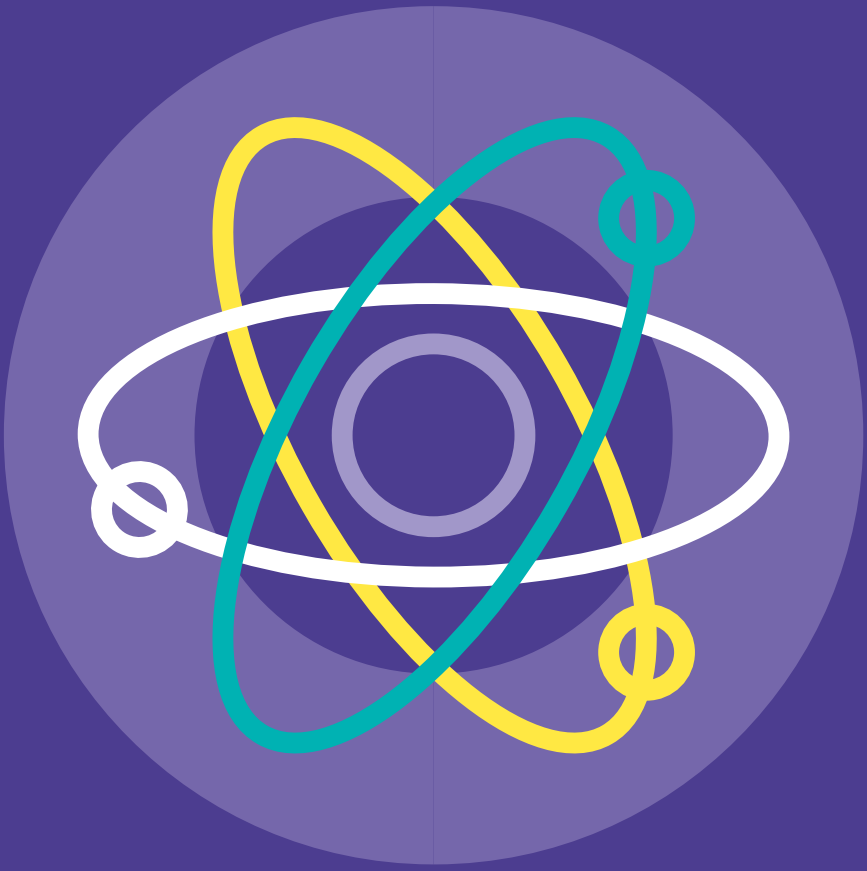
MINISTÈRE
DE L'ENSEIGNEMENT
SUPÉRIEUR,
DE LA RECHERCHE
ET DE L'INNOVATION

Liberté
Égalité
Fraternité

Ouvrir la science !

Partager les données liées aux publications scientifiques

GUIDE POUR LES CHERCHEURS



En matière de recherche scientifique, les publications constituent des vecteurs classiques de diffusion des connaissances. Les résultats qui y sont présentés reposent de plus en en plus sur des données et des analyses sous-jacentes. Le partage des données associées aux publications est donc un élément important participant de la qualité des travaux de recherche. L'objectif de ce guide est de vous familiariser avec les étapes du partage des données liées aux publications en privilégiant une approche concrète. *Alors, on s'y met ?*

1. Pourquoi partager les données liées aux publications ?

L'ouverture des données favorise la **transparence et la reproductibilité** du processus scientifique. En effet elle assure leur disponibilité pour tous et favorise leur réutilisation, permettant ainsi plus de transparence. Elle permet d'étayer plus solidement les résultats exposés dans une publication scientifique. En rendant les données disponibles pour tous, elle favorise leur réutilisation par la communauté scientifique et leur prise en main par les citoyens, et permet leur mobilisation dans le cadre du débat public.

Ouvrir les données, c'est les exposer à la critique. Pour s'y préparer, la stratégie consiste à les documenter mais aussi à mettre en place, tout au long de leur cycle de vie, des actions de gestion visant à préserver leur **qualité**. Les données ainsi mises à disposition bénéficient d'une **traçabilité** accrue et augmentent leur potentiel de réutilisation, y compris pour leur producteur.

