Article

« La gestion des données de recherche en bibliothèque universitaire »

Alex Guindon

Documentation et bibliothèques, vol. 59, n° 4, 2013, p. 189-200.

DOCUMENTATION BIBLIOTHÈQUES

La gestion des données de recherche en bibliothèque universitaire

Alex Guindon
Bibliothécaire responsable des services de données
Université Concordia
alex.guindon@concordia.ca

Résumé | Abstract | Resumen

L'augmentation exponentielle de la quantité de données de recherche produites et réutilisées par les chercheurs pose des défis importants aux bibliothèques universitaires. Il s'agit pour celles-ci de mettre sur pied des services de gestion de données intégrés au cycle de vie de la recherche. Cela n'est possible qu'en assurant une collaboration active avec une série d'acteurs internes et externes, et en développant une formation spécialisée au sein des écoles de sciences de l'information. Cet article décrit le contexte de cette transformation, et identifie les principales activités et les responsables pour chaque étape du cycle de vie de la recherche.

The Management of Research Data by a University Library

The exponential growth in the quantity of data produced and used by researchers has given rise to important challenges for university libraries. They must create a means to integrate and manage the data in accordance with the life cycle of the research activities. This can only be achieved by fostering an active collaboration amongst the internal and external stakeholders and by developing a specialised training in schools of information science. This article describes the context in which this transformation must take place and identifies the principal activities and stakeholders for each stage of the research life cycle.

Acerca de la gestión de datos de investigación en las bibliotecas universitarias

El aumento exponencial del número de datos de investigación que los investigadores producen y reutilizan plantea importantes desafíos para las bibliotecas universitarias, que se ven obligadas a implementar servicios de gestión de datos integrados al ciclo de vida de las tareas de investigación. Sin embargo, esto solo se puede lograr mediante una colaboración activa con determinados participantes internos y externos, y desarrollando una capacitación especializada en las escuelas de ciencias de la información. Este artículo describe el contexto de esta transformación e identifica las principales actividades, así como los responsables de cada etapa del ciclo de vida de la investigación.

Introduction

Il n'est pas exagéré d'affirmer que les bibliothèques universitaires sont en période de rapide transformation. Dans le contexte de la dématérialisation des collections et de l'avènement du tout-numérique, ses rôles traditionnels — développement des collections, catalogage de documents, organisation de l'information, services de référence — se tournent vers de nouveaux objets et reposent sur de nouvelles méthodes. Une des grandes tendances est la création de dépôts institutionnels gérés par la bibliothèque. On y vise la préservation et la diffusion de documents numériques issus de la production intellectuelle des chercheurs : thèses, mémoires, articles, rapports, littérature grise, voire même production artistique (images, vidéos). Or, si la majorité de ces documents sont produits vers la fin du processus de recherche — même la littérature grise et les comptes rendus de conférence sont généralement produits tard dans le cycle de la recherche —, il existe un type d'information qui se retrouve, lui, plutôt en amont de ce cycle : les données de recherche.

Cet article se veut une introduction à la gestion des données de recherche en milieu universitaire. Après avoir présenté le contexte dans lequel se situe cette entreprise, nous en identifierons les principaux enjeux en mettant en évidence l'importance croissante qu'ont les données dans le processus de la recherche scientifique. Enfin, nous tenterons d'identifier les principaux acteurs appelés à jouer un rôle dans ce nouveau champ des sciences de l'information. Pour chaque étape de ce qu'il est convenu d'appeler le cycle de vie de la recherche, nous présenterons les principales activités liées à la gestion des données ainsi que les responsables potentiels de chacune de ces activités.

Un déluge de données

Ces dernières années, de nombreux articles ont mis en évidence la croissance exponentielle de la quantité de données produites par les entreprises, les gouvernements et, ce qui nous intéresse particulièrement, les chercheurs. Il est devenu banal de parler d'un *déluge de données*. Selon certaines estimations (Milner 2009, 83), les besoins en matière de stockage de données numériques augmentent de 127 % par année. Il s'agit non pas d'une évolution du processus de recherche scientifique,

mais bel et bien d'une révolution qui fait de la réutilisation des données scientifiques et de l'infrastructure informatique, qui permet la recherche et l'exploitation intensive de ces données, les nouveaux piliers de l'innovation scientifique (Microsoft 2006). Ce nouveau modèle scientifique est connu sous le vocable d'*eScience*. Richard Luce définit ce nouveau paradigme ainsi :

> *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.*
>
> (2008, 42)

À ce stade-ci, il convient de mieux définir ce que nous entendons par « données de recherche ». Il existe de multiples définitions, mais du point de vue de la gestion et de l'archivage des données, il apparaît utile d'adopter une vision très large. La définition du Conseil de recherches en sciences humaines (CRSH) va en ce sens :

> *Ces données comprennent des ensembles quantitatifs de données sociales, politiques et économiques, des renseignements qualitatifs sous forme numérique, des données de recherche expérimentale, des bases de données d'images et de sons fixes et mobiles, ainsi que d'autres objets numériques utilisés à des fins d'examen analytique.*
>
> (2012)

On voit que le terme « données de recherche » englobe toutes les formes que peuvent prendre les éléments de base sur lesquels repose la recherche, que celle-ci soit quantitative, qualitative ou expérimentale. Il est important de préciser que les produits de la communication scientifique — articles, monographies, thèses ou comptes rendus de conférence — *ne constituent pas* des données. Cette distinction entre données et publications est fondamentale pour les sciences de l'information parce qu'elle indique clairement que nous avons affaire à un nouveau type d'objet avec lequel nous sommes beaucoup moins familiers. Si nos méthodes de traitement des objets traditionnels de la communication scientifique sont bien au point, fortes de l'expérience accumulée au cours de multiples décennies de recherche et de pratique, cela est beaucoup moins vrai pour les nouveaux types d'objets numériques, particulièrement pour les données de recherche scientifique.

Il existe donc un manque d'expertise non seulement chez les bibliothécaires et archivistes, mais au niveau des chercheurs eux-mêmes qui, bien qu'ayant une connaissance poussée et spécialisée des méthodes de collecte de données et de l'utilisation de celles-ci au sein de leur discipline scientifique, ne sont généralement pas au fait des questions entourant l'emploi des métadonnées[1], l'organisation intellectuelle de l'information et la préservation des documents numériques. Qui plus est, le temps et les ressources leur manquent pour pouvoir gérer leurs données de façon optimale, sans compter que la motivation à partager les données de recherche est souvent absente dans un contexte où la « publication » de celles-ci ne contribue que peu ou pas du tout à la reconnaissance professionnelle des chercheurs.

Au vu de tous ces éléments, il n'est pas surprenant de constater que peu de chercheurs décident de partager ou même d'archiver leurs données. Ainsi, les résultats d'un sondage (Perry 2008, 3) effectué auprès de 175 chercheurs canadiens en 2006 révèlent que 54,7 % d'entre eux n'avaient pas l'intention d'archiver leurs données. Dans le domaine des sciences sociales, la réalité est encore plus sombre : en 2009, l'Association des bibliothèques de recherche du Canada estimait (ABRC 2009) que seulement 10 % des professeurs ayant reçu des subventions de recherche du CRSH se conformaient aux exigences de l'organisme en matière d'archivage et de partage de données. En théorie, le CRSH exige que toutes les données soient archivées et qu'elles soient partagées au plus tard deux ans après la fin du projet de recherche. Malheureusement, à ce jour, le CRSH n'a pas la capacité ou la volonté de s'assurer que les chercheurs respectent ces exigences. Une explication possible de ce retard dans les sciences sociales est liée à la nécessité de préserver l'anonymat des sujets de recherche et d'obtenir leur permission pour la diffusion des données.

