Programme Prioritaire de Recherche
AUTONOMIE

DIFFUSER LES DONNÉES DE LA RECHERCHE

Cette fiche, issue des travaux préliminaires et des consultations menées par l'équipe projet du Programme prioritaire de recherche (PPR) Autonomie, piloté par le CNRS, a pour objectif de présenter un certain nombre de ressources pour faciliter le travail des équipes dans l'ouverture et la diffusion de leurs données de recherche. Elle fait partie d'une série de trois fiches pratiques, dont l'une s'intéresse à la gestion des données et l'autre à la réutilisation des données.

Le partage des données

Dans le contexte de la science ouverte, il est désormais obligatoire pour les chercheurs de partager leurs données, c'est-à-dire de les mettre à disposition de la communauté scientifique en les signalant publiquement, par le dépôt des jeux de données sur des plateformes numériques spécialisées : les entrepôts de données.

Les données de la recherche sont définies comme « 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 les résultats de la recherche. » (Rapport de l'OCDE, 2007). Ceci comprend les bases de données quantitatives, mais aussi les retranscriptions d'entretiens, d'observations, journaux de terrains, corpus de textes, enregistrements audios et vidéos, etc.

Le principe de l'ouverture des données repose sur la maxime « aussi ouvertes que possibles, aussi fermées que nécessaires ». Ainsi, la diffusion des données de la recherche ne signifie pas l'ouverture à tous les publics. Si l'ouverture sans restriction peut être la solution retenue, elle n'en est qu'une parmi d'autres.

Outre les impossibilités juridiques du partage de certaines données, de nombreuses raisons peuvent justifier le choix d'en restreindre l'accès : enjeux de confidentialité pour les personnes interrogées, nécessité de produire d'abord des résultats d'analyses primaires, etc.

BON À SAVOIR

Un jeu de données est un ensemble cohérent de données.
Les métadonnées sont utiles pour exploiter un jeu de données produit par d'autres. Ce sont les nombreuses informations relatives au contexte de production, à la méthodologie, à la description du jeu, etc.
Les entrepôts de données sont des espaces qui rendent accessibles les jeux de données et les métadonnées qui y sont associées.
Les entrepôts en auto-dépôt sont des espaces de partage libre de données et sans vérification de la qualité des métadonnées.

Pourquoi diffuser les données de la recherche ?

La recherche est un travail collectif et cumulatif, il est souhaitable de mettre à disposition du reste de la communauté les données mobilisées dans ses travaux. Cela permet à d'autres équipes d'exploiter des jeux de données qu'elles n'auraient pas pu constituer en raison du coût de la collecte ou de la méthode utilisée.

Par ailleurs, partager les données de la recherche peut permettre d'éviter une sur-sollicitation des populations enquêtées et offre aussi un stockage sécurisé des données et un contrôle de leur diffusion aux équipes de recherche habilitées.

La diffusion des données peut aussi relever d'arguments épistémologiques : d'une part, les données partagées peuvent permettre d'appuyer un raisonnement en

mettant à disposition de la communauté les éléments empiriques qui le constituent, d'autre part cela permet d'identifier plus facilement les limites des données et permet aux producteurs de les prendre en compte et d'améliorer un futur recueil.

La loi pour une République Numérique de 2016 prévoit que si rien ne s'y oppose, les données produites grâce à des fonds publics (à hauteur d'au moins 50%) soient mises à disposition, afin de sortir d'une logique de propriété privée et de constituer ces données en bien commun pour la communauté scientifique.

La confidentialité des données récoltées (données à caractère personnel par exemple), l'impossibilité d'anonymisation, l'existence d'une propriété privée s'exerçant sur les données (parce qu'elles sont collectées en partenariat avec une structure privée) ou les enjeux liés au secret-défense ou à l'industrie sont des motifs pouvant justifier la non-publication des données de la recherche.

Enfin, la diffusion de données est l'occasion de visibiliser la production scientifique d'un collectif de recherche alors identifié comme un potentiel partenaire dans le cadre d'un projet de réutilisation de ces données ou d'un projet sur des sujets proches. La diffusion des données peut s'accompagner de la publication de *data papers*[1], une opportunité de valorisation des jeux de données.

Déposer les données de la recherche dans un entrepôt

Il est recommandé de déposer les jeux de données dans des infrastructures dédiées que sont les entrepôts de données. Cela permet de s'appuyer sur une structure compétente pour les questions de stockage et de sauvegarde des jeux de données, de cadrer juridiquement la réutilisation des données (contrats de confidentialité et licences d'utilisation) et d'offrir une plus grande visibilité aux données partagées (attribution de DOI[2], moissonnage[3], etc.).

