Programme Prioritaire de Recherche
AUTONOMIE

LA RÉUTILISATION DES DONNÉES

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 d'exposer un certain nombre de ressources pour faciliter le travail des équipes dans la réutilisation 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 diffusion des données.

Pourquoi réutiliser des données ?

La réutilisation de données est le fait d'analyser des données produites, mises en forme et mises à disposition par d'autres pour un autre usage que celui initialement prévu. Réutiliser des données existantes, en tant que telles ou pour les croiser avec un autre jeu de données, est d'une importance capitale pour les chercheurs. Cependant, nombre d'entre elles ne sont pas assez visibles ou sont trop peu remobilisées.

L'analyse de données existantes peut se justifier par :
- La production d'une nouvelle analyse de données collectées dans le cadre d'un autre projet de recherche ;
- La comparaison de deux jeux de données collectées dans le cadre de travaux de recherche différents ;
- L'utilisation de données pour une première confrontation à un nouveau terrain afin d'adapter le protocole de collecte ou les instruments de celui-ci ;
- L'objectif de reproduire ou de nuancer les conclusions issues de publications reposant sur ces données ;
- La familiarisation à un nouveau champ de recherche ou des nouvelles méthodes dans un contexte pédagogique.

La réutilisation de données dans le cadre de projet de recherche peut également se justifier par :
- L'incitation des financeurs dans le cadre d'appels à projets ;
- Les coûts financiers et humains élevés de la production de nouvelles données, d'autant plus si le protocole de collecte est ambitieux. Réutiliser des données déjà constituées permet d'économiser ces ressources pour les allouer à d'autres postes ;
- L'utilisation de données pour lesquelles les questions éthiques et de confidentialité ont déjà été réglées ;
- La singularité des données mises à disposition en raison de la position particulière du producteur, qui par exemple est le seul à pouvoir produire ce type de données. Il est aussi possible que le contexte de la collecte ait évolué, ayant comme conséquence qu'une collecte, aujourd'hui, ne pourrait concerner la même population, ou ne saurait être rétrospective, ne permettant pas d'études historiques ou des comparaisons temporelles ;
- L'appariement possible de certaines données (notamment à des données administratives) permet d'enrichir considérablement les informations disponibles. On peut citer, à titre d'exemple, le CSNS[1] géré par l'INSEE, qui permet de croiser à l'échelle des individus différentes sources administratives.

Si des données ont déjà été analysées une première fois, cela ne signifie pas que toutes les exploitations pertinentes ont été réalisées. Les chercheurs ont rarement le temps d'exploiter tous les angles de leurs matériaux. Un autre bagage théorique, ou un autre regard disciplinaire, peut permettre de poser un nouveau regard et de nouveaux résultats sur ces données. Enfin, d'autres données peuvent aussi mettre en perspective les premières en apportant un nouveau regard, concordant, discordant ou qui les complète par de nouvelles informations.

1. Code statistique non signifiant.

Quelques conseils

Des difficultés pour accéder aux jeux de données peuvent être rencontrées (données sensibles) ; un passage devant un comité peut s'imposer, par exemple le Comité du secret statistique. Il est donc nécessaire d'anticiper dès le début du projet les besoins de disposer de jeux données sensibles.

En amont du projet, il faut penser à récolter un maximum d'informations sur le jeu de données, pour en avoir une compréhension fine et éviter les erreurs d'interprétation. En effet, utiliser un jeu de données existant suppose d'anticiper certaines questions. Les données peuvent ne pas tout à fait correspondre à la question de recherche initialement identifiée ; pour autant, elles permettent de porter un nouveau regard sur la question, voire invitent à faire un pas de côté par rapport au projet initialement prévu. Si des questions subsistent sur le jeu de données, son contexte de collecte ou de traitement, il ne faut pas hésiter à contacter le producteur.

À noter également :
- Éviter de réutiliser un jeu de données en dehors de tout cadre légal qui protège autant la personne ayant mis à disposition les données que celle qui les réutilise ;
- Respecter la convention d'utilisation ainsi que la licence associée et supprimer le jeu de données à la fin de la réutilisation ;
- Ne pas chercher à identifier des individus ou à apparier des données, si cela n'est pas prévu dans la convention de réutilisation ;
- Réutiliser certaines données suppose d'adapter son organisation du travail afin d'éviter des ruptures de confidentialité, par exemple. À ce titre, le CASD[1] loue un appareil indispensable à l'accès à certains jeux de données, dont le déplacement est strictement encadré ;
- Ne pas diffuser les jeux de données en sa possession ;
- Citer le jeu de données utilisé à l'occasion d'une publication, en utilisant le format de citation défini par les entrepôts de données et l'identifiant du jeu de données lorsqu'il est fourni (par exemple le DOI[2]). Ce dernier permet de retrouver le jeu de données initial et de faciliter le suivi des réutilisations par les services bibliométriques.

1. Centre d'accès sécurisé aux données.
2. *Digital Object Identifier,* identifiant pérenne et unique, permet de référencer, citer et fournir un lien stable vers un fichier en ligne.

