Fiche Bibliothéconomie 1-8

Les données de la recherche

NB : Cette fiche est à lire en regard des fiches 1-6 « Production et diffusion de l'information scientifique et technique » et 1-7 « La science ouverte »

A voir aussi : fiches Biblio 1-5 « La documentation numérique en bibliothèque », 2-6 « Les métadonnées » et 3-12 « Les services d'appui à la recherche ».

L'information scientifique et technique est au cœur de l'activité de recherche.

En 2012, le CNRS organise une journée intitulée *Données de la recherche : enjeux, perspectives, politique(s)*[1]. En 2013, le lancement du chantier Bibliothèque scientifique numérique (BSN) consacré aux données de la recherche est lancé. Les journées d'étude sur le thème se développent.

Aujourd'hui, le comité de pilotage de la science ouverte (CoSo) remplace la BSN. Il propose des orientations et instruit ses sujets sur les questions de la science ouverte. Il impulse et accompagne les actions dans une structure fluide, facilitant l'expression, la remontée des idées, les engagements et les contributions aux différents groupes de travail. Le COSO comprend notamment un collège (groupe de travail) « données de la recherche ».

1. Que sont les données de la recherche ?

La définition la plus souvent reprise pour les données de la recherche est celle proposée en 2007 par l'OCDE. Cette définition caractérise une donnée de recherche par sa finalité et non uniquement par sa typologie : "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." « OECD Principles and Guidelines for Access to Research Data from Public Funding », 2007.

Le périmètre des données de la recherche est circonscrit par ce texte à la notion d'éléments probants, nécessaires à la validation du processus de recherche[2].

On peut donc dire que l'on appelle données de la recherche (DR) l'ensemble des informations collectées, observées ou créées sous une forme numérique ou non, par les

1 https://fredoc.hypotheses.org/124 (Consulté le 25/07/2024)

2 Définition donnée par le site Couperin de la Science ouverte en France : https://scienceouverte.couperin.org/donnees-recherche-definitions/ (Consulté le 25/07/2024)

chercheurs dans le cadre d'un projet de recherche et à partir desquelles ils bâtissent leurs hypothèses.

Elles sont un produit de la recherche, un élément de communication scientifique et regroupent un ensemble hétéroclite de sources et matériaux de recherche[3].

2. Quels types de données de la recherche ?

- ✓ Les données de la recherche sont (désormais) numériques, descriptives et visuelles.
- ✓ Ces données peuvent être : des « enregistrements factuels (chiffres, textes, images, sons) utilisés comme source principale pour la recherche scientifique et généralement reconnus par la communauté scientifique comme nécessaires pour valider les résultats de la recherche. Un ensemble de données de recherche constitue une représentation systématique et partielle du sujet faisant l'objet de la recherche » (OCDE, 2007)[4].
- ✓ Elles ont été produites au cours d'un processus scientifique pour servir une démonstration. Elles concernent également les publications.
- ✓ Ces données peuvent être brutes (ou primaires = données d'observation, d'expérimentation, de simulation), traitées (ou dérivées = traitement, combinaison ou réorganisation des données brutes), analysées (ou interprétées).

3. La vie des données

Il existe de nombreuses représentations de la vie des données.

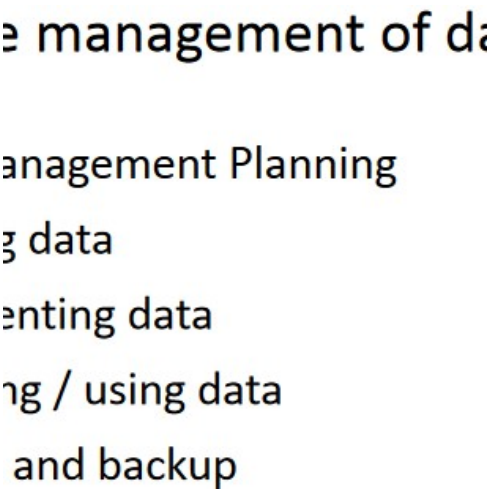
On peut reprendre par exemple, la représentation de Sarah Jones présentée lors de la conférence de Madrid du 25 février 2015 sur les plans de gestion de données et H2020[5] :

3 Julie Duprat. *Les données de la recherche à l'Université Bordeaux Montaigne : Synthèse d'une enquête qualitative auprès des chercheurs*. [Rapport de recherche] Université Bordeaux Montaigne. 2019. hal-02020141(Consulté le 25/07/2024).

4 Organisation de coopération et de développement économiques (OCDE), Principes et lignes directrices de l'OCDE pour l'accès aux données de la recherche financée sur fonds publics, Paris, 2007, disponible sur : https://www.oecd-ilibrary.org/science-and-technology/oecd-principles-and-guidelines-for-access-to-research-data-from-public-funding_9789264034020-en-fr?mlang=fr (Consulté le 25/07/2024).

5 Jones S (2015). Managing research data and Horizon 2020. In: ConsorcioMadroño conference on Data Management Plans and Horizon 2020, ETSI Industriales, 25th February, Madrid, Spain. http://www.consorciomadrono.es/noticias_eventos/2015/JornadaPGD/sarah.pdf Consulté le 25/07/2024).

