arbido

2017/3 Metadaten – Datenqualität

Quiquerez Fanny, bibliothécaire-documentaliste

Minana Floriane, bibliothécaire

La gestion des données de la recherche: enjeux et perspectives

Dans la mouvance du libre accès (*Open Access*), la gestion des données est devenue pour l'ensemble des acteurs de la recherche un défi d'actualité. Des questions relatives à leur exploitation, leur échange et leur valorisation se posent. Les professionnels en information documentaire ont des compétences à faire valoir à ce sujet.

Qu'entend-on par «données de la recherche»?

Nombreuses sont les définitions tentant d'expliquer ce que l'on entend par données de la recherche. Leur définition est très complexe en raison des formes diverses sous lesquelles elles peuvent être enregistrées, la manière dont elles sont obtenues, ou encore, le traitement qu'elles subissent pour être lisibles (Gaillard 2014).

L'Organisation de coopération et de développement économiques (OCDE 2007), dont la définition est citée dans de nombreuses publications, décrit les données de la recherche comme étant des «enregistrements factuels (chiffres, textes, images et sons), qui sont utilisés comme sources principales pour la recherche scientifique et sont généralement reconnus par la communauté scientifique comme nécessaires pour valider des résultats de recherche».

Ces données sont produites par l'ensemble des instituts de recherche, indépendamment des domaines dans lesquels ceux-ci sont spécialisés et de l'institution à laquelle ils sont rattachés (université, entreprise privée, musées, ...). Leurs caractéristiques (les métadonnées de ces données) varient donc d'un domaine à l'autre et il est difficile d'en proposer un schéma standard.

Par ailleurs, il n'existe pas un mais plusieurs types de données qui dépendent notamment de leur niveau de structuration (d'hétérogènes à très bien structurées), de leur contexte de création (capturées ou créées) et des traitements qu'elles subissent. Il est ainsi possible de distinguer cinq grandes catégories de données:

- les données d'observation;
- les données expérimentales;
- les données computationnelles, de modèles ou de simulations;
- les données dérivées ou compilées;
- et pour finir, les données de référence ou canoniques (INIST 2014) 1.

L'enjeu majeur: l'Open Research Data

Si la gestion et la conservation des données de la recherche sont devenues une préoccupation majeure ces dernières années, c'est que les données, au même titre que les résultats de la recherche, font partie des contributions souhaitées au mouvement du libre accès (Gaillard 2014).

Ainsi, tout l'enjeu pour les chercheurs est de pouvoir les partager aussi librement que possible au reste de la communauté scientifique, afin de faciliter les progrès de la recherche. Le mouvement *Open Research Data* implique l'adoption de solutions techniques et juridiques mais il repose surtout sur de bonnes pratiques de gestion des données (Gaillard 2014).

Celles-ci se traduisent notamment par la présence de métadonnées qu'il est impératif de renseigner pour chaque jeu de données. Trois types de métadonnées peuvent servir à décrire les données de la recherche (Jacquemot-Perbal, Cosserat 2015):

- les métadonnées de provenance et de contexte: qui permettent de juger de la confiance et de la reproductibilité de la recherche;
- les mots-clés: qui définissent les sujets traités;
- les identifiants pérennes: qui permettent de garantir leur accessibilité et de relier les données aux publications. On peut notamment citer le DOI (pour «*Digital Object Identifier*»).

L'Institut national de la recherche agronomique (INRA 2017) propose, à travers les principes FAIR2 (pour «*Findable*», «*Accessible*», «*Interoperable*», et «*Reusable*»), un certain nombre de recommandations visant à rendre les métadonnées, au même titre que les données elles-mêmes, trouvables, accessibles, interopérables et réutilisables.

Le choix d’un entrepôt de données

De nombreuses infrastructures permettant l’archivage à long terme des données se sont développées. Celles-ci sont très diverses, et lorsque les chercheurs ne disposent pas d’un entrepôt institutionnel, un outil tel que OpenDOAR leur permet de faire le tour des archives ouvertes existantes, selon le champ disciplinaire auquel appartient la recherche, le pays, etc.

En Suisse, le projet DLCM (Data Lifecycle Management) doit pouvoir proposer d’ici 2018 une solution pour la préservation à long terme des données (Krause, Blumer 2016).