En cas de dépôt des jeux de données dans un entrepôt, l'équipe productrice conserve ses droits moraux d'auteur, et la licence d'utilisation du jeu déposé impose la citation lors d'analyses secondaires. Des entrepôts proposent également de restreindre l'accès aux jeux de données à certains types de publics ou de le soumettre à l'accord préalable des producteurs. Il est aussi possible de déposer les jeux de données dans un entrepôt, mais de les laisser temporairement sous embargo (par exemple le temps de la valorisation scientifique).

Déposer les jeux de données nécessite de garder à l'esprit les principes FAIR. Par les différents outils et l'accompagnement qu'ils proposent, certains entrepôts ont vocation à aider à respecter ces principes.

BON À SAVOIR

Les principes FAIR invitent à ce que les données soient :
- Faciles à trouver, c'est-à-dire, signalées dans les entrepôts de données pertinents ;
- Accessibles à d'autres équipes, dans les conditions déterminées par le producteur ;
- Interopérables, à savoir, facilement interprétables par des machines ;
- Réutilisables, donc décrites et documentées pour faciliter les analyses secondaires.

Quelques points de vigilance

D'abord, se renseigner sur la propriété intellectuelle des données. Si celles-ci ont été collectées dans le cadre d'une collaboration entre plusieurs équipes de recherche ou entre une équipe et une institution tierce, ces parties prenantes ont un droit sur ces données et peuvent s'opposer à leur diffusion, sauf si des dispositions contraires sont prévues par la convention de partenariat (signée en amont de la constitution des données).

S'agissant de la gestion des données personnelles – si de telles données ont été recueillies – seul l'accord préalable des personnes concernées peut permettre de les diffuser en l'état. Cet accord peut être obtenu en même temps que le consentement à participer à la recherche ou à postériori avant la diffusion des données. Sans lui, il est alors nécessaire de procéder à une anonymisation, c'est-à-dire retirer de tous les jeux de données les informations permettant une identification directe ou indirecte des individus.

La description du jeu de données, de son contexte de collecte et de diffusion est indispensable à la bonne réutilisation des données. La plupart des entrepôts proposent des standards de métadonnées afin de guider les producteurs sur les informations à renseigner

1. Publication consacrée à la valorisation de données et mise à disposition de la communauté de recherche.
2. *Digital Object Identifier,* identifiant pérenne et unique, permet de référencer, citer et fournir un lien stable vers un fichier en ligne.
3. Technique informatique qui permet le transfert et l'affichage d'information automatiquement.

pour caractériser les jeux de données. Outre ces métadonnées, il est conseillé d'accompagner la mise à disposition des jeux de données d'une documentation complémentaire (fichier *lisez-moi*, sommaire, contexte de collecte, méthodologie, traitement, limite, etc.) et d'une indexation à l'aide de mots-clés issus de référentiels.

Enfin, lorsqu'ils sont nombreux, organiser les différents fichiers selon des normes prédéfinies facilite la navigation dans la documentation. Cela peut passer par l'adoption de conventions dans la structuration des fichiers, de retranscriptions d'entretiens, d'observation ou dans la dénomination des fichiers et des dossiers. L'organisation des fichiers peut aussi être un moyen d'améliorer l'intelligibilité en distinguant les données selon leur type, leur temporalité, leur origine géographique, etc.

Finalement, il est conseillé de prévoir la diffusion des données dès le début du projet, notamment dans le cadre du PGD[4], afin d'anticiper au mieux tous ces aspects.

Ressources

Les entrepôts où diffuser les données

Quetelet-Progedo-Diffusion est un service de diffusion de données qui héberge des jeux de données quantitatives mobilisables en SHS, par exemple les enquêtes « Conditions de vie des étudiants ».

Le CASD[5] est la solution à privilégier pour diffuser des données quantitatives particulièrement sensibles. Il héberge essentiellement des jeux de données administratives ou de la statistique publique, mais les chercheurs peuvent demander à y héberger des données sensibles.

Le CDSP[6] héberge des jeux de données quantitatives et peut en assurer la documentation sous certaines conditions. À titre d'exemples, ont été prises en charge les enquêtes sur la santé et les consommations lors de l'appel à la préparation à la défense (ESCAPAD) ou une enquête sur les contrôles d'identité à Paris.

La plateforme CoCoON[7] est un entrepôt hébergeant des corpus d'enregistrements oraux avec leurs annotations. On y trouve par exemple des atlas linguistiques (d'Alsace ou d'Haïti), des corpus d'enregistrements produits lors d'enquêtes et des archives de la parole.

Certaines universités et certains EPST[8] ont eux-mêmes développé des entrepôts pour leurs équipes et les laboratoires associés. On peut ainsi citer DOREL pour l'université de Lorraine ou Didómena pour l'EHESS[9].

Enfin, un recensement interdisciplinaire des entrepôts français est réalisé par Cat.OPIDoR et un recensement international par Re3Data.

Les entrepôts en auto-dépôt

Recherche Data Gouv a vocation à servir d'entrepôts aux équipes qui ne trouvent pas de solutions appropriées. Elle vise à centraliser et à rendre visible les métadonnées d'un grand nombre d'entrepôts institutionnels par moissonnage.