- ✓ Créer ou collecter les données (Create). Les données peuvent également avoir été produites par d'autres et réutilisées
- ✓ Documenter les données (Document), normalement tout au long du travail effectué sur les données.
- ✓ Exploiter les données et les interpréter pour produire des résultats (Use)
- ✓ Stocker et sauvegarder les données (Store)
- ✓ Partager les données selon le principe « aussi ouvert que possible, aussi fermé que nécessaire » (Share)
- ✓ Conserver les données (Preserve)



Les bonnes pratiques de gestion des données s'appliquent à chaque étape du cycle de vie des données.

4. Quelles données partager ?

A priori toutes les données, qu'elles soient brutes, dérivées ou analysées, peuvent être partagées. Les données de la recherche issues de financements publics sont d'ailleurs des données publiques au sens de la loi[6]. Il existe tout de même des limites qui concernent les données personnelles, les données sensibles et les données portant sur des intérêts commerciaux. Les communautés de chercheurs, chacun dans leur domaine, doivent également contribuer à définir quelles données / jeux de données doivent être archivés, conservés, partagés et comment.

La possibilité de manipuler, d'agréger ces données, d'entreprendre de nouvelles recherches à partir de grandes masses de données (big data) requiert la mise en place d'outils de gestion, de conservation et de partage[7].

5. La gestion des données : modèles et outils

5.1. Les plans de gestion de données

6 Article L211 du code du patrimoine : https://www.legifrance.gouv.fr/codes/id/LEGISCTA000006129161/(Consulté le 25/07/2024)
)

7 Séminaire *Outils et services pour la gestion et l'ouverture des données scientifiques : retours d'expérience*, INRA (Paris), 16 février 2017, disponible sur https://seminaire.inra.fr/data (Consulté le 25/07/2024)
.

Les plans de gestion de données (ou data management plan, DMP) sont des documents évolutifs dont le but est de décrire les différentes étapes du travail à réaliser sur les données au cours d'un projet de recherche. Ces documents sont demandés par les financeurs dans le cadre de réponses aux appels à projets, ou par les institutions. Les DMP doivent garantir que les problématiques liées à la gestion des données au cours d'un projet ont bien été anticipées. Ce document est à mettre en lien avec les politiques de science ouverte. En effet, un des objectifs d'une bonne gestion de données est de faciliter leur réutilisation par d'autres équipes de recherche, et si possible leur ouverture, en fin de projet.

Le plan de gestion de données est un document regroupant un ensemble de questions sur les données. Voici quelques exemples :

Qui sont les personnes responsables ? Quelle politique sera appliquée (politique de financement, politique institutionnelle…) ? Quels types de données seront collectées / générées ? Comment seront organisés les fichiers / les dossiers ? Quelle description des données (métadonnées, documentation) ? Comment ces données seront-elles stockées, sauvegardées, sécurisées ? Comment seront-elles partagées (propriété intellectuelle, licence de réutilisation) ? Quelles modalités de conservation ? Coût et ressources nécessaires pour la gestion et le partage ? …

Les plans de gestions doivent prévoir et répondre à ces questions.

Quelques exemples de modèles :

- ✓ Cirad Découvrir des plans de gestion des données de la recherche[8].
- ✓ Inist-CNRS : DMP Opidor et la plateforme d'autoformation d'apprentissage numérique DoRANUM[9] qui propose des ressources d'auto-formation à la gestion des données de la recherche ainsi que des conseils pour la rédaction des DMP.

5.2 Le principes FAIR

Les principes FAIR[10] (faciles à trouver, accessibles, interopérables et réutilisables) sont à la base des bonnes pratiques de gestion et d'ouverture des données de la recherche.

1) Facile à Trouver : les données sont mises en ligne dans un espace bien identifié des communautés signalées par un identifiant unique et pérenne (DOI), décrites par des métadonnées riches et indexées, qui facilitent leur repérage.

8 Plus d'infos : https://coop-ist.cirad.fr/actualites/decouvrir-des-plans-de-gestion-de-donnees-de-la-recherche Consulté le 25/07/2024)

9 Aller plus loin : https://dmp.opidor.fr/ (voir en particulier le modèle de DMP de l'ANR) et https://doranum.fr/plan-gestion-donnees-dmp/ Consulté le 25/07/2024)

10 https://www.ccsd.cnrs.fr/principes-fair/ Consulté le 25/07/2024)

2) Accessibles : la conservation des données est garantie sur le long terme. Les données sont accessibles gratuitement selon le principe, « aussi ouvert que possible, aussi fermé que nécessaire ».

3) Interopérables : les données sont disponibles dans des formats et standards ouverts et partagés, permettant une lecture par les machines.

4) Réutilisables : les métadonnées décrivent le contexte de production du jeu de données et les conditions de réutilisation, notamment via le moyen d'une licence creative commons (CC-by).