La situation n'est pas unique au Canada : un rapport publié en Grande-Bretagne (Waller & Sharpe 2006) relève une série de faits préoccupants en ce qui concerne la préservation des documents numériques (notamment les données de recherche). Ainsi, 28 % des répondants (archivistes, bibliothécaires, responsables des technologies de l'information (TI), mais aussi producteurs de données) indiquent avoir perdu des données (29 % des répondants n'étaient pas en mesure de répondre à la question) alors que 48 % craignent la disparition de données (27 % n'étaient pas en mesure de répondre à cette question). Plus récemment, une étude américaine (Jahnke, Asher & Keralis 2012) basée sur une série d'entrevues menées auprès de professeurs et d'étudiants aux cycles

1. Bien que le terme « métadonnées » soit ici employé dans un sens général, nous l'utiliserons plus loin dans le contexte précis des données de recherche. Sans entrer dans les détails techniques, disons simplement que, dans ce contexte, on peut suivre le modèle de l'Open Archival Information System (OAIS) qui définit quatre types de métadonnées : 1) le contenu d'information (l'objet lui-même ainsi que les informations minimales permettant de l'utiliser); 2) l'information de pérennisation (les détails techniques et administratifs portant sur l'archivage); 3) l'information d'empaquetage (définit les relations entre l'objet lui-même et les métadonnées associées) et ; 4) l'information descriptive (le dictionnaire de données, par exemple).

supérieurs dans le domaine des sciences sociales confirme les défis que pose la gestion des données. On y note que peu de chercheurs se préoccupent de l'archivage de leurs données de recherche et qu'aucun de ceux interviewés n'avait reçu une formation en gestion de données. Qui plus est, les participants sont davantage préoccupés par la publication d'articles et ne s'intéressent à la création de métadonnées ou de documentation que dans la mesure où cela les aide à publier leurs résultats. Bref, en l'absence d'incitatifs professionnels — l'évaluation des professeurs reposant avant tout sur la publication d'articles — ou d'obligations liées aux subventions de recherche, il serait surprenant que les chercheurs consacrent les efforts nécessaires à une tâche complexe et de longue haleine telle que l'archivage et le partage des données scientifiques[2].

Dans ce contexte, il est clair que les chercheurs ne peuvent seuls mener à bien toutes les tâches liées à la gestion des données. Comme nous le verrons plus loin, il s'agit d'une tâche trop lourde pour être assumée par un groupe unique. Il s'agit plutôt d'une entreprise qui doit être menée par un ensemble d'acteurs incluant chercheurs, assistants de recherche, bibliothécaires, personnel des dépôts de données et personnel du soutien technique. Cet article vise notamment à jeter un éclairage sur les tâches que les bibliothécaires universitaires peuvent jouer dans ce domaine. Quoique notre expérience à ce jour soit limitée, nous disposons d'atouts importants qui devraient nous permettre de participer pleinement à ce nouveau champ de pratique. D'abord, les bibliothèques et les centres d'archives sont des institutions reconnues pour leur rôle dans la préservation et le partage des connaissances. Nous disposons non seulement de connaissances pertinentes — par exemple, en matière d'organisation de l'information, de préservation des documents et de création de métadonnées —, mais nous avons aussi la confiance des chercheurs dans ces domaines d'expertise.

Les enjeux de la gestion des données de recherche

Même si nous n'en sommes qu'aux balbutiements de ce que d'aucuns appellent « science des données » (*data science*), on peut au moins se consoler en se disant qu'il y a bel et bien une prise de conscience dans le monde universitaire : comme le soulignait en 2009 un éditorial de la revue *Nature* (« Data's Shameful Neglect » 2009), il y a urgence. Au Canada, plusieurs groupes issus du milieu de la recherche se sont donné le mandat de sensibiliser les décideurs, de créer des réseaux de bonnes pratiques et d'assurer la formation de spécialistes de la gestion de données. Pour ce qui est des bibliothèques universitaires, c'est notamment le Sous-comité sur la gestion des données de l'ABRC qui joue ce rôle[3].

Comme il est impensable de faire des progrès significatifs sans la participation active de la communauté scientifique, une des premières tâches du groupe a été de faire la promotion de la gestion des données en identifiant les enjeux liés à l'archivage et au partage des données. Le Sous-comité (ABRC 2009) met de l'avant six avantages liés à la bonne gestion et à la réutilisation des données.

1. *Accélérer le progrès scientifique*
   La pleine réalisation du potentiel de l'eScience implique non seulement que les données soient archivées de façon adéquate, mais aussi que des normes reconnues de métadonnées soient adoptées, que les ensembles de données soient bien documentés et qu'ils soient aisément repérables.
2. *Accroître la visibilité et les retombées de la recherche*
   Les résultats des premières enquêtes sur les effets positifs du partage des données sont encourageants : ainsi, l'étude de Piwowar, Day et Fridsma (2007) indique que les articles dans le domaine des essais cliniques sur le cancer, pour lesquels les données ont été publiées, reçoivent 70 % plus de citations que ceux qui ne sont pas accompagnés de leurs données de recherche. Ces résultats s'ajoutent aux études portant sur l'effet positif du libre accès sur le niveau de citation des publications scientifiques[4]. À plus long terme, l'enjeu central sera sans doute de revoir le modèle d'évaluation du travail universitaire pour que la publication des données de recherche et leur citation par des pairs soient reconnues au même titre que les publications traditionnelles.
3. *Assurer le respect des politiques des organismes subventionnaires*
   Le Canada est signataire de la Déclaration sur l'accès aux données de la recherche financée par des fonds publics[5] de l'Organisation de coopération et de développement économiques (OCDE) et la plupart des organismes subventionnaires (notamment le CRSH, le Conseil de recherches en sciences naturelles et en génie (CRSNG), et les Instituts de recherche en santé du Canada

2. Les lecteurs qui voudraient explorer davantage le point de vue des chercheurs sur la gestion des données pourront consulter, en plus de l'étude de Jahnke *et al.* (2012), les articles de Perry (2008) et de Scaramozzino, Ramirez et McGaughey (2012).

3. L'organisme Données de recherche Canada (<http://rds-sdr.cisti-icist.nrc-cnrc.gc.ca>) et le Comité national canadien pour CODATA (<http://www.codata.org/canada>/) se penchent aussi sur la question.

4. Le Open Citation Project a créé une importante bibliographie à ce sujet : <http://opcit.eprints.org/oacitation-biblio.html>.

5. <http://acts.oecd.org/Instruments/ShowInstrumentView.aspx?InstrumentID=157&Lang=fr&Book=False>.

(IRSC)) ont des politiques concernant l'archivage ou le partage des données. Par ailleurs, il y a fort à parier que le pays suivra la voie tracée par le Royaume-Uni et les États-Unis où les chercheurs sont tenus de présenter un plan de gestion de données au moment de leurs demandes de subvention.

