SCIENCE OUVERTE

DONNÉES DE LA RECHERCHE

PASSEPORT POUR LA
SCIENCE
OUVERTE

TABLE DE MATIÈRES

LÉGENDE

Le texte souligné renvoie au glossaire.
▼ signale des outils donnés à titre d'exemple.

La version numérique de ce guide est disponible sur www.ouvrirlascience.fr

Inscrit dans la collection Passeport pour la science ouverte, ce guide aborde les principales notions relatives à la gestion et à la diffusion des données de la recherche. Il est à destination des chercheuses et chercheurs comme vous, quelle que soit votre discipline ! Vous trouverez au fil de votre lecture des explications pour comprendre ce que sont les données de la recherche, les enjeux liés à leur gestion raisonnée et l'intérêt de leur diffusion, mais également comment être accompagné au mieux dans leur gestion et leur partage.

Isabelle Blanc
Administratrice ministérielle des données,
des algorithmes et des codes sources
Ministère de l'Enseignement supérieur et de la Recherche

DONNÉES DE LA RECHERCHE DE QUOI PARLE-T-ON ?

Les données de la recherche dans tous leurs états

Plusieurs définitions existent ; la plus couramment utilisée est celle de l'Organisation de Coopération et de Développement Économique (OCDE) qui définit les données de la recherche 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 ».

BON À SAVOIR

Les codes sources et logiciels ne doivent pas être considérés comme des données : ils présentent des enjeux, pratiques et recommandations de partage et d'ouverture particuliers. Consultez le livret *Codes et logiciels*.

Les données de la recherche peuvent être des enregistrements sonores, des images vidéo, des images satellitaires, des images issues de microscopes, un corpus textuel, des transcriptions, un tableau de résultats d'une enquête ou d'un test, des relevés de température d'une série temporelle ou toute autre mesure sur le terrain, le contenu d'une base de données...

Les données de la recherche sont caractérisées notamment par :

- Leur mode d'obtention : données produites dans le cadre d'expérimentation ou d'analyse par des instruments, données d'observation, données collectées lors d'une enquête ou d'un prélèvement sur le terrain... Vous pouvez produire vous-même des données ou réutiliser des données produites par ailleurs.
- Leur type : données textuelles, audiovisuelles, données numériques, d'imagerie, d'observation, de séquences génomiques... produites par certains instruments de mesure, d'analyse ou d'observation.
- Leur format : données avec format ouvert ou propriétaire.
- Leur contexte de production : partenariat industriel, laboratoire en zone à régime restrictif.
- Leur régime juridique : données à caractère personnel (Règlement général pour la protection des données), couvertes par un secret (professionnel, défense ou industriel par exemple), soumises à un accord de confidentialité, encadrées par des obligations contractuelles (contrat d'accès).
- Leur criticité : données sensibles, confidentielles.

Toutes ces notions servent à décrire les données et constituent des métadonnées.

Les **métadonnées scientifiques** apportent des informations sur la donnée, notamment : protocole et contexte d'obtention, références temporelles, paramètres d'instruments utilisés, outils et logiciels d'analyse, etc. en utilisant les vocabulaires contrôlés du domaine de recherche.
Les **métadonnées documentaires** apportent des informations plus spécifiques comme l'établissement et les producteurs, les conditions d'utilisation et d'accès, l'identifiant pérenne du jeu de données, l'identifiant des publications et codes logiciels liés aux données, etc.

Un jeu de données ou *dataset* est un ensemble cohérent de données dans le cadre d'un même projet, sur un même objet d'étude ou recueillies sur un même lieu. Toutes les données d'un *dataset* peuvent donc être décrites avec une majorité de métadonnées communes.

Les étapes-clés pour l'ouverture des données de la recherche

Au cours d'un projet de recherche, les données sont collectées, générées ou réutilisées, puis elles sont stockées afin de pouvoir être traitées et analysées. Elles seront ensuite structurées, nettoyées et triées pour ne conserver que celles qui sont pertinentes pour être diffusées ou publiées.

Par ailleurs, certaines données, notamment d'observation temporelle sont aussi archivées pour être conservées sur le long terme.

Ces différentes étapes jalonnent un projet de recherche et constituent le cycle de vie de la donnée. Une bonne gestion de ses données vise à les rendre Faciles à trouver, Accessibles, compréhensibles par les humains et les machines, c'est-à-dire Interopérables, et Réutilisables. C'est ce qu'on appelle les principes FAIR. Ces principes recouvrent les différentes manières dont les données de la recherche se construisent, se conservent, se présentent, se partagent et se réutilisent (voir aussi "Principes FAIR" p. 13-15).

L'enjeu du processus de FAIRisation des données est de permettre *in fine* leur réutilisation par l'équipe productrice comme par d'autres et directement par des machines, afin de nourrir d'autres recherches, des méta-analyses et modèles à grande échelle (climat, biodiversité, pandémie, apprentissage par l'intelligence artificielle…).

Statut juridique des données de la recherche

L'un des objectifs de la recherche publique est d'assurer l'accès libre aux données de la recherche scientifique. Cela est reconnu par l'article L.112-1 du Code de la recherche. Le régime juridique applicable à la science ouverte est désormais gouverné par le principe général selon lequel les données de la recherche doivent être aussi ouvertes que possible et pas plus fermées que nécessaire. Lorsqu'elles correspondent à des données publiques, les données de recherche sont par ailleurs soumises à un principe d'ouverture par défaut (Open Data), introduit par la loi pour une République numérique et inscrit désormais dans le code des relations entre le public et l'administration (CRPA).

L'article L. 533-4 du Code de la recherche , issu de l'article 30 de la loi pour une République numérique prévoit par ailleurs, après publication, un principe de libre réutilisation des données de recherche, lorsque celles-ci :

- sont issues d'une recherche inancée à plus de 50 % par des fonds publics,
- ne sont pas protégées par un droit spécifique ou par une réglementation,
- ont été rendues publiques par l'établissement ou l'organisme de recherche.

Les établissements publics sont les garants de la mise en œuvre de l'ouverture des données publiques. Consultez le ▼**guide d'application de la loi pour une République numérique pour les données de la recherche** pour en apprendre davantage sur sa mise en pratique.

La gestion des données de la recherche doit donc être raisonnée, l'ouverture étant le principe général et la fermeture, l'exception. La décision de maintenir des données fermées doit s'appuyer sur des motifs découlant d'autres dispositifs juridiques qui constituent des exceptions au principe général d'ouverture. L'établissement d'un Plan de gestion de données constitue un moment privilégié pour examiner ces questions juridiques.

Voir aussi « Respect des exceptions », p. 20

LES ÉTAPES-CLÉS POUR L'OUVERTURE DES DONNÉES DE LA RECHERCHE

DÉBUT DU PROJET DE RECHERCHE

PLANIFICATION
Mise en œuvre d'un plan de gestion des données et de leur FAIRisation

Réfléchissez avec vos collègues du projet aux questions clés : Quel(s) type(s) de données allez vous produire ? Quel volume ? Comment allez-vous les décrire ? Les traiter ? Les analyser ? Les partager ? Les conserver ? Existe-t-il un cadre législatif particulier pour le protocole de diffusion ?

RÉUTILISATION

Découvrez des données dans différents catalogues et entrepôts. Vérifiez les conditions de réutilisation des données, définies par leur licence. **Citez** ces données *via* leur identifiant pérenne.

DONNÉES CITABLES FAIR

PUBLICATION POUR OUVERTURE
Préparation pour ouverture ou partage (accès restreint)

Pensez à lier les données aux publications associées et vice versa grâce à leurs identifiants pérennes.