Tous ces entrepôts doivent satisfaire à certains critères afin de garantir que les données soient exploitables (c’est-à-dire que leur intégrité et authenticité soient préservées et qu’elles soient retrouvables et réutilisables). Pour répondre à ces exigences et permettre l’interopérabilité des systèmes, il est donc nécessaire de définir des jeux de métadonnées standardisés. Quand aucun entrepôt n’existe pour la discipline et le pays choisis, il est encore possible d’opter pour des solutions développées de manière entièrement automatisée. On peut notamment citer Zenodo, créé par le CERN dans le cadre du projet européen OpenAIREplus. Cependant, il faut savoir que les machines ne sont pas encore capables, du fait de l’automatisation du traitement des données, de fournir les mêmes prestations, avec les mêmes garanties, que les infrastructures gérées par les professionnels de l’information. Par conséquent, ces solutions ne peuvent pas être considérées comme des «entrepôts de confiance» (Pôle Digitalist 2017).

Le *Data Management Plan*, ou Plan de gestion des données

À partir du mois d’octobre de cette année, les chercheurs souhaitant soumettre leurs projets au Fonds national suisse de la recherche scientifique (FNS) devront établir un *Data Management Plan* (communément appelé DMP, ou encore Plan de gestion des données en français), dans le respect des principes FAIR cités plus haut (FNS 2017). Ce document permet de planifier la gestion des données de la recherche avant, pendant, et après un projet. Il implique notamment une réflexion sur la manière dont seront décrites les données produites ou collectées au cours de la recherche. Cette étape est fondamentale puisque la question des métadonnées constitue l’une des principales conditions au partage et à la compréhension des données (Gaillard 2014). Ainsi, chaque chercheur devra définir les standards et formats des métadonnées, comment et par qui elles seront renseignées, les règles de nommage des jeux de données et leur arborescence.

Le DMP, qui doit s'adapter aux exigences des différents types de recherches, peut prendre des formes très variées, comme en témoigne le Digital Curation Centre (2017). Plusieurs modèles existent pour sa réalisation: des modèles sont proposés par de nombreuses universités, et des outils en ligne tels que DMP Online ou Data One permettent également de réaliser ce document.

Le rôle des spécialistes en information documentaire

Dans la mouvance du libre accès, afin de définir les pratiques de gestion et de conservation des données de leurs chercheurs et en favoriser leur partage, de plus en plus d'universités s'engagent à développer des politiques de gestion des données de la recherche. Les spécialistes en information documentaire contribuent pleinement à ce mouvement, non seulement en participant à la définition de ces politiques, mais aussi en formant les chercheurs à la planification et à la gestion de leurs données, ou encore en prenant part au développement des entrepôts et en gérant les métadonnées (Gaillard 2014).

Afin d'assurer la curation des données, de nouveaux métiers tels que les data archivists ou data curators ont émergé, nouveaux visages des professionnels en archivistique. En effet, ceux-ci savent depuis longtemps préserver des données (les documents contenant ces données) et les retrouver par leurs métadonnées (les inventaires d'archives). Les chercheurs peuvent ainsi s'appuyer sur leurs compétences pour assurer l'exploitabilité de leurs données.

Bibliographie

DIGITAL CURATION CENTRE (DCC), 2017. Example DMPs and guidance. DCC [en ligne]. [Consulté le 23 juillet 2017]. Disponible à l'adresse: http://www.dcc.ac.uk/resources...

FONDS NATIONAL SUISSE DE LA RECHERCHE SCIENTIFIQUE (FNS), 2017. Open Research Data?: les requêtes devront inclure un plan de gestion des données. FNSNF: Fonds national suisse de la recherche scientifique [en ligne]. [Consulté le 23 juillet 2017]. Disponible à l'adresse: http://www.snf.ch/fr/pointrech...

GAILLARD, Rémi, 2014. De l'Open data à l'Open research data: quelle(s) politique(s) pour les données de recherche? [en ligne]. [Villeurbanne]: Ecole nationale supérieure des sciences de l'information et des bibliothèques. Mémoire d'étude. [Consulté le 25 juillet 2017]. Disponible à l'adresse: http://www.enssib.fr/bibliothe...