4. *Limiter la répétition des travaux de recherche*
   Comme l'originalité de la recherche est un critère de base de l'évaluation des publications scientifiques, il est crucial que les chercheurs soient au fait des études effectuées dans leur domaine. L'idéal serait de mettre en place un réseau disciplinaire de centres de données reposant sur des normes de métadonnées partagées, des systèmes interopérables, des identificateurs d'objet numérique pour les ensembles de données et des outils de découverte en ligne.
5. *Faciliter la reproduction et la validation des résultats de la recherche*
   L'utilité de l'archivage et de la publication des données pour la validation et la reproduction des résultats est évidente dans le domaine des sciences naturelles puisqu'il s'agit de données expérimentales. Le concept de reproduction des résultats est moins pertinent en sciences sociales ou humaines, des domaines où les données sont généralement basées sur différentes techniques d'observation et ne sont pas recueillies dans des conditions expérimentales contrôlées. Par contre, disposer d'un ensemble de données permet tout au moins de reproduire les analyses et de vérifier qu'il n'y a pas eu d'erreurs méthodologiques.
6. *Intensifier la coopération entre les chercheurs*
   Il s'agit non seulement de créer des réseaux entre chercheurs d'une même discipline, mais aussi de favoriser la coopération entre spécialistes de domaines différents. Cette approche transdisciplinaire n'est possible que si les données respectent des normes de métadonnées reconnues et sont accompagnées d'une documentation détaillée. En effet, les techniques d'organisation et de description des données reposant sur une connaissance tacite particulière à un domaine d'expertise, elles ne sauraient suffire à la réutilisation des résultats par des chercheurs ne disposant pas de ces connaissances. Dans un même ordre d'idées, la collaboration entre disciplines scientifiques implique des systèmes informatiques et des formats de fichiers interopérables.

Le cycle de vie de la recherche et le processus de gestion des données

Pour bien saisir les diverses tâches et responsabilités liées à la gestion des données de recherche, il importe d'avoir une vue d'ensemble du processus de recherche lui-même et de comprendre ce qui se passe du point de vue des données à chaque étape de ce qu'il est maintenant convenu d'appeler le cycle de vie de la recherche. Comme nous l'avons mentionné plus tôt, les bibliothécaires ont l'habitude d'exercer leur expertise en aval du processus de recherche, c'est-à-dire une fois les résultats publiés dans un article, une monographie ou présentés lors d'une conférence scientifique. Mais, pour ce qui est de la gestion des données, le travail commence dès le début du processus de recherche. En ce sens, l'approche du cycle de vie de la recherche (et du cycle de vie des données qui en découle) se rapproche davantage du travail des archivistes que de celui des bibliothécaires. 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 que les anglophones appellent *data science*. Mais si le cycle de vie des documents en archivistique est généralement linéaire et unidirectionnel — on passe des archives courantes aux archives intermédiaires et finalement aux archives définitives —, le cycle de vie des données et celui du processus de recherche qui le sous-tend sont réellement circulaires et pas toujours unidirectionnels.

Plusieurs auteurs et institutions ont proposé des modèles décrivant soit le cycle de la recherche, soit celui des données qui en découle. Il n'existe pas un modèle unique qui peut décrire de façon détaillée et définitive le déroulement complexe de la recherche scientifique de la première hypothèse jusqu'à la publication des résultats. D'abord, ce processus n'est pas identique d'une discipline à l'autre. Ensuite, même au sein d'une discipline particulière, les projets peuvent différer et les détails du déroulement de la recherche ne sont pas nécessairement les mêmes. Ainsi, certains projets débuteront par la réutilisation de données préexistantes alors que d'autres sont fondés uniquement sur la création de données originales. En somme, l'objectif de ces modèles, dans la perspective de la gestion des données, est de présenter un schéma simple qui permette aux praticiens (bibliothécaires, archivistes et autres) de : 1) situer dans le temps les différentes actions nécessaires à la documentation, la préservation et la réutilisation des données ; 2) établir une liste de ces tâches pour chaque étape de la recherche ; 3) assigner les responsabilités pour chacune de ces tâches. Sarah Higgins souligne que la nature des documents numériques — le fait qu'ils puissent si facilement être copiés, déplacés ou effacés — ajoute à l'importance d'un modèle basé sur le cycle de vie de la recherche :

> *If data is not managed from the point of creation onwards, and the correct activities undertaken at the relevant points in the data's lifecycle, the ability to look after it successfully may be greatly diminished.*
>
> (2012, 18)

Un des objectifs de cet article est de présenter un modèle simple du cycle de vie de la recherche, en l'occurrence celui proposé par Ann Green et Myron Gutmann (2007), pour donner des exemples d'activités de gestion des données qui peuvent être entreprises tout au long du cycle et pour identifier les principaux partenaires, au sein des universités, qui prendront part à la gestion des données de recherche. Il ne s'agit pas de dresser une liste exhaustive des activités requises à la bonne gestion des données, mais plutôt de donner une idée des grandes étapes du processus et de démontrer la nécessité de compter sur la collaboration de plusieurs partenaires.

Un modèle du cycle de vie de la recherche en sciences sociales

Le modèle de Green et Gutmann (2007) est inspiré de la recherche en sciences sociales, mais son niveau d'abstraction assez élevé rend possible son utilisation pour décrire, du point de vue de la gestion des données, toute activité de recherche peu importe la discipline concernée. Comme nous l'avons mentionné plus haut, en plus des modèles de cycle de vie de la recherche, il existe plusieurs modèles basés sur le cycle de vie des données elles-mêmes. Ces deux types de modèles sont intimement liés et représentent en fait le même processus vu de deux angles légèrement différents. Les modèles qui portent sur la recherche décrivent le processus de la recherche du point de vue des chercheurs. On y recense les principales activités de recherche ayant une incidence sur la création, la réutilisation, l'analyse, l'archivage et le partage des données. Ils permettent d'avoir une vision globale du processus de recherche et d'établir les grandes lignes d'un programme de gestion des données. Les modèles portant sur le cycle de vie des données de recherche, quant à eux, adoptent le point de vue des spécialistes de l'information (bibliothécaires, archivistes, informaticiens)[6]. Ils sont généralement plus élaborés et permettent d'arrimer un grand nombre d'activités à chacune des étapes du cycle de vie. Puisque cet article se veut une introduction à la discipline de gestion des données de recherche, il est

Figure 1
Cycle de vie de la recherche en sciences sociales
(Adapté de Green et Guttman 2007)

davantage pertinent de présenter un modèle tiré du cycle de vie de la recherche qui permettra d'identifier les grands types de tâches à accomplir plutôt que de s'attarder à une description précise d'un grand nombre d'activités souvent assez techniques.