La diffusion ouverte des données assure la **reconnaissance** de leurs producteurs et leur **visibilité** ainsi que celle de leur établissement, au même titre que la publication scientifique assure la visibilité de ses auteurs. Un jeu de données ouvert augmente sa visibilité et donc ses chances d'être utilisé par un autre projet de recherche puis cité de manière analogue à une publication. Les personnes qui y sont associées voient leur **implication valorisée**. Il est également à noter qu'une publication accompagnée des données est **davantage citée**.

Les données produites pendant un projet de recherche peuvent avoir de la valeur et un intérêt au-delà du projet, et parfois de la discipline initiale. Les rendre disponibles permet d'exploiter pleinement leur potentiel, favorisant ainsi **l'interdisciplinarité** et **la collaboration** académique.

2. Comment partager des données liées aux publications ?

Lors de la rédaction d'un article scientifique, les auteurs adoptent naturellement une approche pédagogique consistant à bien définir toutes les notations et conventions utilisées dans l'article, à détailler les hypothèses, le cadre de travail ainsi que l'état de l'art afin d'en faciliter la lecture et en permettre la compréhension. Les données font partie de cette approche. Si l'on veut que les données partagées puissent être utiles à la communauté, les mêmes attentions doivent être portées à la publication des données, y compris en amont de leur partage.

Préparation et documentation des données

Décrire les données afin de les rendre intelligibles à toute personne n'ayant pas participé à leur production constitue une étape préalable à leur diffusion. Les informations sur la provenance des données, les hypothèses ou contraintes liées à leur production et les protocoles expérimentaux qui y sont associés doivent faire partie de l'ensemble des informations descriptives de ces données, encore appelées métadonnées. Il existe des standards de métadonnées génériques ou propres à certaines communautés. Pour accompagner ce processus de gestion continue de données, on peut par ailleurs s'appuyer sur un **Plan de Gestion des Données**, qui est un document permettant de définir les modalités de suivi et de description des données.

Au moment de partager les données, plusieurs éléments sont à prendre en compte. Certaines données sont concernées par des contraintes juridiques qui empêchent leur partage ou rendent nécessaires des traitements d'anonymisation ou des demandes d'autorisation. Chaque établissement de recherche possède sa propre politique d'ouverture des données et sa prise en compte, dans le cadre de la législation en vigueur, est un préalable important au choix du moyen qui permettra de partager les données.

Il recommandé **de ne pas confier les données à partager aux éditeurs des revues** qui proposent de les publier sous forme de «*supplementary data*» ou de «*supplementary materials*» associés à l'article. Une telle publication se fait encore souvent dans format et un environnement qui ne permettent pas de documenter correctement les données et rendent difficile leur réutilisation. Elle peut aussi s'accompagner d'une demande de transfert exclusif de droit contraire à la loi française et à l'esprit de la science ouverte. Enfin, dans certains cas, elle contribue à rendre les utilisateurs captifs au sein d'environnements maîtrisés par de grands acteurs commerciaux de l'édition scientifique.

Il est donc plutôt recommandé d'utiliser pour le partage des données des **entrepôts de données institutionnels, généralistes ou disciplinaires**, qui permettent d'éviter ces écueils et offrent un environnement dédié à la documentation, l'ouverture et la réutilisation de la donnée de recherche. Établir correctement l**e lien entre le jeu de données déposé dans l'entrepôt et l'article disponible sur une plateforme de publication** devient alors une nécessité et une démarche à anticiper. On utilisera plutôt des entrepôts de données institutionnels, généralistes ou disciplinaires.

Choix de l'entrepôt

- Dans le cas de disciplines structurées pour le partage des données (astronomie, génomique, etc…), les producteurs de données ont à disposition des **entrepôts spécifiques à leur discipline**. Ils utiliseront alors naturellement l'ensemble des standards et bonnes pratiques déjà en place pour documenter et mettre en forme leurs données. La pratique de sa communauté est le meilleur guide mais des annuaires de ces entrepôts existent.

- En alternative, les producteurs de données pourront se tourner vers l'**entrepôt institutionnel** auquel ils sont affiliés, s'il existe, ou utiliser l'**entrepôt pluridisciplinaire Recherche Data Gouv**. Dans ces deux cas, des exigences minimales seront imposées par les entrepôts et la charge de s'assurer de la qualité de la documentation des données sera davantage portée par le déposant.