DONNÉES VALIDÉES

DÉPÔT

Déposez les fichiers. **Définissez** les droits de réutilisation (licence) et d'accès aux données : ouverture ou accès restreint. **Assurez** la mise à jour des versions des jeux des données.

COLLECTE
CRÉATION
STOCKAGE

Interrogez-vous sur le besoin de produire de nouvelles données et sur la possibilité de réutiliser celles de recherches précédentes.

Privilégiez des logiciels libres et des formats ouverts pour permettre une meilleure compatibilité entre outils.

Consultez comment sont gérés les modalités de stockage, les protocoles de sécurité, les droits d'accès et de récupération des données en cas d'incident.

DOCUMENTATION

Inventoriez les données préexistantes utilisées et les nouvelles données collectées. **Décrivez-les** grâce à des métadonnées scientifiques : date de création, provenance, méthode de collecte ou protocole d'obtention ; utilisez les vocabulaires contrôlés de votre communauté, précisez le type, le format.

DONNÉES EN COURS DE TRAITEMENT

ANALYSE
TRAITEMENT
CALCUL

Réfléchissez au besoin des traitements des données et à l'impact environnemental et social de votre recherche.

ARCHIVAGE

Tout document a vocation à être conservé de manière pérenne ou être détruit à l'issue de sa durée d'utilité administrative.

Au-delà de la vie du projet, certaines données seront conservées sur le long terme. **Renseignez-vous** auprès du service d'archives de votre établissement.

PRÉPARATION
AU DÉPÔT
Structuration
et description

Triez vos données, sélectionnez les données à déposer et publier. Confirmez le choix de l'entrepôt en vérifiant qu'il est de confiance. Préparez les fichiers des données ; renseignez les métadonnées documentaires (contributeurs, établissement, licences, financeurs du projet, ...); et les fichiers associées comme le fichier README les décrivant.

POURQUOI LES DIFFUSER ?

Le partage et l'ouverture des données de la recherche favorisent leur **réutilisation** par vous-même et par les autres : des membres de l'équipe de votre projet, de votre équipe de recherche, communauté scientifique dans son ensemble.

La diffusion de vos données contribue à **augmenter la visibilité de vos travaux** et vous permet d'être davantage cité. Selon une étude publiée dans la revue PLOS ONE, **les articles scientifiques avec des données ouvertes associées étaient 25 % plus citées**.

Diffuser les données de la recherche participe à la transparence de la démarche scientifique et accroît la confiance des citoyens envers la science. La diffusion des données contribue à la **reproductibilité** de la science, atteste de la manière dont elles ont été produites, analysées et traitées et constitue aussi un gage fort d'intégrité scientifique et éthique.

Il existe également un intérêt à diffuser les données n'ayant pas abouti à une publication ou répondu à une hypothèse scientifique initiale. Elles peuvent être utiles à d'autres chercheurs pour explorer de nouvelles hypothèses, effectuer de nouvelles recherches, y compris dans d'autres disciplines ou mettre en évidence des résultats négatifs.

BON À SAVOIR

Le numérique est un secteur en forte croissance. Consommateur de ressources abiotiques et responsable de pollutions multiples, il contribue par ses nombreux impacts environnementaux et sociaux au dépassement des limites planétaires. Les données numériques de la recherche participent à cette croissance, il est donc primordial pour ne pas accroître l'empreinte environnementale : 1) de permettre la réutilisation (principes FAIR) des données existantes avant de chercher à en produire de nouvelles, 2) de documenter le plus finement et clairement possible l'utilisation et les impacts de ses données.
Pour allier science ouverte et enjeux environnementaux, il est crucial de rendre les données trouvables et accessibles, mais aussi de détruire les données qui ne seront plus utiles car non décrites. Ces bonnes pratiques de partage et de destruction quand elles ne sont plus utiles permettent de réduire l'empreinte numérique des données.

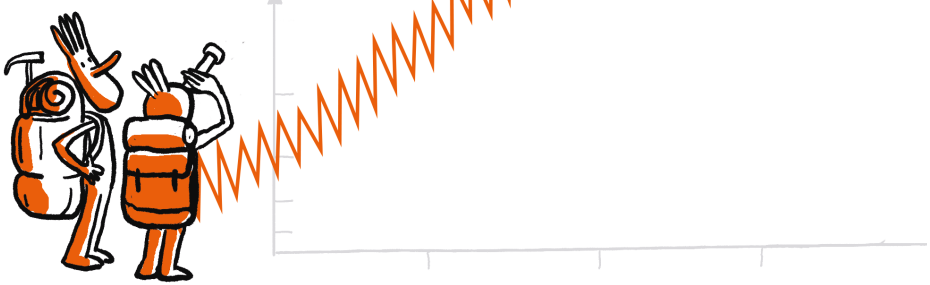
La collecte et l'analyse des données sont des étapes très coûteuses. **Des données non partagées et non diffusées sont donc des données perdues pour une équipe de recherche**. Le rapport de la Commission européenne *Cost of not having FAIR research data* paru en 2019 estime que le coût de **la mauvaise gestion des données de la recherche se chiffre à 3 milliards d'euros pour la France**, dû aux pertes de temps, à l'absence d'optimisation des coûts de stockage, aux frais de licence ou encore aux problèmes de duplication de la recherche.

Certaines données de la recherche ont un caractère unique. C'est le cas par exemple des données de suivi de long terme des paramètres environnementaux. Ces données étant des archives publiques au sens du code du Patrimoine, elles constituent un patrimoine scientifique national. Grâce à la description précise, au partage et à l'ouverture de données d'observation *in situ*, il est possible de constituer des séries temporelles et de réaliser des analyses sur plusieurs décennies afin d'évaluer par exemple l'impact du changement climatique.

Le partage et l'ouverture des données sont de plus en plus au cœur des politiques publiques. Ainsi, la diffusion des données s'inscrit-elle dans les recommandations du Plan national pour la science ouverte et des feuilles de route institutionnelles. Diffuser ses données permet de répondre aux obligations légales, aux demandes des financeurs et de certaines revues.

EXEMPLE

La courbe de Keeling est un graphique qui montre l'évolution de la concentration du CO_2 dans l'atmosphère au sein de l'Observatoire de Mauna Loa (Hawaii) de 1952 à nos jours. Ces mesures ont eu lieu dans le cadre d'un programme de l'Institut de recherche Scripps et se poursuivent actuellement dans le cadre d'un programme du *National Oceanographic and Atmospheric Administration* (NOAA).



COMMENT LES DIFFUSER ?

Préparer la diffusion de ses données

Prévoir la gestion de ses données

La diffusion des données se prépare en amont du projet de recherche. Pour cela, le plan de gestion de données ou PGD est un outil qui permet de décrire leur gestion, leur stockage, leur analyse, leur préservation et d'anticiper leur ouverture en fonction des cadres juridiques, contractuels… liées aux données du projet. Le PGD est évolutif et s'adapte à chaque étape du projet de recherche.