Le modèle proposé par Green et Gutmann (Figure 1) se décline en cinq phases : 1) découverte et planification ; 2) collecte initiale de données ; 3) préparation des données et analyse ; 4) publication et partage ; 5) gestion à long terme. Puisqu'il s'agit bel et bien d'un *cycle* de vie, les étapes 5 et 1 sont liées. Ainsi, à l'étape 1, les chercheurs peuvent faire usage des données existantes dont le partage a été effectué à l'étape 4 et l'archivage à l'étape 5. Nous allons maintenant explorer ces étapes une à une en présentant quelques-unes des tâches particulières à chacune d'entre elles ainsi qu'en identifiant les principaux partenaires qui joueront un rôle dans l'exécution de ces tâches. Il s'agit ici de présenter les partenaires internes (au sein de l'université), c'est-à-dire le chercheur, ses assistants de recherche, le bibliothécaire responsable des données, les bibliothécaires responsables de certaines disciplines, le bureau des technologies de l'information (TI), etc. Il existe aussi des partenaires externes tels que les dépôts disciplinaires, les agences subventionnaires (CRSH, CRSNG, etc.), les réseaux ou consortiums d'universités ou de centres de recherche ; mais dans le cadre de cet article, hormis quelques mentions des dépôts disciplinaires, nous préférons nous en tenir aux seuls partenaires internes par souci de concision. Le tableau à la page suivante (Figure 2) regroupe ces différents éléments pour chaque étape du cycle de vie.

6. Deux excellents exemples de ce type de modèle sont celui de l'Interuniversity Consortium for Political and Social Research (ICPSR 2012) et celui, plus généraliste, du Data Curation Centre (DCC) tel que présenté par Sarah Higgins (2008, 2012). Le document produit par l'ICPSR s'adresse aux chercheurs ou aux bibliothécaires qui les appuient afin qu'ils puissent préparer leurs données pour les déposer dans un centre spécialisé (comme l'ICPSR lui-même). Le modèle du DCC s'adresse aux institutions universitaires (principalement en Grande-Bretagne) qui veulent établir des centres d'archives de données.

Figure 2
Éléments liés aux étapes du cycle de vie de la recherche

Phases du cycle de vie	Tâches	Partenaires
Découverte et planification	• Identification de sources de données secondaires • Préparation du plan de gestion des données • Choix d'un dépôt de données • Préparation du formulaire de consentement • Identification des risques liés à la confidentialité	• Chercheur • Bibliothécaire de données • Bibliothécaire disciplinaire • Comité d'éthique
Collecte initiale de données	• Liens entre bases de données (données originales et données secondaires) • Choix d'un schéma de métadonnées • Formation des assistants • Structure des fichiers • Élaboration de la documentation des données • Choix des logiciels • Contrôle des versions de la base de données et vérification de son intégrité	• Chercheur • Assistant de recherche • Bibliothécaire de données • Bibliothécaire disciplinaire ou de métadonnées • Bureau des TI (bibliothèque et université)
Préparation des données et analyse	• Collaboration entre les chercheurs (environnement de recherche virtuel) • Contrôle des versions de la base de données et vérification de son intégrité • Vérification finale des données • Copie de sécurité de l'ensemble de données final	• Chercheur • Assistant de recherche • Bibliothécaire de données • Bureau des TI (bibliothèque et université) • Dépôt institutionnel
Publication et partage	• Dé-identification des données • Préparation du Paquet d'informations à verser (PIV) • Préparation du Paquet d'informations archivé (PIA) • Préparation du Paquet d'informations diffusé (PID) • Création d'identificateurs d'objet numérique (DOI, handle, etc.)	• Chercheur • Bibliothécaire de données • Dépôt disciplinaire, institutionnel ou associé à une revue scientifique
Gestion à long terme	• Liens entre données et publications • Gestion des citations • Migration des formats • Exécution de la stratégie de sauvegarde des données • Mise en valeur des ensembles de données • Vérification de l'intégrité des données	• Bibliothécaire de données • Bibliothécaire disciplinaire • Bureau des TI (bibliothèque et université) • Dépôt disciplinaire, institutionnel ou associé à une revue scientifique

Étape 1 : Découverte et planification

Lors de cette étape initiale, le chercheur élabore sa problématique de recherche sur la base des connaissances théoriques et empiriques de sa discipline. On ne saurait surestimer l'importance de cette étape du point de vue de la gestion des données. En effet, les décisions qui seront prises ici auront une influence déterminante tout au long du cycle de la recherche. Une bonne planification permettra au chercheur et aux autres partenaires d'économiser énormément de temps et d'efforts. De plus, certaines décisions sont quasi irréversibles et, dans le pire des scénarios, pourront rendre difficiles, voire impossibles, l'archivage et le partage des données. Bien qu'une intervention plus tardive, en aval du cycle de vie, puisse souvent permettre de mitiger les problèmes dus à une mauvaise planification, cela se fait au prix d'efforts additionnels et au risque de se retrouver avec des données incomplètes ou de la documentation partielle.

L'une des premières décisions du chercheur dans le processus de la recherche consiste à déterminer s'il réutilisera des données secondaires (données existantes produites par d'autres chercheurs ou par lui-même lors d'un cycle de recherche précédent), ou si sa recherche nécessitera la production ou la collecte de nouvelles données. Dans le cas où il y a réutilisation de données secondaires, le bibliothécaire responsable des données, avec l'aide possible de bibliothécaires disciplinaires, peut jouer un rôle important dans l'identification de ces données et leur obtention. Dans le cas où de nouvelles données sont produites (ces deux scénarios ne sont pas mutuellement exclusifs), d'autres décisions s'imposent. Idéalement, il faut choisir dès cette étape un dépôt pour l'archivage et le partage des données. Il peut s'agir d'un centre de données spécialisé (ICPSR, GenBank, Astronomical Data Archives Center, etc.), d'un dépôt institutionnel, du dépôt associé à la revue où l'on compte publier l'article de recherche (bien qu'à ce jour peu de revues offrent ce service) ou même d'une solution en ligne gérée par le chercheur lui-même (comme Dataverse). Ici encore, les bibliothécaires peuvent aider le chercheur à repérer les différentes possibilités qui lui sont offertes.

Par ailleurs, la plupart des projets de recherche sont financés par de grands organismes subventionnaires fédéraux ou provinciaux. Comme mentionné précédemment, ces agences ont souvent des politiques quant à l'archivage et au partage des données qui déterminent, du moins en théorie, ce que les chercheurs doivent faire de leurs données à la fin du cycle de recherche. De plus, il est probable que dans un proche avenir, les agences canadiennes exigent des chercheurs qu'ils produisent des plans de gestion de données comme elles le font déjà aux États-Unis et en Grande-Bretagne. Dans cette éventualité, les services de données au sein des bibliothèques universitaires devront offrir de l'aide ou même des outils en ligne pour faciliter la préparation de ces documents. Plusieurs universités et centres de données offrent déjà ce genre de services[7]. Au Canada, ces outils devront évidemment être adaptés aux nouvelles exigences des agences subventionnaires. Soulignons l'importance de ces plans de gestion des données qui sous-tendent le processus de gestion des données tout au long du cycle de vie de la recherche et imposent aux chercheurs de définir dès le départ des éléments cruciaux tels que le choix d'un dépôt, la durée de l'archivage des données et les critères de partage des données. La rédaction d'un tel plan oblige le chercheur à réfléchir à ces questions et à commencer à planifier la création et la gestion des données de recherche dans l'optique de leur préservation et de leur partage. Ils permettent aussi d'entamer le dialogue entre chercheurs, bibliothécaires et responsables de dépôts de données.