INSTITUT DE L'INFORMATION SCIENTIFIQUE ET TECHNIQUE (INIST), 2014. Une introduction à la gestion et au partage des données de la recherche: quelques définitions. Institut de l'information scientifique et technique [en ligne]. [Consulté le 21 juillet 2017]. Disponible à l'adresse: http://www.inist.fr/donnees/co...

INSTITUT NATIONAL DE LA RECHERCHE AGRONOMIQUE (INRA), 2017. Principes FAIR. INRA: Science & Impact [en ligne]. [Consulté le 23 juillet 2017]. Disponible à l'adresse: https://www6.inra.fr/datapartage/Technologies/Principes-FAIR

JACQUEMOT-PERBAL, Marie-Christine et COSSERAT, Françoise, 2015. Gestion et diffusion des données de la recherche. Institut de l'information scientifique et technique [en ligne]. [Consulté le 9 août 2017]. Disponible à l'adresse: http://www.inist.fr/IMG/pdf/ur...

KRAUSE, Jan et BLUMER, Eliane, 2016. Projet DLCM: les données de la recherche en Suisse. Authorea [en ligne]. [Consulté le 23 juillet 2017]. Disponible à l'adresse: https://www.authorea.com/users...

ORGANISATION DE COOPÉRATION ET DE DÉVELOPPEMENT ÉCONOMIQUES (OCDE), 2007. Principes et lignes directrices pour l'accès aux données de la recherche financée sur fonds publics [document PDF]. OECD [en ligne]. Avril 2007. [Consulté le 21 juillet 2017]. Disponible à l'adresse: http://www.oecd.org/fr/sti/sci...

PÔLE DIGITALIST (DLH), 2017. Datapartage - Déposer dans Zenodo-Inra. Inra.fr [en ligne]. [Consulté le 31 juillet 2017]. Disponible à l'adresse: https://www6.inra.fr/datapartage/Partager-Publier/Valoriser-ses-donnees/Deposer-dans-Zenodo-Inra

1 Pour aller plus loin sur le sujet, l'Institut de l'information scientifique et technique (INIST) a mis en ligne plusieurs tutoriels à propos du libre accès aux résultats de la recherche, dont un permettant de se familiariser avec les données de la recherche.

2 Ces derniers sont très largement développés au sein d'un tableau, précisant chacun des concepts et comment les mettre en œuvre.

Fanny Quiquerez

Bibliothécaire au centre de documentation de la Haute école de santé de Genève depuis 2013, Fanny Quiquerez est au bénéfice d'un Bachelor of Science HES-SO en Information documentaire obtenu en 2012 auprès de la Haute école de gestion de Genève. Intéressée par les nouveaux enjeux liés au domaine, elle complète actuellement sa formation par un Master en Sciences de l'information qu'elle a débuté en septembre 2016 auprès de la même institution.

Floriane Minana

Formée aux métiers du livre en France, Floriane Minana a travaillé en bibliothèques municipales quelques années avant de s'installer en Suisse en 2014. Actuellement en poste en bibliothèque scolaire dans le canton de Vaud, elle suit le Master en Sciences de l'information proposé par la Haute école de gestion de Genève afin de compléter sa formation.

Abstract

Français
La gestion des données de la recherche représente l'une des préoccupations actuelles majeures dans le milieu de la recherche. Elle apparaît dans la continuité du mouvement du libre accès (*Open Access*) et témoigne d'une volonté de partager librement les données. Leur exploitation, leur échange et leur valorisation nécessitent notamment la création de métadonnées de qualité. Cet article propose de faire un tour d'horizon de la problématique et du rôle que peuvent jouer les spécialistes en information documentaire.

Deutsch
Die Verwaltung von Forschungsdaten über die Zeit stellt eine der grössten Herausforderungen für die Forschung dar. Dieses Anliegen steht in der Fortsetzung der Bewegung für freien Zugang (Open Access) und zeugt von einer Bereitschaft, Forschungsdaten frei zur Verfügung zu stellen und mit anderen zu teilen. Die Nutzung, der Austausch und die Aufwertung von Forschungsdaten bedingen insbesondere die Erstellung von qualitativ guten Metadaten. Im vorliegenden Artikel wird ein Überblick über die Fragestellungen in diesem Bereich geboten und die möglichen Rollen von I+D-Spezialistinnen und -Spezialisten in diesem Kontext werden ausgelotet.