L'entrepôt national Recherche Data Gouv :

La plateforme nationale Recherche Data Gouv propose un entrepôt de données pluridisciplinaire qui sera opérationnel dès 2022 : Il assure la souveraineté française sur les données, est conforme aux droits français et communautaire, garantit la pérennité et l'indexation des données stockées, suivant les principes FAIR. C'est l'entrepôt de choix quand aucun entrepôt disciplinaire n'existe.

Quel que soit l'entrepôt choisi pour partager les données, celui-ci se doit en particulier d'offrir les fonctionnalités suivantes :

- L'assignation d'un **identifiant pérenne** (*Persistent Identifier* : PID) de type DOI qui permet de citer les données (par exemple http://dx.doi.org/10.15497/RDA00027) et constitue la brique de base pour établir le lien avec d'autres produits de la recherche comme les publications.
- La **description des données** à un niveau suffisant pour en faciliter la découverte, la compréhension et la réutilisation (métadonnées descriptives standardisées, vocabulaires disciplinaires contrôlés).
- L'utilisation de **licences** et la définition de **règles d'accès** permettant d'inscrire la réutilisation dans un cadre légal bien défini et compatible avec le droit français et européen.
- Une **durée de conservation** minimale de plusieurs années cohérente avec la politique de conservation des données de l'établissement de rattachement.

Lier les données aux publications

Plusieurs options sont disponibles pour établir la liaison entre un article et des données qui lui sont associées **avant la publication de l'article** considéré. Il est alors facile de créer le lien entre l'article et les données associées (Fig. 1). De même, le référencement des publications associées aux données (y compris les *data papers*) est généralement possible dans tous les entrepôts de données, même après le dépôt initial.

À l'inverse, indiquer le lien explicite vers les données **après la publication d'un article** est le plus souvent impossible à l'heure actuelle. Une solution de secours consiste à faire référence aux données dans la version de l'article déposée dans une archive ouverte (HAL par exemple) qui permet de renseigner à tout moment des identifiants pérennes liés aux publications dans des champs spécifiques « Données associées » de l'enregistrement. Ce schéma permet donc la liaison réciproque entre publications et données, mais seulement pour sa version déposée dans l'archive ouverte.

Les *Data papers* :

Un *data paper* est une publication dont le but est la description d'un jeu de données scientifiques brutes. Contrairement à un article de recherche classique, le *data paper* consiste en une description détaillée des données scientifiques, des métadonnées, ainsi que les circonstances et méthodes de leur collecte mais sans analyse ou interprétation de ces données. Le contenu du *data paper* est consacré à la description détaillée des données, notamment les métadonnées, permettant à n'importe quelle personne tierce d'exploiter ces données. Les données décrites doivent être accessibles (dans la mesure du possible), déposées dans un entrepôt approprié, et munies d'un identifiant pérenne type DOI.

Comment lier des données à un article avant publication dans une revue classique ?

Vos données ne sont pas encore finalisées ?

Vous pouvez pré-réserver un PID !

1. Préréservez un PID pour vos données auprès de l'entrepôt de votre choix

2. Indiquez ce PID dans l'article à publier

3. Quand les données seront finalisées, n'oubliez pas de les déposer dans l'entrepôt de données !

4. Et quand l'article sera publié, indiquez le PID article dans l'entrepôt où sont déposées les données si celui-ci le permet.

Vos données sont finalisées ?

Vous pouvez les déposer dans un entrepôt !

1. Déposez vos données dans l'entrepôt de votre choix : un PID est attribué à vos données

2. Indiquez ce PID dans l'article à publier

Vos données ont été valorisées par la publication d'un data papers ?

Vos données ont donc déjà été déposées dans un entrepôt !

Indiquez le attribué à vos données par l'entrepôt + le PID du Data paper dans l'article à publier

Figure 1 : diversité des processus de liaison des données de recherche à un article avant sa publication.

Un *data paper* est publié sous forme d'un article revu par les pairs, gage de sa qualité, et peut être cité au même titre qu'un article « classique ». Par conséquent, l'auteur d'un *data paper* doit être convaincant quant à la qualité et la portée scientifique des données (notamment leur potentiel de réutilisation). Il peut être publié dans des revues spécifiques (*data journal*) ou dans des revues scientifiques traditionnelles qui permettent ce format.

En Bref

Partager des données liées à une publication

À PRIVILÉGIER

Déposer ses données **avant** de publier son article, et ainsi lier ses données à l'article en mentionnant l'**identifiant pérenne** des données.