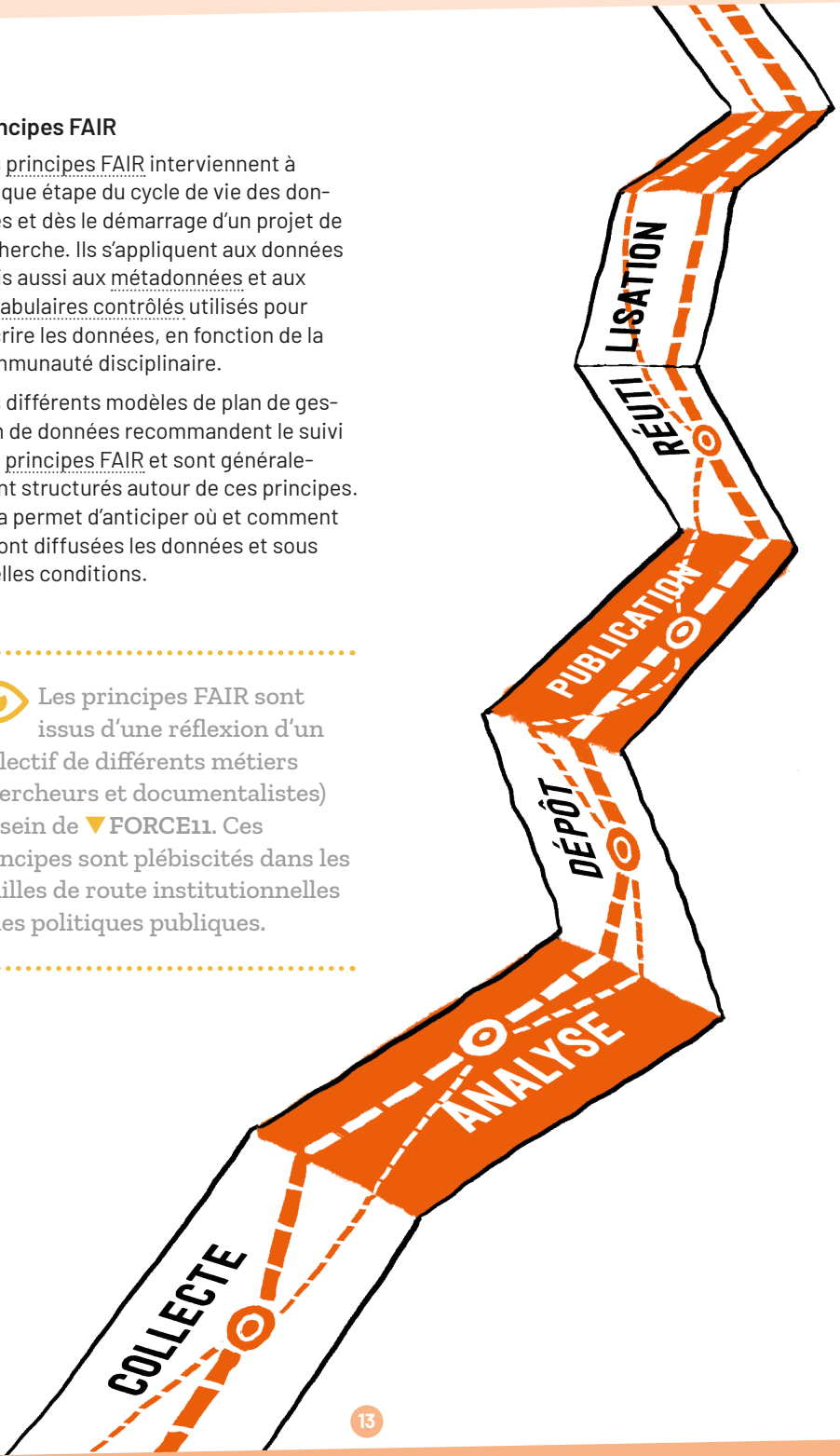
Le plan de gestion de données se présente sous la forme d'un document structuré en rubriques selon un modèle souvent recommandé ou imposé par la tutelle ou l'agence de financement. Vous trouverez des modèles sur ▼ **DMP OPIDoR**. Le PGD vise à synthétiser la description et l'évolution des jeux de données du projet de recherche. Il décrit les données, leur gestion au cours du projet, et définit les modalités de leur diffusion, réutilisation et de conservation. Il est d'autant plus important de bien l'alimenter puisque c'est un document de pilotage de la gestion des données tout au long et après le projet.

Principes FAIR

Les principes FAIR interviennent à chaque étape du cycle de vie des données et dès le démarrage d'un projet de recherche. Ils s'appliquent aux données mais aussi aux métadonnées et aux vocabulaires contrôlés utilisés pour décrire les données, en fonction de la communauté disciplinaire.

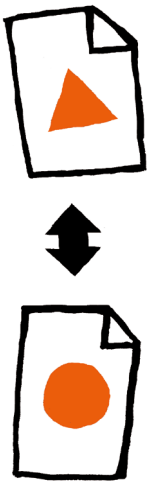
Les différents modèles de plan de gestion de données recommandent le suivi des principes FAIR et sont généralement structurés autour de ces principes. Cela permet d'anticiper où et comment seront diffusées les données et sous quelles conditions.

Les principes FAIR sont issus d'une réflexion d'un collectif de différents métiers (chercheurs et documentalistes) au sein de ▼ FORCE11. Ces principes sont plébiscités dans les feuilles de route institutionnelles et les politiques publiques.



Faciliter la découverte des données pour les humains et les machines à travers l'indexation des métadonnées

- Les données ont un *identifiant pérenne* ou *Persistent IDentifier* (PID) en anglais (par exemple le *Digital Object Identifier* ou DOI) afin de disposer d'un accès stable à la ressource.
- Les données sont décrites par des métadonnées scientifiques et des métadonnées documentaires.
- Les données, ou au moins leurs métadonnées, sont indexées ou enregistrées dans un outil de recherche, par exemple à travers le dépôt des données dans un entrepôt ou leur référencement dans un catalogue de données.



Rendre ses données ***interopérables*** pour permettre leur exploitation quel que soit l'environnement informatique utilisé par les humains et les machines

- Les données sont décrites dès le début du cycle de vie à l'aide de vocabulaires contrôlés.
- Les métadonnées doivent autant que possible faire référence aux autres données qui peuvent être mises en relation (par exemple : données naturalistes liées aux données climatiques et pédologiques) et ainsi permettre des liens entre elles.
- Les formats de fichiers utilisés sont ouverts et documentés pour permettre l'exploitation et une pérennisation des données par différents outils.

Permettre l'***accès*** aux données et aux métadonnées

- Sur Internet, à travers un protocole standard, libre et ouvert, par exemple https.
- Sur authentification pour les données en accès restreint.
- Les métadonnées doivent rester accessibles même si les données sont temporairement inaccessibles ou si l'accès aux données est restreint.

Permettre la ***réutilisation*** des données pour de futures recherches

- Les métadonnées ont une pluralité d'attributs utiles pour la compréhension et la réutilisation des données.
- Une licence de réutilisation est attribuée aux données.
- La description des données indique leur provenance.
- La structure des données suit les standards de la communauté scientifique pour faciliter leur analyse.

Implications des choix lors de la gestion des données

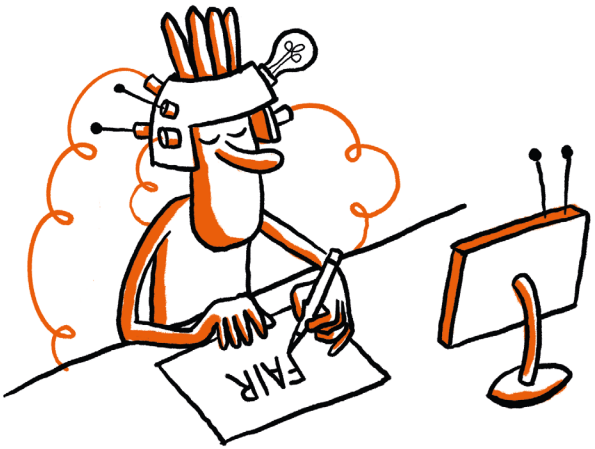
Certains choix vont avoir un impact sur la qualité de complétion des critères FAIR ainsi que sur l'environnement, via la diffusion des données : par exemple le choix d'un entrepôt, des vocabulaires ou du format.