Un autre élément important à considérer à cette étape pour tous les projets qui portent sur des sujets humains est la création d'un formulaire de consentement à l'intention des participants. Il faut réfléchir à tout ce qui a trait à l'anonymat de ces répondants. Encore une fois, il est crucial de garder en tête la question du partage des données dès cette étape initiale, car une fois les formulaires de consentement signés par les participants, il est impossible de changer les règles de confidentialité à moins de créer de nouveaux formulaires et de demander aux participants d'approuver les nouvelles règles *a posteriori*. Ainsi, si l'on désire partager ses données avec d'autres chercheurs, voire avec le grand public, il est impératif de rédiger les formulaires de façon à ce que cette possibilité soit mentionnée de façon explicite. Il importe aussi d'identifier tous les risques liés à l'identification des participants et de commencer à réfléchir à des stratégies pour minimiser ces risques dans le cadre du partage des données. Ces questions touchent directement au champ d'expertise des comités d'éthique de la recherche des différentes universités. Ces comités sont des partenaires incontournables pour la gestion des données de recherche puisqu'ils définissent les règles qui régissent la protection des participants et la confidentialité des données. On devra donc établir un dialogue entre chercheurs, bibliothécaires et membres du comité d'éthique pour définir des règles (et les formulaires de consentement qui en découlent) qui permettent à la fois de protéger l'anonymat des participants et de faciliter le partage et la réutilisation des données. Soulignons que les règles actuelles devront évoluer puisqu'elles n'ont généralement pas été développées dans une optique de partage des données.

Étape 2 : Collecte initiale de données

Du point de vue des données, cette étape est fondamentale. C'est ici que le chercheur commence sa collecte de données et qu'il établit des liens entre sources de données secondaires et nouvelles données[8]. Dans l'intérêt du chercheur et de son équipe de recherche, et toujours en prévision d'un partage ultérieur des données, on doit prendre des décisions quant au schéma de métadonnées utilisé et à la documentation nécessaire à l'interprétation et à la validation des données. On doit également choisir les logiciels qui seront utilisés pour la cueillette des données (pour la saisie automatique d'entrevues ou les sondages sur le Web, par exemple), pour leur nettoyage et leur analyse (SPSS, SAS, STATA pour les sciences sociales, logiciels spécialisés liés aux instruments et langages de programmation en sciences de la nature). Finalement, on doit décider quelles variables mesurer et quelles variables secondaires dériver des résultats obtenus.

Idéalement, on doit documenter tout ce travail dès le départ, et ce, d'une façon systématique et précise. Malheureusement, dans les faits, les chercheurs négligent cette documentation par manque de temps ou de personnel formé pour accomplir ces tâches. Ce manque d'intérêt relatif à la documentation et à l'organisation des données s'explique facilement dans un contexte où l'évaluation du travail des chercheurs passe d'abord et avant tout par la publication rapide des résultats de recherche, sans égard au partage des données. Dans un nouveau paradigme où on accorderait davantage d'importance à la publication des données, les attitudes seraient sans doute différentes. Il reste que, même dans la situation actuelle, une mauvaise documentation des données et des fichiers risque d'apporter son lot de confusion, de duplication du travail et même éventuellement d'erreurs d'interprétation, surtout dans le cas d'équipes de recherche regroupant plusieurs chercheurs et assistants. Par exemple, en l'absence de procédures

7. Voir par exemple des outils tels que le DMPTool du University of California Curation Center <https://dmp.cdlib.org/>, le DMP Online du Data Curation Centre <https://dmponline.dcc.ac.uk/> ou des gabarits comme celui de l'Université de Melbourne: <http://www.eresearch.unimelb.edu.au/__data/assets/word_doc/0007/163366/RDMS_2009-05-05_DMP_Template_v0_2.docx>.

8. Ce travail peut prendre plusieurs formes selon la discipline et le type de recherche. Il s'agit d'enrichir les nouvelles données en leur donnant un contexte fourni par des bases de données préexistantes, soit de créer un véritable arrimage entre les objets (ou sujets en sciences sociales) au moyen d'identificateurs (champs) communs. Incidemment, pour les études portant sur des sujets humains, la multiplication des bases de données et la possibilité de les croiser posent des défis importants au niveau de l'anonymat des répondants.

précises pour identifier les versions successives des fichiers de données, les différents membres de l'équipe risquent de confondre les fichiers ou même de détruire des données cruciales.

Outre les membres de l'équipe de recherche, qui sont les intervenants susceptibles de jouer un rôle à cette étape de la recherche et quel peut être leur apport exactement ? Pour ce qui est du choix d'un schéma de métadonnées, le bibliothécaire responsable des données ou un bibliothécaire spécialisé en métadonnées (si la bibliothèque a la chance d'en avoir un) est souvent la personne idéale pour conseiller le chercheur. Bien que le nombre de normes de métadonnées bien établies et utilisées à grande échelle demeure assez limité (on mentionnera le Data Documentation Initiative (DDI) en sciences sociales, le Content Standard for Digital Geospatial Metadata (CSDGM) pour les données géospatiales et le Ecology Metadata Language (EML) en sciences de la vie), il existe en fait une grande quantité de ces schémas[9]. De plus, il existe des normes générales pour les objets numériques telles que le Dublin Core. La quantité et la complexité de ces normes rendent la collaboration de l'équipe de recherche avec les spécialistes de l'information quasi incontournable. En pratique, puisque le choix d'un schéma de métadonnées est souvent tributaire du centre de données choisi pour l'archivage et le partage des données, ce dialogue se fera avec les responsables du dépôt choisi. En effet, les dépôts de données exigent généralement un type de métadonnées précis ou, du moins, un nombre minimal d'éléments descriptifs qui en sont dérivés.

Le bibliothécaire de données pourra aussi jouer un rôle important pour ce qui est de l'organisation et de la nomenclature des fichiers de données. Il s'agit ici d'établir une convention pour les noms de fichiers qui tienne compte des multiples versions et des dates de création des fichiers, et d'élaborer une structure pour l'organisation des fichiers. En outre, il faut s'assurer que les noms de fichiers soient lisibles par les différentes plateformes informatiques (Windows, Unix, Linux, Mac). Le bibliothécaire, en collaboration avec le bureau des TI, peut aussi conseiller les chercheurs en ce qui concerne l'adoption d'outils pour la gestion des versions de fichiers et l'intégrité de celles-ci. En effet, il existe des solutions techniques qui automatisent la gestion des multiples versions des fichiers et qui assurent la préservation intégrale du contenu : on est donc certain qu'il n'y aura ni modification accidentelle ni perte de données pour des raisons techniques (en informatique, cette vérification de l'intégrité d'un fichier est appelée *checksum*).

Finalement, cette étape du processus de recherche où l'on démarre le projet et où l'on embauche les assistants se prête bien à des activités de formation en gestion des données. Ces ateliers peuvent être dirigés par le bibliothécaire de données et s'adressent généralement aux étudiants des cycles supérieurs, qui sont souvent employés comme assistants de recherche. Cependant, les nouveaux chercheurs peuvent aussi bénéficier de ces formations. Puisque la nature des projets de recherche est assez différente selon les disciplines, la formation devra idéalement s'adresser à un groupe de chercheurs assez homogène. On pourrait par exemple donner une formation en sciences sociales et une autre en sciences de la nature, voire même des formations plus spécialisées. Cela implique que les bibliothécaires qui dirigent ces activités soient non seulement compétents en gestion des données, mais qu'ils aient aussi une bonne compréhension du type de recherche effectué dans le domaine sur lequel porte la formation. Une solution serait évidemment de compter sur une équipe de bibliothécaires regroupant les bibliothécaires spécialistes d'un sujet (qui auraient idéalement une certaine expérience de recherche) et le bibliothécaire responsable des données. On touche ici à un des grands défis de la gestion des données au sein des bibliothèques universitaires : la nécessité de compter sur des professionnels alliant connaissances techniques et expérience pratique de recherche.