Déposer ses données dans un **entrepôt de données** dédié indépendant (disciplinaire ou institutionnel).

À ÉVITER

Déposer ses données dans un entrepôt **après** avoir publié son article.

Confier les données à l'éditeur de la revue pour une diffusion sous forme de « supplementary materials » sur sa plateforme de publication.

Glossaire

- **Données de la recherche :** les documents factuels (notes numériques, documents textuels, images et sons) utilisés comme sources primaires pour la recherche scientifique, et qui sont communément acceptés dans la communauté scientifique comme étant nécessaires pour valider les résultats de la recherche. Pour aller plus loin : https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0347

- **Entrepôts de données :** plateformes sur lesquelles sont déposés, décrits et conservés des jeux de données de la recherche. Les entrepôts peuvent être généralistes ou disciplinaires.

- **FAIR :** ensemble de principes visant à soutenir la recherche en facilitant la réutilisation des données. Faciles à trouver (Findable), Accessibles (Accessible), Interopérables (Interoperable), Réutilisables (Reusable). Pour aller plus loin : https://www.ouvrirlascience.fr/fair-principles/

- **Métadonnées :** ensemble d'informations structurées qui décrit, explicite, localise une ressource informationnelle, dans le but d'en faciliter la recherche, l'usage, et la gestion. Pour aller plus loin : https://groups.niso.org/apps/group_public/download.php/17446/Understanding%20Metadata.pdf

- **Open data :** données ouvertes, dont l'accès est libre et sans restriction. Pour aller plus loin : https://sparceurope.org/new-sparc-europe-report-analyses-open-data-open-science-policies-europe/

- **PID :** identifiant unique pérenne

- **Licence :** mention définissant les conditions de réutilisation des données

Références citées

1. Colavizza G, Hrynaszkiewicz I, Staden I, Whitaker K, McGillivray B (2020) The citation advantage of linking publications to research data. PLOS ONE 15(4): e0230416. https://doi.org/10.1371/journal.pone.0230416

2. https://doranum.fr/metadonnees-standards-formats/fiche-synthetique/

3. https://doranum.fr/plan-gestion-donnees-dmp/minute/

4. https://repositoryfinder.datacite.org/

5. https://doranum.fr/aspects-juridiques-ethiques/les-licences-de-reutilisation-dans-le-cadre-de-lopen-data-2/

Pour aller plus loin

- Guide de bonnes pratiques sur la gestion des données de la recherche : https://mi-gt-donnees.pages.math.unistra.fr/guide/00-introduction.html
- Dedieu, L.; Barale, M. 2020. Déposer des données dans un entrepôt, en 6 points. Montpellier (FRA) : CIRAD, 4 p. https://doi.org/10.18167/coopist/0070
- Dedieu, L. 2014. Rédiger et publier un data paper dans une revue scientifique, en 5 points. Montpellier (FRA) : CIRAD, 7 p. https://doi.org/10.18167/coopist/0057
- Deboin, M.C. 2021. Citer un jeu de données scientifiques, en 4 points. Montpellier (FRA) : CIRAD, 4 p. https://doi.org/10.18167/coopist/0058
- How to cite datasets and link to publications https://www.dcc.ac.uk/sites/default/files/documents/publications/reports/guides/How_to_Cite_Link.pdf

Partager les données liées aux publications scientifiques.
Guide pour les chercheurs
mars 2022

Direction de la publication :
Ministère de l'Enseignement supérieur,
de la Recherche et de l'Innovation

Coordination éditoriale :
Collège des données de la recherche
du Comité pour la science ouverte

Rédacteurs :
Baptiste Cecconi (Observatoire de Paris)
Jean-Yves Chatelier
Bénédicte Kuntziger (CCSD - CNRS)
Yvette Lafosse (Inist-CNRS)
Jean-François Martin (Institut Agro)
Kenneth Maussang (Université de Montpellier)
Claire Sowinski (Inist-CNRS)
Corentin Spriet (CNRS)
Coralie Wysoczynski (Inist-CNRS)
Carlo Maria Zwölf (Observatoire de Paris)
Les rédacteurs s'expriment en leur nom propre
et non au titre de leur employeur.

Conception graphique :
Opixido

MINISTÈRE
DE L'ENSEIGNEMENT
SUPÉRIEUR,
DE LA RECHERCHE
ET DE L'INNOVATION

Liberté
Égalité
Fraternité

Ouvrir la science !