Certaines plateformes ou entrepôts vont suggérer d'utiliser des standards de données et de métadonnées. C'est le cas de ▼**GBIF**, plateforme pour des données sur la biodiversité qui propose le standard ▼**Darwin Core** pour les données et ▼**EML** (*Ecological Metadata Language*) pour les métadonnées. Cela contribue à remplir les critères « I »et « R » des principes FAIR. ▼**Progedo** propose le standard de métadonnées ▼**DDI** (*Data Documentation Initiative*) pour décrire les données issues d'enquêtes et d'autres méthodes d'observation en sciences sociales, comportementales, économiques et de la santé.

Le choix d'un entrepôt et la manière de documenter les métadonnées influencent aussi le potentiel de réutilisation automatisée des données. Des données tabulaires nécessiteront par exemple une transformation automatisée pour qu'elles soient lisibles par les machines et ainsi incluses et liées aux autres données du web de données.

▼**FAIR-Aware** est un outil en ligne permettant de tester votre niveau de connaissance des critères FAIR.

Des outils permettent d'obtenir des suggestions d'amélioration du niveau FAIR d'un jeu de données à l'aide de son identifiant pérenne ou PID. C'est le cas de ▼**FAIR-Checker** ou ▼**F-UJI**. Ces outils, limités par leur approche automatique, peuvent induire des biais. Ils sont donc à utiliser avec précaution.



SUR LE TERRAIN

JOSHUA G.

Doctorant en sciences biomédicales
à l'Université de Lyon

Le traitement du signal est une composante majeure de la recherche dans de multiples domaines. Au cours de ma thèse, j'ai eu l'occasion d'acquérir, manipuler et traiter des données d'imagerie biomédicale issues de plusieurs modalités différentes. J'ai donc pu me rendre compte qu'au-delà de la logique des algorithmes, le format des données lui-même est un enjeu majeur souvent ignoré dans les publications.

De plus en plus de chercheurs mettent à disposition leurs outils, mais le manque d'homogénéité dans les formats des données rend souvent leur utilisation particulièrement laborieuse et chronophage, parfois décourageante. De nouvelles solutions sont développées par la communauté scientifique, et tous profiteraient de la possibilité de les utiliser simplement.

Dans mon cas, il s'agit du format BIDS (*Brain Imaging Data Structure*), qui concentre ses efforts pour la communauté des neurosciences à travers un standard unique et un écosystème grandissant d'applications qui en tirent parti.

C'est en partie pour cela que l'harmonisation des données de la recherche est un enjeu important à mon sens. C'est pour cette raison que je cherche à m'investir dans les initiatives communautaires qui naissent de ce besoin.

Qui contacter ?

Au niveau local, près des équipes de recherche, de nombreux experts peuvent vous accompagner à chaque étape d'un projet de recherche :

- Les ▼**ateliers de la donnée de Recherche Data Gouv** proposent une expertise, un accompagnement et des formations autour des données. Ils fédèrent les différents acteurs de la donnée (chercheurs, enseignants-chercheurs, documentalistes, bibliothécaires, informaticiens, archivistes, juristes, etc.) à l'échelle d'un ou plusieurs établissements. S'il n'existe aucun atelier de la donnée sur votre site ou un dispositif d'accompagnement dans votre établissement, vous pouvez consulter le répertoire ▼**SOS-PGD** qui peut vous aider à trouver le bon interlocuteur.
- Des référents « science ouverte » ou des référents « données » peuvent avoir été désignés au sein de votre laboratoire, de votre structure de recherche ou de votre bibliothèque universitaire.
- Des interlocuteurs spécifiques à certains domaines existent également. En sciences humaines et sociales, les ▼**Maisons des Sciences de l'Homme** peuvent ainsi fournir un accompagnement. Pour les sciences sociales, il est aussi possible de faire appel aux ▼**Plateformes universitaires de données** (PUD) encadrées par ▼**Progedo**.

Au niveau national, l'écosystème ▼**Recherche Data Gouv** fédère de nombreux acteurs, dont les ▼**ateliers de la donnée** et six ▼**centres de référence thématiques** :

- ▼**CDS** en astronomie et astrophysique.
- ▼**Data Terra** pour le système Terre et environnement.
- ▼**PNDB** (Pôle National de Données de Biodiversité) en écologie-biodiversité.
- ▼**Huma-Num** et ▼**Progedo** en sciences humaines et sociales.
- ▼**IFB** (Institut français de bioinformatique) en biologie-santé.

Vous trouverez également des ressources d'autoformation sur ▼**DoRA-Num**, le ▼**réseau URFIST**, ▼**Couperin**, ▼**CoopIST** ou sur la plateforme ▼**Recherche Data Gouv**.

Renseignez-vous au sein de votre établissement : existe-t-il des référents « données » ? Un atelier de la donnée ? Rapprochez-vous de l'entité en charge de la science ouverte de votre établissement ou de votre bibliothèque universitaire. Si vous êtes doctorant, vous pouvez aussi contacter votre école doctorale.