Étape 3 : Préparation des données et analyse

À ce stade, le chercheur effectue les dernières vérifications et modifications des données, et entreprend le travail d'analyse. Il commence aussi la mise en forme des résultats et la rédaction d'articles et de présentations. Au sein d'une équipe de recherche, le travail d'analyse implique une collaboration constante entre différents chercheurs et assistants parfois issus d'institutions diverses. Cela suppose un accès facile mais contrôlé aux données et cela, si possible, à partir d'ordinateurs situés hors campus. Dans plusieurs cas, ce scénario idéal n'est pas respecté, car l'institution ne dispose pas de l'infrastructure technique nécessaire. En l'absence de ce qu'il est convenu d'appeler un « environnement de recherche virtuel » (ERV), les différents membres de l'équipe doivent souvent créer de multiples copies des fichiers de données et de résultats d'analyse. Évidemment, cela implique des risques — confusion entre les différentes copies et versions des fichiers, risques associés à la sécurité et la confidentialité des données — et une diminution certaine de l'efficacité.

Revenons rapidement à la question des environnements de recherche virtuels. Pour comprendre le rôle des différents acteurs dans la mise sur pied et le soutien technique d'un ERV, il importe de bien saisir de quoi il s'agit. De façon générale, on peut définir un ERV comme étant « *a set of online tools and other network resources and technologies interoperating with each*

9. Le Digital Curation Centre recense des normes de métadonnées en biologie, en sciences de la terre, en sciences sociales et humaines, en sciences physiques ainsi que des schémas généraux : <http://www.dcc.ac.uk/resources/metadata-standards>.

other to support or enhance the processes of a wide range of research practitioners within and across disciplinary and institutional boundaries » (JISC 2006, 1).

Toujours selon le Joint Information Systems Committee (JISC 2006, 2), les principaux types de fonctions qui peuvent être facilités par un ERV comprennent la gestion des données (production, recherche et analyse), la création collective des produits de la recherche ainsi que les activités liées à la gestion de projet. Comme on peut le voir, l'utilité de ce type de technologie va au-delà de la gestion des données au sens strict : elle englobe tout le cycle de vie de la recherche. Du point de vue des données, un tel système ne doit pas nécessairement remplir toutes ces fonctions, mais il doit au minimum permettre la collaboration à distance et en temps réel pour ce qui est de la production, du traitement et de l'analyse des données[10]. La mise sur pied et la gestion d'un tel système, qui repose généralement sur l'arrimage de plusieurs technologies différentes comme une plateforme de dépôt (EPrints, DSpace, Fedora), une technologie de gestion des versions de fichiers, une solution pour la gestion des métadonnées, etc., n'est pas simple et implique un rôle central pour le bureau des TI de l'université. Le rôle du bibliothécaire des données dans ce contexte est plutôt celui d'un intermédiaire entre chercheurs et ERV, et passe par la formation et l'aide aux usagers. L'adoption d'ERV par les bibliothèques ou par les universités elles-mêmes en est encore à un stade très précoce. Par contre, le potentiel de ces technologies est immense et ouvrirait une nouvelle avenue de développement pour les bibliothèques qui s'intéressent à la gestion des données. Comme le souligne Wusteman (2008), si les bibliothécaires ne font pas valoir leurs connaissances et leurs aptitudes (expertise en métadonnées et en organisation de l'information, lien de confiance avec les professeurs), il est probable que les chercheurs développent leurs propres outils et, de ce fait, court-circuitent les bibliothèques.

À la fin de cette étape du cycle de vie, le chercheur établit une version définitive de l'ensemble de données et il est important d'en garder une copie de sauvegarde en lieu sûr. Il ne s'agit pas encore ici d'un produit destiné au partage avec d'autres chercheurs, mais uniquement d'une version définitive et stable de l'ensemble de données à partir de laquelle l'analyse a été faite. En général, la préservation de cet ensemble de données dans un ERV ou dans le dépôt institutionnel de l'université est suffisante. Le bibliothécaire responsable des données et le personnel du dépôt institutionnel sont les interlocuteurs logiques pour faciliter le dépôt.

10. Pour un exemple concret d'application d'un ERV à la gestion des données à la bibliothèque de l'Université Cornell, voir la communication de Steinhart (2010).

Étape 4 : Publication et partage

Une fois l'analyse terminée, le chercheur publie généralement ses résultats sous la forme d'un article scientifique. Il s'agit donc du moment idéal pour publier en parallèle les données qui sous-tendent les résultats de la recherche. Cette « publication » vise à la fois l'archivage des données à plus ou moins long terme et le partage de celles-ci, soit avec le public en général (sans restriction quant à leur utilisation), soit dans un cadre contrôlé limitant l'accès à des utilisateurs autorisés qui s'engagent à respecter des règles d'utilisation et de confidentialité précises. Si le choix d'un dépositaire pour les données n'a pas déjà été effectué à l'étape 1, il devra se faire maintenant, ce qui n'est pas idéal, car on s'expose à un refus ou à des demandes de modifications des données ou des métadonnées qui peuvent être onéreuses. Une série de tâches est associée à la publication des données, certaines essentielles comme la dé-identification des participants et la publication des métadonnées, d'autres importantes bien que pas absolument nécessaires, telle la création d'identificateurs d'objet numérique. Soulignons qu'à cette étape du cycle de vie de la recherche ainsi que lors de la cinquième étape (gestion à long terme), les responsabilités du chercheur et du bibliothécaire de données sont moins importantes, car c'est le dépôt de données qui prend le relais. Cela est encore plus vrai si les données sont déposées dans un dépôt disciplinaire externe à l'université d'attache du chercheur et du bibliothécaire. Dans le cas de l'utilisation du dépôt institutionnel de l'université, il est probable que le bibliothécaire de données ait un rôle plus actif à jouer. Passons maintenant en revue les principales activités qui interviennent à cette étape.

Le processus de dépôt, d'archivage et de partage des données est généralement encadré par le modèle OAIS (Open Archival Information System), formalisé en tant que norme ISO (ISO 14721:2012)[11]. Une explication détaillée de ce modèle complexe dépasse largement le cadre de cet article. Par contre, il est utile d'en présenter quelques concepts clés ayant trait à ce qu'OAIS appelle son *modèle d'information*. Quant au modèle OAIS dans son ensemble, il suffira ici de dire qu'il s'agit d'un schéma conceptuel de haut niveau applicable non seulement à l'archivage des données, mais également à la gestion des objets numériques de façon générale. Un grand nombre de dépôts numériques à travers le monde l'emploient, notamment Scholars Portal[12], le dépôt numérique de l'Ontario Council of University Libraries (OCUL). Pour décrire les diverses tâches qui interviennent à cette étape du cycle de vie de la recherche, nous retiendrons les trois grands types de « Paquets

11. Le document décrivant le modèle OAIS, créé par le Consultative Committee for Space Data Systems, peut être consulté en ligne : <http://public.ccsds.org/publications/archive/650x0m2.pdf>.