5.3 Les entrepôts de données

La diffusion des données de la recherche passe largement par leur dépôt dans un entrepôt de données, base de données visant à rendre accessibles un ensemble de jeux de données de la recherche. Comme pour les archives ouvertes de publications scientifiques, les entrepôts de données peuvent être disciplinaires (exemple : data terra), institutionnels (exemple : data.sciencespo) ou nationaux ou internationaux (exemple : Zenodo, qui regroupe archive ouverte et entrepôt de données.)

La France a lancé en juillet 2022 un entrepôt national pour les données de la recherche : *recherche data gouv*. Dans le cadre du deuxième plan national pour la science ouverte et de cette nouvelle infrastructure, les établissements sont appelés à accompagner l'usage de cet entrepôt et à développer des services d'accompagnement à la gestion des données.

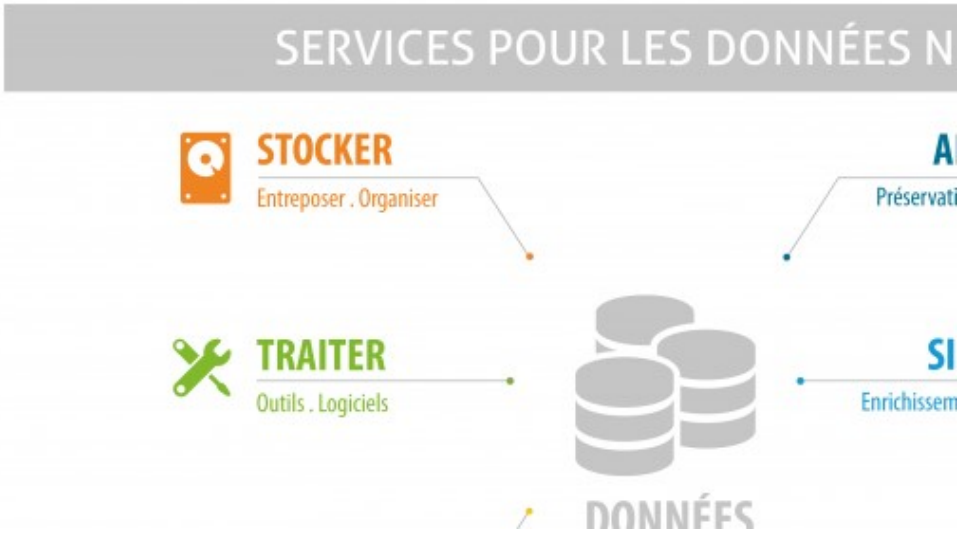
5.3 Le rôle des SCD dans l'accompagnement à la gestion et à l'ouverture des données de la recherche

Aujourd'hui, les SCD jouent très souvent un rôle dans l'accompagnement des chercheurs à la gestion des données de la recherche[11]. Cet accompagnement peut se traduire par des formations sur les bonnes pratiques de gestion et d'ouverture des données, une aide à la rédaction de plans de gestion de données, des conseils tout au long du projet de recherche, le développement d'outils institutionnels ou encore l'aide au dépôt de jeux de données en entrepôt. Le terme de *data librarian* s'est ainsi développé depuis quelques années.

11 Voir notamment les nombreux retours d'expérience menés par le GTSO données : https://gtso.couperin.org/gtdonnees/6-retours-dexperience/ (Consulté le 25/07/2024)

5.4 Exemple de vie des données dans la TGIR HumaNum (infrastructure de recherche soutenue par le CNRS)

À chaque étape du cycle de vie des données correspond un service dédié : : le traitement (boîte à outils partagée), la conservation (Huma-Num-Box), l'accès et l'interopérabilité des données de la recherche en sciences humaines et sociales par des services numériques de traitement, de partage (Nakala) et d'accès unifié (ISIDORE), ainsi qu'une procédure d'archivage à long terme.



Pour aller plus loin :

Site web du CIRAD, Gérer les données de la recherche : https://coop-ist.cirad.fr/gerer-des-donnees (Consulté le 25/07/2024)

Dossier *Politique nationale de l'IST : des infrastructures en cohérence*, Ar(abes)ques, n° 84, février-mars 2017. Disponible sur http://www.abes.fr/Publications-Evenements/Arabesques/Arabesques-n-84 (Consulté le 25/07/2024)

OCDE (Organisme de coopération et de développement économique), *Principes et lignes directrices de l'OCDE pour l'accès aux données de la recherche financée sur fonds publics*, 2007. Disponible sur : https://www.oecd-ilibrary.org/science-and-technology/oecd-principles-and-guidelines-for-access-to-research-data-from-public-funding_9789264034020-en-fr?mlang=fr (Consulté le 25/07/2024

Rémi Gaillard, *De l'Open data à l'Open research data : quelle(s) politique(s) pour les données de recherche ?*, mémoire d'étude pour l'accès au diplôme de conservateur, disponible sur : https://www.enssib.fr/bibliotheque-numerique/notices/64131-de-l-open-data-a-l-open-research-data-quelles-politiques-pour-les-donnees-de-recherche (Consulté le 25/07/2024

Le site dédié au Plan national pour la science ouverte : https://www.ouvrirlascience.fr/ (Consulté le 25/07/2024