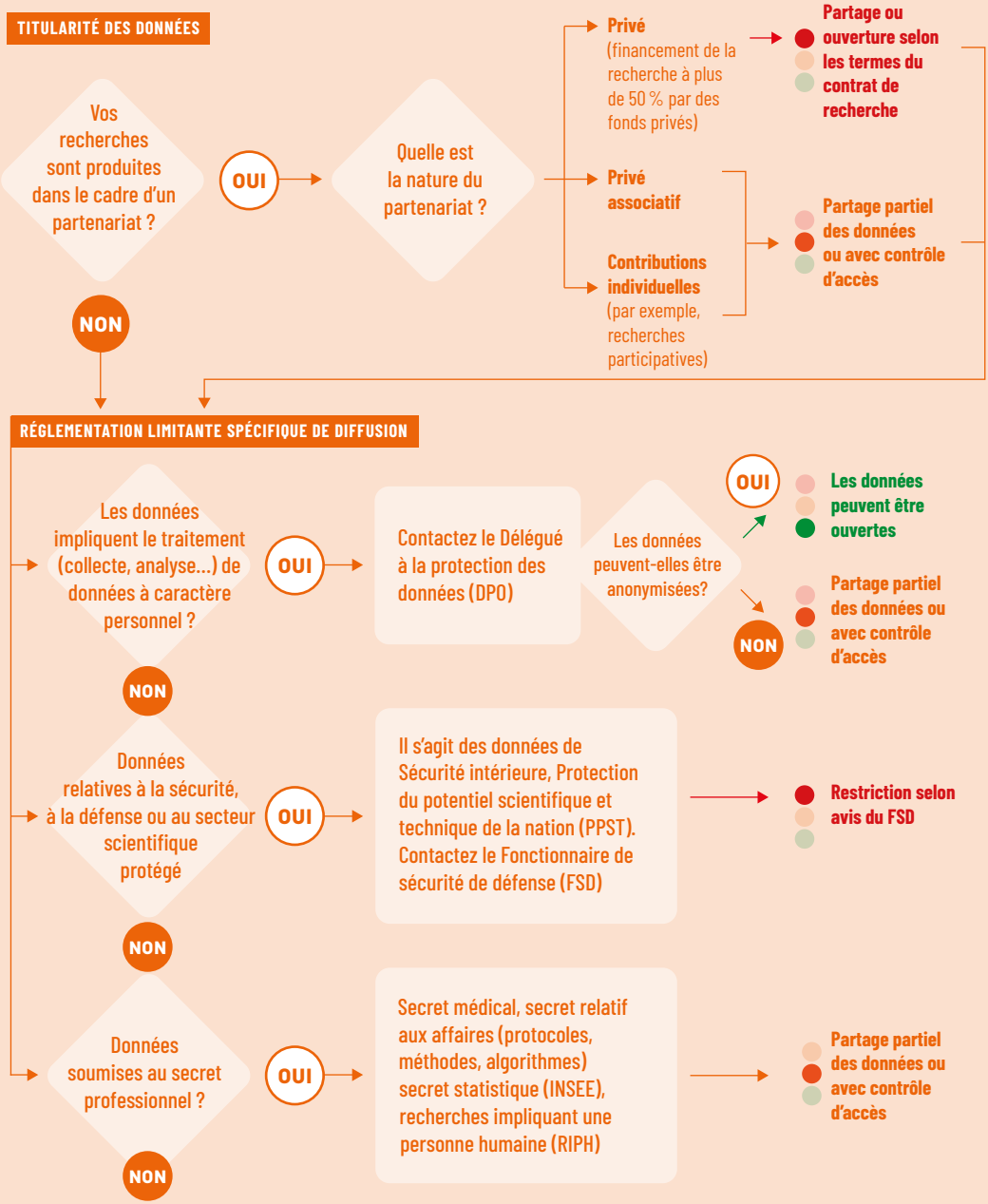
Diffuser les données de la recherche

Questions juridiques, le respect des exceptions

Les données de recherche sont soumises à un principe général d'ouverture par défaut, mais ce principe comporte des exceptions, fixées par la loi.

Il faut donc se demander si des droits assurant la prise en compte d'intérêts légitimes existent afin de déterminer le régime applicable aux données collectées et produites au cours d'une activité de recherche. Il peut être question de la protection de la propriété intellectuelle (droit d'auteur et droit *sui generis* des bases de données), de la biodiversité, de données à caractère personnel ou encore d'un secret (secret professionnel, secret lié à la défense nationale, secret d'affaires, etc.). Certaines données sectorielles (données de santé, données environnementales, données archéologiques, etc.) obéissent, en outre, à des régimes juridiques particuliers.

Vous avez produit des données dans le cadre de vos activités de recherche,
mais vous vous interrogez sur les modalités de partage et diffusion ?
Voici quelques questions pour vous guider :



Vous avez répondu NON à toutes ces questions ? Une fois achevées et validées scientifiquement, les données doivent être ouvertes. Le personnel d'appui de proximité peut vous accompagner dans la gestion et la préparation de la diffusion de vos données.

Données couvertes par la protection du patrimoine scientifique et technique de la Nation, par le secret défense…

Le Code du patrimoine et les dispositions relatives à l'accès aux archives publiques encadre la durée pendant laquelle ces données doivent rester confidentielles. À l'issue de cette période, les données peuvent être diffusées librement. Cela s'applique par ailleurs aux autres types d'exceptions.

Données produites par des laboratoires situés en zone à régime restrictif…

Ces dernières ne sont pas automatiquement exclues du principe d'ouverture par défaut : pour déterminer les données à garder confidentielles, il convient de se rapprocher des personnes habilitées à se prononcer sur les restrictions de diffusion comme notament le Fonctionnaire de Sécurité de Défense ou FSD. Ensuite, comme dans tout autre projet, il revient aux équipes d'identifier les données publiques et achevées qui peuvent être ouvertes.

Données couvertes par le secret professionnel…

Les brevets protègent les inventions et non les données sous-jacentes qui ont permis leur avènement. Néanmoins, avant l'obtention d'un brevet, il convient d'être vigilant à ce que la diffusion des données ne conduise pas à une révélation de l'invention. Une fois le titre de propriété intellectuelle obtenu, les données associées à l'invention peuvent être ouvertes, si rien ne s'y oppose par ailleurs.

Données à caractère personnel
Leur gestion doit être anticipée au maximum. En effet, pour pouvoir traiter, collecter, enregistrer, modifier, ou même transmettre des données à caractère personnel, il vous faudra contacter le délégué à la protection des données (DPD ou DPO - *Data Protection Officer*) de l'organisme d'appartenance de votre directeur d'unité pour une inscription au registre des traitements (conformément au Règlement général sur la protection des données - RGPD) ou pour une demande d'autorisation à la Commission nationale d'informatique et des libertés (CNIL).

Les données sensibles constituent une catégorie spécifique de données à caractère personnel qui fournissent par exemple des informations spécifiques sur l'origine ethnique, les opinions politiques, les convictions religieuses, la santé, la vie ou l'orientation sexuelle d'une personne. Elles peuvent aussi concerner des données génétiques ou biométriques, générées afin d'identifier une personne physique de manière unique. La collecte et le traitement des données sensibles sont en principe interdits, mais le RGPD prévoit des exceptions en faveur des activités de recherche.

Protocoles pour la diffusion des données à caractère personnel

Les données contenant des informations personnelles peuvent être rendues publiques après traitement : chiffrement anonymisation ou pseudonymisation, en fonction du niveau de confidentialité et de la nature des données traitées. Ce niveau de confidentialité sera à déterminer en collaboration avec le DPO de votre structure et des objectifs du projet de recherche.

- L'anonymisation rend l'identification d'une personne impossible de manière définitive. ▼ **Amnesia** est un outil qui vous permet d'anonymiser vos jeux des données.
- La pseudonymisation empêche d'identifier une personne si l'on n'utilise pas de données tierces. Contrairement à l'anonymisation, la pseudonymisation est réversible. Elle consiste à substituer des données identifiantes (nom, prénom, ...) par des données indirectement identifiantes (alias, numéro, ...).

Des données réellement anonymisées ne constituent plus des informations à caractère personnel et peuvent donc être ouvertes, à condition de pouvoir établir que la ré-identification, même indirecte, des personnes n'est plus possible. La pseudonymisation constitue une mesure de protection des personnes, mais les données restent soumises à la réglementation sur les données personnelles et ne peuvent donc pas être ouvertes.

Les licences

Lorsqu'on publie des données, il est fortement recommandé d'y apposer une licence afin de définir comment celles-ci peuvent être réutilisées et modifiées.

En France, un décret liste les licences que les administrations peuvent utiliser pour diffuser des données publiques. Il s'agit de de la licence ▼ **Etalab** qui apportent la sécurité juridique nécessaire aux producteurs et aux ré-utilisateurs des données concernées, autorisant leur reproduction, redistribution, adaptation et exploitation commerciale tout en rendant obligatoire la mention de leur paternité.

En complément de la licence Etalab, il est recommandé d'ajouter les licences ▼ ***Creative Commons***. Elles permettent de personnaliser le degré d'ouverture souhaité et avec la licence CC-BY de créditer les producteurs des jeux de données.

Une liste de licences est généralement proposée par l'entrepôt de données qui va préserver et diffuser les données déposées.

Questions pratiques

Les contributeurs dans l'équipe

Tout au long du cycle de vie des données, différents acteurs contribuent à leur ouverture : le chercheur qui s'interroge sur les données à ouvrir lors de la rédaction de son plan de gestion de données, les professionnels d'appui à la recherche qui les accompagnent aux différentes étapes de la gestion des données, le délégué à la protection des données qui conseille le chercheur sur les conditions d'ouverture des données personnelles, le responsable scientifique du projet qui dépose des données dans un entrepôt en vue de leur réutilisation, les éditeurs qui proposent de publier des *data papers*…

Pour créditer au mieux les différents contributeurs lors de la diffusion des résultats, vous pouvez consulter ▼**CRediT**, une taxonomie identifiant jusqu'à quatorze rôles au sein d'un projet de recherche.