12. <http://spotdocs.scholarsportal.info/display/sp/home>.

d'informations » présentés par le modèle d'information, soit le Paquet d'informations à verser (PIV), le Paquet d'informations archivé (PIA) et le Paquet d'informations diffusé (PID). Sans entrer dans les détails, disons simplement qu'un paquet d'informations contient à la fois l'objet numérique à préserver (l'ensemble de données dans le cas qui nous intéresse), les documents qui décrivent cet objet ainsi que les métadonnées spécifiant les modalités de préservation et autres informations techniques.

Le PIV constitue essentiellement l'ensemble des documents — ensemble de données, dictionnaire des données, documents méthodologiques, description du projet de recherche, entente signée stipulant les conditions de dépôt (propriété intellectuelle, paramètres de partage, durée de préservation, etc.) — soumis par le chercheur au dépôt de données. En théorie, l'ensemble de données sera prêt à être archivé et partagé, et les documents descriptifs seront complets et d'excellente qualité, ce qui permettra aux autres chercheurs de les réutiliser de façon autonome. En pratique, il est courant pour les responsables du dépôt d'avoir à communiquer avec le chercheur pour obtenir des précisions, corriger des erreurs, voire obtenir des documents manquants. La préparation du PIV est donc une tâche partagée par le chercheur (ou son équipe), qui assume la plus grande partie du travail, et les archivistes du dépôt de données, qui effectuent un contrôle de qualité et communiquent avec les chercheurs le cas échéant. Le bibliothécaire de données joue ici un rôle consultatif et peut conseiller le chercheur quant aux documents requis et aux principales métadonnées exigées par le dépôt.

La dé-identification des données représente une activité cruciale pour la préparation du PIV. Tous les ensembles de données portant sur des sujets humains comportent des risques d'identification des participants à l'étude. Ceux-ci exigeant généralement la confidentialité, il est indispensable de s'assurer que les données partagées ne puissent être utilisées directement ou indirectement pour identifier les individus ayant participé à la recherche. En plus d'éliminer les informations nominales (nom, adresse, téléphone, etc.), le chercheur (avec l'aide des archivistes du dépôt) procède, si nécessaire, à une opération de manipulation statistique qui consiste, par exemple, à regrouper les informations de types géographiques précis (code postal, aire de diffusion du recensement, etc.) en de plus grands ensembles comme la ville ou la région. On peut également éliminer des variables ou effacer les valeurs extrêmes de certaines variables qui pourraient éventuellement permettre d'identifier des individus. Le guide de l'ICPSR (2012, 37-38) à l'usage des chercheurs qui désirent y déposer leurs données décrit ces techniques en plus amples détails.

La préparation du Paquet d'informations archivé (PIA) relève strictement du domaine du dépôt de données. Il s'agit de transformer les documents constituant le PIV en des documents standardisés qui pourront être sauvegardés à long terme et seront aisément repérables par la communauté scientifique concernée. Cela implique généralement le respect de normes de métadonnées, par exemple le Dublin Core pour la description générale de l'étude et le Data Documentation Initiative (DDI) pour décrire les données en détails jusqu'au niveau des variables. Les fichiers de métadonnées seront idéalement dans un format ouvert et interopérables comme le XML. On s'assurera également d'archiver l'ensemble de données dans un format ouvert qui ne risque pas la désuétude (fichier ASCII, par exemple). Finalement, il est de plus en plus courant d'attribuer un identificateur d'objet numérique (DOI ou handle) à l'ensemble de données afin d'en faciliter le repérage, l'identification et la citation.

Le Paquet d'informations diffusé (PID) est lui aussi du ressort du dépôt de données. En termes simples, il s'agit de l'ensemble de documents (données et métadonnées) produit par le dépôt en réponse à une demande d'un utilisateur. Dans le scénario le plus simple, ce PID est simplement une copie du PIA correspondant. Par contre, il est courant pour les dépôts de données de fournir différents formats de fichiers correspondant à l'ensemble de données (ASCII, SAV, CSV, par exemple) pour répondre aux besoins précis des utilisateurs. Pour les OAIS plus sophistiqués, l'interface de commande permet aux utilisateurs de choisir, à partir d'un ensemble de données, un sous-ensemble de variables ou même un sous-ensemble de cas qui correspondent à des valeurs particulières de certaines variables. Par exemple, dans le cas d'une enquête sur la rémunération, un chercheur peut soumettre une requête pour obtenir un fichier qui comprend uniquement les individus dont le salaire brut est situé entre 25 000 $ et 50 000 $. Évidemment, ce genre de PID suppose un système informatique capable non seulement d'interpréter les requêtes complexes, mais aussi de générer de façon dynamique les PID qui répondent aux critères spécifiés par l'utilisateur.

Une autre tâche importante à l'étape de la diffusion des données consiste à publier simultanément les métadonnées qui décrivent les projets de recherche et les ensembles de données correspondants. La plupart des grandes archives de données emploient le protocole OAI-PMH pour assurer un maximum d'interopérabilité entre leur dépôt et les agrégateurs (Google Scholar, OAIster et autres « moissonneurs » d'archives ouvertes). Le protocole OAI-PMH peut être utilisé avec tous les schémas XML (Dublin Core ou DDI, par exemple).

Étape 5 : Gestion à long terme

À ce stade du cycle de vie, on met l'accent sur la préservation des données à long terme ainsi que sur l'aide au repérage des ensembles de données. Le chercheur n'a plus vraiment de rôle actif à jouer par

rapport à l'ensemble de données archivé, sinon peut-être de répondre à l'occasion à des questions sur ses données provenant d'autres chercheurs. Il peut se consacrer à l'élaboration d'un nouveau projet et à l'identification de données secondaires qui pourraient lui être utiles (retour à l'étape 1 du cycle de vie de la recherche). La majeure partie du travail de préservation et de mise en valeur des données repose désormais sur le dépôt. Par contre, le bibliothécaire de données et les bibliothécaires disciplinaires peuvent certainement contribuer au repérage (par exemple, en s'assurant que les métadonnées soient bien visibles pour les moteurs de recherche) et à la mise en valeur des données archivées, surtout s'ils travaillent conjointement avec le dépôt disciplinaire ou si le dépôt institutionnel de l'université est utilisé comme archive de données.

Au niveau du dépôt, les responsables doivent s'assurer de la mise en œuvre de la stratégie de préservation des données. Certains aspects de cette stratégie, notamment la création périodique de copies de sauvegarde et la vérification de l'intégrité des fichiers au moyen de l'exécution régulière de *checksum*, sont généralement automatisés. Par contre, certaines tâchent nécessitent généralement une intervention humaine. Ainsi, si le dépôt conserve des fichiers de type propriétaire, il faut s'assurer d'effectuer une migration de ceux-ci vers les versions plus récentes des logiciels correspondants pour éviter l'obsolescence des formats.

