 1700 descripteurs pour indexer les ressources par leur genre/forme
- Le **RVMEM** : plus de 820 descripteurs de moyens d'exécution en musique pour indexer les partitions musicales et les enregistrements sonores
- Le **RVMGD** : plus de 1050 descripteurs de groupes démographiques pour définir le public cible et les auteurs ou collaborateurs des ressources

Des mots pour tout

rvmweb.bibl.ulaval.ca



UNIVERSITÉ
LAVAL