Le choix d'un entrepôt de données

Le choix de l'entrepôt est déterminant car les entrepôts sont plus ou moins compatibles avec les principes FAIR. Pour qu'une donnée soit facilement accessible, il faut qu'elle soit mise à disposition dans un entrepôt de données. Pour qu'une donnée soit facilement trouvable, il faut qu'elle soit également référencée dans des catalogues ou plateformes d'accès par le biais d'un identifiant pérenne.

Lors de la publication d'un jeu de données, l'entrepôt lui attribue un identifiant unique et pérenne. Plus une donnée sera décrite à travers des métadonnées riches et détaillées (titre, producteurs, date, résumé, format, identifiant pérenne, conditions d'accès et d'utilisation, métadonnées géographiques, temporelles, etc.), mieux elle sera indexée et plus elle sera facile à trouver.

Pour répondre à un enjeu de qualité des données, certains entrepôts modèrent les données avant leur publication et proposent au déposant des pistes d'amélioration de la description des jeux de données, sur la base de critères définis dans un guide de curation.

Pour favoriser la visibilité, le partage et la réutilisation des données produites ou collectées dans le cadre de projets scientifiques, une offre variée d'entrepôts de données existe : thématiques ou disciplinaires, de confiance ou certifiés, institutionnels ou souverains, généralistes…

Les plateformes ▼**Cat OPIDoR**, ▼**re3data.org** et ▼**FAIRsharing.org** répertorient de nombreux entrepôts. Afin d'identifier l'entrepôt de données le plus adapté, il est utile de s'informer sur son modèle économique, ses fonctionnalités et ses caractéristiques afin de vérifier qu'il réponde à vos besoins scientifiques, documentaires et techniques (champ disciplinaire, type de données acceptées, limite de volume). Lors de la soumission d'un article, l'éditeur peut vous demander de lui communiquer les données associées à la publication afin de les diffuser ensuite. Retrouvez les bonnes pratiques dans le guide ▼**Partager les données liées aux publications**.

BON À SAVOIR

Certains entrepôts offrent la possibilité de publier des données sous embargo afin que seules les métadonnées soient publiques dans un premier temps. Cela permet de signaler et de citer les données, tout en ne donnant pas accès aux fichiers eux-mêmes qui ne sont alors ni consultables publiquement, ni téléchargeables. Il reste néanmoins possible au déposant de donner accès aux fichiers des données sous embargo à des personnes identifiées. L'entrepôt de données de Recherche Data Gouv vous permet de définir les droits d'accès. Par exemple : lors de la soumission d'un article : lorsque la revue vous demande vos données pour les évaluateurs, vous pouvez, en les déposant dans l'entrepôt de Recherche Data Gouv, donner un accès temporaire et ciblé aux fichiers des données. Ces fichiers ne seront ensuite ouverts que lorsque vous le déciderez après la publication de votre article.

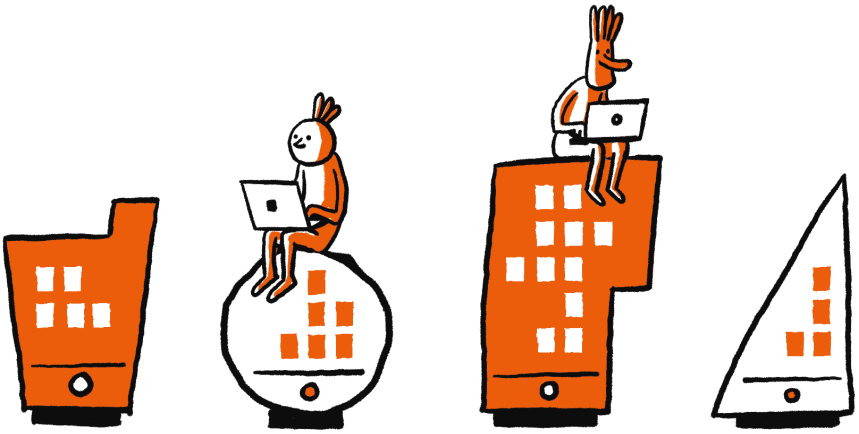
▼ **Science Europe** via le guide ▼ ***Criteria for the Selection of Trustworthy Repositories***, indique qu'un entrepôt de confiance doit répondre aux quatre critères suivants :

- attribuer systématiquement un identifiant pérenne à une donnée et/ou à un jeu de données,
- proposer un cadre de description des données normalisé et ouvert,
- définir les conditions d'accès et le cadre de réutilisation par l'attribution de licences,
- garantir un niveau de préservation et d'accessibilité sur le long terme aux données et métadonnées par la mise en place d'une politique et d'une gouvernance dédiées.

Des questions complémentaires doivent guider votre choix vers un entrepôt de confiance :

- Existe-t-il un entrepôt utilisé par vos pairs au sein de votre domaine de recherche ?
- L'entrepôt répond-il aux enjeux de la politique nationale de science ouverte et à la ligne directrice des principes FAIR ?
- Fournit-il un identifiant pérenne aux données ?
- Quelle est la durée de conservation ?
- Quel est le type de modération ?
- Propose-t-il la possibilité d'un embargo ?
- Est-il recommandé par les agences de financement ?
- L'entrepôt est-il certifié ?

Pour les communautés ne disposant pas d'un entrepôt thématique, l'entrepôt de données de ▼ **Recherche Data Gouv** offre un service pluridisciplinaire souverain, de confiance, pour la publication des jeux de données. Il dote votre jeu de données d'un identifiant et d'une licence qui vous permettra d'être cité.



BON À SAVOIR

Des certifications permettent de labelliser les entrepôts pour une durée déterminée, sur la base d'une liste de critères établis par des organismes reconnus. Jusqu'en 2023, seuls trois entrepôts français avaient reçu une certification du ▼***CoreTrustSeal*** (CTS) : le ▼**Centre de Données Astronomiques de Strasbourg (CDS)**, l'▼**IFREMER-SISMER** et l'▼**Institut d'astrophysique spatiale** (IDOC). Néanmoins de nombreux entrepôts sont dits « de confiance » lorsqu'ils fournissent un certain niveau de services et de garanties. La certification CTS ou non d'un entrepôt ne doit donc pas être un critère bloquant pour le dépôt de ses données.

ATTENTION !

Le dépôt des données dans un entrepôt ne signifie pas que vos données sont conservées à très long terme. En effet, il faut distinguer les notions de stockage, de sauvegarde et d'archivage pérenne.

Le stockage est une étape élémentaire de dépôt sur un support numérique pendant la durée du projet, alors que la sauvegarde a pour objectif de dupliquer la donnée sur différents supports numériques.

L'archivage est un processus qui permet, à la fin du projet, une conservation à très long terme des données sélectionnées. Les entrepôts de données diffusent des données mais seulement quelques-uns proposent un archivage des données en partenariat avec des services d'archives, comme ▼**Quetelet-Progedo**.

SUR LE TERRAIN

JULIETTE G.