En plus de ces activités de nature technique, cette étape du cycle de vie est aussi le moment où, d'un point de vue bibliothéconomique, l'on se préoccupe de la citation des ensembles de données et de la création de liens entre ces ensembles et les articles qui les citent. Ce travail peut reposer sur les archivistes ou les bibliothécaires employés par le dépôt de données, mais on peut aussi imaginer des scénarios dans lesquels les bibliothécaires de données ou les bibliothécaires disciplinaires de l'institution où la recherche a été effectuée se chargent de ce travail avec ou sans la collaboration du dépôt. Pour ce qui est des citations, il s'agit de voir à ce que celles-ci respectent les modèles généralement acceptés tels que ceux proposés par DataCite ou Dataverse, ou à ce qu'elles se conforment à des modèles précis adoptés par les revues scientifiques comme l'*American Sociological Review*. Finalement, les bibliothécaires peuvent établir une liste de publications qui citent un ensemble de données, ce qui accroît la visibilité et la reconnaissance académique des chercheurs. Notons que des dépôts disciplinaires comme l'ICPSR (avec sa *Bibliography of Data-Related Literature*) et les grands producteurs de bases de données bibliographiques comme Thomson Reuters (avec le nouveau *Data Citation Index*) ont lancé de véritables bases de données faisant le lien entre publications et ensemble de données. Ici encore, les bibliothécaires peuvent jouer un rôle : l'ICPSR encourage en effet ses utilisateurs à lui soumettre toutes les références bibliographiques rencontrées afin que celles-ci soient ajoutées à la *Bibliography of Data-Related Literature*.

Conclusion

Nous espérons avoir démontré l'importance, autant du point de vue des chercheurs que de celui des bibliothèques universitaires, de s'attaquer dès maintenant à la question de la gestion des données de recherche. Pour les chercheurs, il s'agit d'embrasser pleinement la révolution de l'*eScience* et d'en tirer tout le potentiel de coopération entre collègues et de réutilisation des quantités massives de données créées chaque jour. 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.

Cela étant dit, nous n'en sommes qu'au début de cette transformation de la recherche et la science des données (*data science*) fait ses premiers pas. Le rôle exact que chacun des partenaires (chercheurs, bibliothécaires, comités d'éthique, bureau des TI, dépôts institutionnels et disciplinaires, entreprises du domaine des sciences de l'information) sera appelé à jouer reste à définir. Et les défis pour les bibliothèques universitaires sont indéniables. Le plus important est sans doute d'assurer la formation nécessaire aux bibliothécaires de données et autres archivistes spécialisés. Si l'on fait exception de quelques écoles pionnières aux États-Unis, ce type de formation ne fait pas encore partie du cursus des écoles de sciences de l'information. Qui plus est, 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. Par ailleurs, le développement de services de données nécessitera la mise sur pied d'une infrastructure technologique de préservation et de partage des données, et la création d'une culture de collaboration entre bibliothécaires et équipes de recherche. En fait, ces défis sont trop importants pour être relevés par des institutions individuelles, sauf peut-être les plus grandes et les plus riches. Il apparaît essentiel d'établir des partenariats au niveau provincial et même national. Il faudra également obtenir un appui financier et politique des différents paliers de gouvernement. Soulignons aussi qu'en l'absence d'exigences claires quant au partage des données de la part des organismes

subventionnaires et de moyens de vérifier leur application, le progrès ne saurait qu'être laborieux. Pour terminer sur une note positive, la multiplication au Canada des sommets, ateliers et rencontres portant sur la question de la gestion des données de recherche et la création de programmes de formation dans le domaine (surtout aux États-Unis) reflète un intérêt grandissant et une prise de conscience de l'importance de ces enjeux pour l'avenir des bibliothèques universitaires. ⊙

Sources consultées

Association des bibliothèques de recherche du Canada (ABRC). 2009. *Les données de recherche : un potentiel insoupçonné*. <http://www.carl-abrc.ca/uploads/pdfs/data_toolkit_low_res-f.pdf> (consulté le 25 avril 2013).

Conseil de recherches en sciences humaines (CRSH). 2012. *Politique sur l'archivage des données de recherche*. <http://www.sshrc-crsh.gc.ca/about-au_sujet/policies-politiques/statements-enonces/edata-donnees_electroniques-fra.aspx ?> (consulté le 25 avril 2013).

Data's Shameful Neglect. 2009. *Nature*, 461 (7261) : 145.

Green, Ann G. & Myron P. Gutmann. 2007. Building partnerships among social science researchers, institution-bssed repositories and domain specific data archives. *OCLC Systems & Services* 23 (1) : 35-53.

Higgins, Sarah. 2008. The DCC curation lifecycle model. *International Journal of Digital Curation* 3 (1) : 134-140.

Higgins, Sarah. 2012. The lifecycle of data management. In *Managing Research Data*, sous la direction de Graham Pryor. Londres : Facet Publishing, 17-46.

Interuniversity 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 Arbour : ICPSR.

Jahnke, Lori, Andrew Asher & Spencer D. C. Keralis. 2012. *The Problem of Data*. Washington, D.C. : Council on Library and Information Resources (CLIR).

Joint Information Systems Committee (JISC). 2006. *Virtual Research Environments Programme : Phase 2 Roadmap.* <http://www.jisc.ac.uk/publications/programmerelated/2006/pub_vreroadmap.aspx> (consulté le 25 avril 2013).

Luce, Richard. 2008. A new value equation challenge : the emergence of eResearch and roles for research libraries. In *No Brief Candle : Reconceiving Research Libraries for the 21st Century.* Washington, D.C. : Council on Library and Information Resources (CLIR), 42-50. <http://www.clir.org/pubs/reports/pub142/reports/pub142/pub142.pdf> (consulté le 29 avril 2013).

Microsoft. 2006. *Towards 2020 Science.* <http://research.microsoft.com/en-us/um/cambridge/projects/towards2020science/downloads/T2020S_ReportA4.pdf> (consulté le 25 avril 2013).

Milner, John. 2009. A UK research data service (UKRDS) : the way forward for research data management ? *Serials : The Journal for the Serials Community* (22) 1 : 83-85.

Perry, Carol Marie. 2008. Archiving of publicly funded research data : a survey of Canadian researchers. *Government Information Quarterly* (25) 1 : 133-148.

Piwowar, Heather A., Roger S. Day & Douglas B. Fridsma. 2007. Sharing detailed research data is associated with increased citation rate. *Plos One* (2) 3. <http://www.plosone.org/article/info :doi %2F10.1371 %2Fjournal.pone.0000308> (consulté le 25 avril 2013).

Scaramozzino, Jeanine Marie, Marisa L. Ramirez & Karen J. McGaughey. 2012. A study of faculty data curation behaviors and attitudes at a teaching-centered university. *College & Research Libraries* (73) 4 : 349-365.

Steinhart, Gail. 2010. *DataStaR : A Data Staging Repository to Support the Sharing and Publication of Research Data*. Communication présentée à la 31st Annual International Association of Scientific and Technological University Libraries Conference. <http://docs.lib.purdue.edu/iatul2010/conf/day2/8> (consulté le 25 avril 2013).

Waller, Martin & Robert Sharpe. 2006. *Mind the Gap : Assessing Digital Preservation Needs in the UK*. York, G.-B. : Digital Preservation Coalition. <http://www.dpconline.org/component/docman/doc_download/340-mind-the-gap-assessing-digital-preservation-needs-in-the-uk> (consulté le 25 avril 2013).

Wusteman, Judith. 2008. Virtual research environments : what is the librarian's role ? *Journal of Librarianship and Information Science* (40) 2 : 67-70.