Ressources

Afin d'accompagner les chercheurs dans la réutilisation de données, voici une liste non-exhaustive de ressources. Celles-ci peuvent aider à identifier des données pertinentes, à accéder aux jeux de données, trouver des outils ou des services sur les méthodes de réutilisation. Il est à signaler que certains portails de données ont mis en place des moteurs de recherche dont l'interrogation peut porter non seulement sur les titres et les descriptions de leurs jeux de données, mais aussi sur les variables de ces jeux de données. C'est notamment le cas de data.progedo ou de la base de données de questions du Ethmig Survey Data hub, spécialisé dans les enquêtes sur les minorités ethniques et migrantes.

Des portails interdisciplinaires

Zenodo, un entrepôt généraliste, permet d'accéder à de nombreux jeux de données de toutes disciplines. Néanmoins, fonctionnant en auto-dépôt, la documentation peut parfois être lacunaire et l'accès soumis à l'accord des personnes ayant diffusé le jeu.

Recherche Data Gouv propose un catalogue de jeux de données de la recherche française qui signale les données déposées dans des entrepôts nationaux ou internationaux, thématiques ou disciplinaires.

Isidore est un moteur de recherche qui collecte, enrichit et offre un signalement et un accès unifié à de nombreuses ressources, y compris à des données. Néanmoins, le portail est enrichi par moissonnage automatique ; il n'est donc pas exhaustif et l'accès aux jeux de données recensés n'est pas garanti.

DataCite Commons recense toutes les ressources scientifiques ayant reçu un DOI, y compris les jeux de données.

Des portails spécialisés en SHS

Quetelet-Progedo-Diffusion met à disposition plus de 1 500 jeux de données qualitatifs et quantitatifs issus d'enquêtes, de la statistique publique ou de données administratives. Il regroupe les données diffusées par les ADISP[3], le CDSP[4] de SciencesPo et le Datalab de l'Ined[5].

Le CASD spécialisé dans les données fortement sensibles, met à disposition des équipes de recherche de nombreux jeux de données administratives ou d'enquêtes protégés (par exemple les différents volets de l'enquête CARE[6]). Cet accès est payant (quelques centaines d'euros) et peut être

3. Archives de données issues de la statistique publique.
4. Centre de données socio-politiques.
5. Institut national d'études démographiques.
6. Capacités, aides et ressources des seniors.

soumis à l'accord du producteur ou du Comité du secret statistique. Par ailleurs, le CASD propose des appariements de jeux de données et les encadre.

Data.sciencepo héberge les jeux de données du CDSP, mais aussi des jeux de données mis à disposition par les équipes de SciencesPo. Il regroupe aujourd'hui près de 400 jeux de données.

BeQuali concerne les données qualitatives et permet l'accès à plus d'une vingtaine de corpus d'entretiens avec une vaste documentation pour chacun.

Le CESSDA[7] moissonne le catalogue de ses représentants dans chaque pays d'Europe, dont celui de Quetelet-Progedo-Diffusion pour la France. S'il n'héberge pas les données, il permet d'avoir une vision large des données disponibles en Europe.

Des portails pour les données de la statistique publique ou administratives

Les services statistiques ministériels, par exemple du Ministère des solidarités et de la santé (DREES[8]) ou du Ministère du travail (DARES[9]) mettent directement à disposition des jeux de données anonymisés ou agrégés sur leur site internet ou sur une plateforme rattachée (data.drees par exemple).

Data Gouv compte près de 42 000 jeux de données numériques produits par les acteurs publics (ministères, collectivités, établissements publics, etc.) ou privés (entreprises, associations, citoyens, etc.).

Data Europa compile les données publiques nationales en libre accès de 36 pays ; elle compte plus d'1,4 millions de jeux de données.

7. Consortium of European Social Science Data Archives.
8. Direction de la recherche, des études, de l'évaluation et des statistiques.
9. Direction de l'animation de la recherche, des études et des statistiques.

Plus de portails

Il existe de nombreux autres entrepôts mettant à disposition des jeux de données en France ou à l'international. En fonction des disciplines ou des données recherchées, il peut être pertinent d'aller interroger des plateformes plus spécialisées. Cat.OPIDoR réalise un recensement des portails et entrepôts français ; Re3Data fait ce même travail à l'échelle internationale.

Se former et être conseillé

Divers services dispensent un accompagnement dans la réutilisation de données existantes et proposent des formations ou des outils.

Les PUD[10] ont pour rôle d'assister les chercheurs dans toutes les démarches relatives aux données, notamment en ce qui concerne la bonne compréhension de celles-ci.

Pour certains jeux de données, des groupes de travail rassemblent différentes équipes réutilisatrices pour faciliter les échanges sur des problèmes méthodologiques, partager des solutions, etc. Ce type d'organisation est notamment mis en place pour les données issues de l'échantillon démographique permanent ou à l'occasion de la production de certaines enquêtes de la statistique publique.

La chaîne Youtube de UK Data Service propose des vidéos avec des conseils et des outils pour réutiliser les données : comment trouver et interpréter les métadonnées ? Comment analyser différents types de données (recensement, données longitudinales, etc.) ?

10. Les plateformes universitaires de données.

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