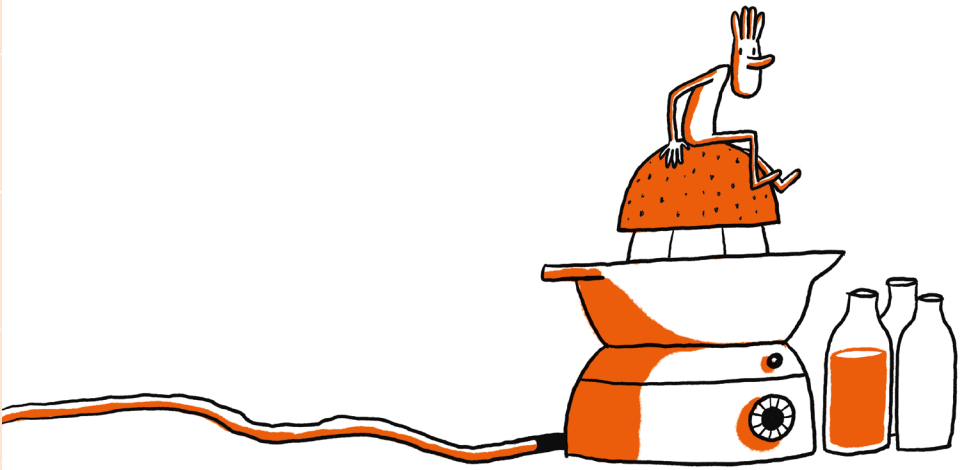
Doctorante en hydrologie
à l'Université Gustave-Eiffel

Je travaille sur le risque inondation en France, particulièrement sur l'Arc méditerranéen. Le but de ma thèse est d'appliquer et d'évaluer un modèle qui pourrait prévoir les impacts des crues soudaines.

Dans le cadre de ma thèse, j'ai dû suivre une formation sur les enjeux de la science ouverte. Avant cela, je n'y étais pas très sensibilisée, mais maintenant cela m'apparaît comme une évidence. C'est justement notre travail de chercheur de rendre disponible nos résultats pour toute personne qui en aurait le besoin, et d'être totalement transparent en fournissant toutes les données qui nous ont permis de tirer nos conclusions.

J'ai eu l'opportunité de publier un premier article dans une revue scientifique en open access, et j'ai publié en parallèle le jeu de données sur la plateforme ▼**Recherche Data Gouv**. Mon article est actuellement dans le processus de revue par les pairs, les reviewers pourront ainsi jeter un œil à mon jeu de données pour s'assurer de la fiabilité de mon travail.

Dans la mesure du possible, j'ai décidé de toujours procéder ainsi : dès que je publierai des résultats, ce sera toujours en open access, et je fournirai les codes et les données que j'ai utilisés. Je suis convaincue qu'une véritable prise de conscience est en train de s'opérer sur la science ouverte, et je ferai de mon mieux pour accompagner le mouvement.



CHEDID S.

Doctorant en Ingénierie géotechnique à l'Université de Nantes et à l'Université Gustave-Eiffel

Je travaille au sein de l'équipe de la centrifugeuse géotechnique (Lab. GERS-CG), où j'aborde deux aspects de recherche : expérimental et numérique. Pendant mon projet doctoral, après avoir publié les résultats dans une revue internationale, plusieurs chercheurs m'ont contacté pour obtenir l'accès aux données expérimentales.

Conscient de l'importance des résultats des essais en centrifugeuse pour les travaux numériques et après l'accord de mes encadrants, j'ai décidé de publier les données sur une plateforme ouverte au grand public ▼ **Recherche Data Gouv**, avec un accès facile et gratuit. Cette diffusion des données va permettre à d'autres chercheurs d'utiliser mes données, évitant ainsi la collecte de nouvelles données à partir de zéro et permettant d'autres analyses approfondies. De plus, cette démarche avec la science ouverte contribuera à renforcer la transparence et la diffusion de mon travail et à accroître la confiance du public dans les résultats obtenus.

D'un point de vue personnel, la diffusion des données m'a permis de rendre mon travail plus visible et d'obtenir une plus grande reconnaissance au sein de la communauté scientifique. Cela va contribuer à accroître ma réputation et celle de l'Université Gustave-Eiffel (Lab. GERS-CG) et mon impact en tant que jeune chercheur.

ET APRÈS ? PRÉPARER L'AVENIR

Valoriser ses données

Parallèlement au dépôt des données dans un entrepôt, vous pouvez choisir de valoriser vos données de la recherche dans un *data paper*. Un *data paper* est un article décrivant un jeu de données original, à fort enjeu de réutilisation. Il comporte la description fine du jeu de données (contexte de production, producteurs, droits associés, ...) et un accès au jeu de données, souvent sous la forme d'un lien pérenne vers l'entrepôt de données.

Les *data papers* suivent le même processus éditorial et d'évaluation que les articles scientifiques classiques. Il existe différentes revues publiant des *data papers*. Elles peuvent être multidisciplinaires, disciplinaires ou thématiques. Elles peuvent être spécialisées dans les *data papers* ou bien être des revues classiques proposant une section *data paper*. ▼**CoopIST** propose des clés pour comprendre comment se structure le contenu d'un *data paper*, comment choisir une revue pour le publier et comment l'évaluer. Vous trouverez différents critères et exemples de revues en fonction des disciplines.

Lier ses données au reste de ses travaux scientifiques

Grâce à un identifiant pérenne, la citation d'un jeu de données est facilitée et stabilisée puisque ce type d'identifiant pointera de manière univoque sur celui-ci. Dans une publication, les données associées, les auteurs et les contributeurs associés seront liés sans équivoque grâce aux identifiants pérennes de manière durable et stable, et ce quelle que soit la forme des informations enregistrées pour les décrire dans les différentes institutions.

▼ **DataCite** est une organisation non lucrative qui attribue les identifiants de jeux de données au niveau international. L'agence française d'attribution des DOI (*Digital Object Identifier*) pour les données, DataCite France est porté par l'Inist-CNRS. La fourniture de certains identifiants pérennes est adossée à des services complémentaires, comme la mise en forme automatique de citations, ce que permettent les DOI. Le DOI est attribué automatiquement par l'entrepôt dans lequel les données vont être déposées.

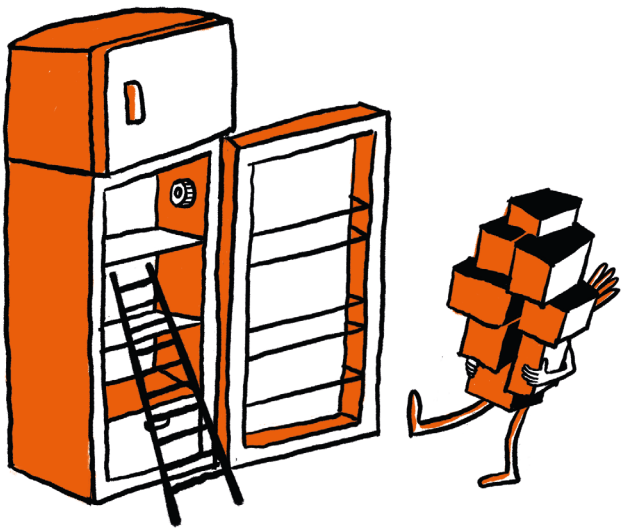
Identifier les différentes versions d'un jeu de données

Diffuser un jeu de données est une première étape. Ce jeu peut être complété s'il évolue dans le temps. La plupart des entrepôts permettent le versionnage des jeux de données pour marquer cette évolution.