Nakala est une solution française généraliste qui permet un hébergement et un accès aux fichiers de tout type en SHS, dont des jeux de données d'enquêtes ou historiques.

Zenodo est un entrepôt européen à vocation internationale généraliste et interdisciplinaire. Parmi les jeux de données stockés, se trouvent plus de cent-cinquante-mille jeux de données, dont une grande partie associée à des publications.

Les services d'accompagnement

Certains entrepôts prennent en charge tout ou partie de la documentation des données, d'autres accompagnent les chercheurs dans le travail de documentation ou valident la diffusion et sa conformité à la législation. Certains pratiquent également l'auto-dépôt, c'est-à-dire que les producteurs sont libres de partager les jeux de données, sans que l'entrepôt ait un droit de regard. Il est alors possible de solliciter des services dédiés à l'accompagnement pour le partage de données.

Les services documentaires d'universités et autres structures de recherche sont compétents afin d'accompagner et d'orienter les chercheurs dans leurs démarches de dépôts. Les DPD[10] de chaque structure sont des interlocuteurs autour des enjeux de données personnelles. Les PUD[11] encadrées par la Progedo sont aussi compétentes sur les démarches de mises à disposition.

4. Plan de gestion de données.
5. Centre d'accès sécurisé aux données.
6. Centre de données socio-politiques.
7. Collections de corpus oraux numériques.
8. Établissements publics à caractère scientifique et technologique.
9. École des hautes études en sciences sociales.
10. Délégués à la protection des données (ou DPO : *Data protection officer*).
11. Plateformes universitaires de données.

Recherche Data Gouv met en place des ateliers de la donnée labéllisés (ou en cours de labéllisation) et fédèrent les compétences à l'échelle d'un territoire pour accompagner les équipes de recherche dans la gestion, la structuration et la mise à disposition de leurs données. On peut citer la Cellule data Grenoble Alpes, pilotée par l'université Grenoble-Alpes et qui mobilise 22 personnes pour accompagner près de 90 unités de recherche (consulter la liste de l'ensemble des ateliers de la donnée).

Se former à la mise à disposition

Dans la perspective de préparer les données en vue d'une diffusion, de nombreuses ressources sont disponibles pour accompagner la montée en compétences des chercheurs.

Un arbre de décision est proposé par les équipes du consortium Couperin et du CIRAD, pour orienter les équipes dans les problématiques juridiques liées à l'ouverture des données.

DORANum conçoit des ressources variées pour former les chercheurs aux défis de la mise à disposition. Il est labélisé centre de ressources par Recherche Data Gouv.

Le réseau des URFIST[12] propose des formations diverses, notamment en ce qui concerne les données de recherche. Ce réseau est également labéllisé Centre de ressources.

Concernant la gestion des données personnelles, de nombreux guides sont disponibles : La CNIL[13] propose des définitions claires des concepts importants et le Vadémécum pour la réutilisabilité des données, issu de l'activité du groupe de travail « (Ré)utilisatibilité » au sein du consortium Cahier, propose des bonnes pratiques afin de faciliter la réutilisation des données.

L'ouvrage *La diffusion numérique des données en SHS – Guide des bonnes pratiques éthiques et juridiques* (2018) rédigé par le groupe de travail Ethique et Droit présente des bonnes pratiques pour faciliter et répondre aux difficultés de l'ouverture des données, notamment au travers de retours d'expériences.

Le guide *Partager les données liées aux publications scientifiques*, produit par le comité pour la science ouverte, est un court document donnant quelques conseils et éléments de repérages sur l'ouverture des données.

Des outils pour les jeux de données quantitatives

Le package sdcMicro (gratuit) disponible sur R offre un support pour l'anonymisation des données quantitatives. Une interface explicative du fonctionnement est utilisable via la fonction « sdcApp() » dans la console R.

Le logiciel QAmyData (gratuit) permet de faciliter la standardisation de fichiers, uniformiser les noms de fichiers, vérifier la bonne mise en forme de bases de données et de leurs métadonnées. Il est nécessaire de se former au logiciel, mais celui-ci permet la personnalisation des informations et de leurs critères ainsi que leur vérification automatique.

12. Unité régionale de formation à l'information scientifique et technique.
13. Commission nationale de l'informatique et des libertés.

Contacts

L'équipe projet du PPR Autonomie se tient à votre disposition pour vous orienter vers des ressources ou des personnes compétentes. Des actions sont conduites afin de favoriser l'émergence et le développement d'une communauté de pratiques autour des expériences de gestion, diffusion et réutilisation des données.

N'hésitez pas à nous suivre pour rester informé de nos actualités !

Linkedin

Site du PPR Autonomie

S'inscrire à la newsletter du PPR Autonomie

Contacter l'équipe projet