Archiver de manière pérenne

Il s'agit de conserver, de garantir l'accès et de préserver l'intelligibilité sur le très long terme (au-delà de 30 ans) de données qui ont été préalablement sélectionnées. Cette sélection est régie par une législation et une réglementation spécifiques liées à des enjeux d'accessibilité, de lisibilité et de constitution du patrimoine scientifique. En effet, toutes les données n'ont pas vocation à être archivées sur le long terme, pour des raisons de coût et de préservation de l'environnement notamment. Les équipes d'appui à la gestion des données et en particulier les archivistes peuvent vous conseiller à ce sujet.

En France, le ▼**CINES** (Centre informatique national de l'enseignement supérieur) se charge de la conservation pérenne et propose plusieurs solutions d'archivage. Tout projet d'archivage doit être analysé en amont avec le service des archives de l'établissement avant de s'adresser au CINES : quelles données ont une valeur scientifique reconnue pour justifier leur archivage sur le long terme ? Quelle est leur volumétrie ? Quel est leur format ? Quelle durée de conservation est souhaitée ? Quel budget doit être alloué et pour quelle durée d'archivage ? Quel est l'impact environnemental de la conservation pérenne de mes données ?



SUR LE TERRAIN

NAOMI T.

Maîtresse de conférences en linguistique et germaniste à l'Université de Leiden

Presque toutes mes données, articles et réflexions sont aujourd'hui en accès ouvert, mais cela n'a pas toujours été le cas. Il est important de le dire : pratiquer la science ouverte est un processus, et on n'a pas besoin de tout rendre public tout de suite !

De mon côté, cela a commencé par mes corpus annotés. Les données annotées sont issues de débats parlementaires français, allemands et britanniques. Mon projet a ainsi consisté en la mise en valeur de transcriptions de débats parlementaires en France, en Allemagne et au Royaume-Uni en format XML-TEI sous licence CC-BY 4.0 afin de faciliter leur diffusion et réutilisation le plus rapidement et le plus largement possible.

J'ai publié ces données en accès ouvert sur ORTOLANG (Outils et Ressources pour un Traitement Optimisé de la LANGue) dès le début de ma thèse et avant même d'avoir publié mes résultats.

La réutilisation des données parlementaires est un enjeu démocratique de taille : alors que les transcriptions de séances parlementaires sont toutes disponibles sur les sites respectifs des parlements, l'exploitation des données à des fins de recherche reste compliquée.

Cette démarche m'a permis de valoriser plus largement les résultats de ma recherche. Deux *data papers* décrivent le processus afin de le rendre transparent et reproductible. Je suis très heureuse d'avoir été lauréate du prix Science ouverte des données de la recherche dans la catégorie « réutilisation des données » grâce à ce travail.

Si je n'avais qu'un seul conseil : lancez-vous !

POUR ALLER PLUS LOIN

RESSOURCES GÉNÉRALES

Collection *Passeport pour la science ouverte*.
https://www.ouvrirlascience.fr/passeport-pour-la-science-ouverte/?menu=3

▼ **Ouvrir la science** : ressources.
https://www.ouvrirlascience.fr/category/ressources/

PLATEFORMES D'INTÉRÊT

▼ **DoRANum** propose des ressources pour accompagner la communauté scientifique dans la gestion et le partage des données. Vous y trouvez des contenus d'autoformation par thématiques (enjeux, dépôt, plan de gestion, métadonnées…) ou par disciplines : https://doranum.fr

Le Groupe de travail Science ouverte - Données de ▼ **Couperin**
https://gtso.couperin.org/groupe-donnees/

▼ **Recherche Data Gouv** propose des guides, classes virtuelles et tutoriels : https://recherche.data.gouv.fr/fr/aide-en-ligne

▼ **CoopIST** du CIRAD : https://coop-ist.cirad.fr/gerer-des-donnees

▼ **Réseau des URFIST** propose de formations sur place ou à distance : https://sygefor.reseau-urfist.fr/#/

▼ **EcoInfo –** groupement de services du CNRS sur la réduction des impacts environnementaux et sociétaux négatifs des technologies du numérique : https://ecoinfo.cnrs.fr/

▼ **Labos 1point5 –** collectif de membres du monde académique, de toutes disciplines et sur tout le territoire, partageant un objectif commun, celui de « mieux comprendre et réduire l'impact des activités de recherche scientifique sur l'environnement, en particulier sur le climat » : https://labos1point5.org/

GUIDES PRATIQUES

Partager les données liées aux publications scientifiques. Guide pour les chercheurs (2022). Ministère de l'Enseignement supérieur et de la recherche, Comité pour la science ouverte. https://www.ouvrirlascience.fr/partager-les-donnees-liees-aux-publications-scientifiques-guide-pour-les-chercheurs/

Guide d'application de la loi pour une République numérique pour les données de la recherche (2022). Ministère de l'Enseignement supérieur et de la recherche, Comité pour la science ouverte. https://www.ouvrirlascience.fr/guide-dapplication-de-la-loi-pour-une-republique-numerique-pour-les-donnees-de-la-recherche/

Guide de bonnes pratiques sur la gestion des données de la recherche (2023). Atelier Données, groupe de travail inter réseaux de la MITI – CNRS. https://mi-gt-donnees.pages.math.unistra.fr/guide/00-introduction.html

Faire entrer la science ouverte dans son projet ANR : un guide pratique (2023). Groupe de travail Science ouverte - Données de Couperin. https://doi.org/10.5281/zenodo.7657818

JEUX SÉRIEUX

Dép'Osez : un jeu sur les entrepôts de données. Inist-CNRS, DoRANum (2023). DOI : 10.13143/AABD-HS57

GopenDoRe : un jeu sur la gestion des données de la recherche. Inist-CNRS, DoRANum (2022). DOI :10.13143/91td-qe92

RESSOURCES THÉMATIQUES

Droit

• Robin, A. (2022). *Droit des données de la recherche-Science ouverte, innovation, données publiques*. Larcier.

• Documentation RGPD de la CNIL, en particulier pour les données sensibles : https://www.cnil.fr/fr/definition/donnee-sensible

• Liste des licences utilisables pour les données, codes et logiciels de la recherche. https://www.data.gouv.fr/fr/pages/legal/licences/

Sobriété numérique

• Référentiel GreenDate pour un impact maîtrisé des données ouvertes. OpendataFrance, version beta : https://opendatafrance.gitbook.io/greendata-pour-un-impact-maitrise-des-donnees/greendata/1.1-contexte

• Publications d'EcoInfo : https://cnrs.hal.science/ECOINFO/browse/last

SOURCES

Bouchet-Moneret, F., *Les données personnelles de recherche et le RGPD* (2021). https://hal.univ-lorraine.fr/hal-03636697

Colavizza, G., Hrynaskiewicz, I., Staden, I.a, *et al.*, *The citation advantage of linking publications to research data* (2020). https://doi.org/10.1371/journal.pone.0230416

Code des relations entre le public et l'administration (CRPA) : applicable à compter du 1er janvier 2016. Légifrance (legifrance.gouv.fr)

Article L112-1 du Code de la recherche. Légifrance (legifrance.gouv.fr)

Article L. 533-4 du Code de la recherche. Légifrance (legifrance.gouv.fr)

Cost of not having FAIR research data - Cost-Benefit analysis for FAIR research data (2018). http://publications.europa.eu/resource/cellar/d375368c-1a0a-11e9-8d04-01aa75ed71a1.0001.01/DOC_1

Confederation of Open Access Repositories. *COAR Community Framework for Best Practices in Repositories* (2020). https://doi.org/10.5281/zenodo.4110829

Cotte, E., Sébire, F., *Modèle de logigramme de l'Institut Pasteur relatif aux questions juridiques liées à la diffusion des données de la recherche* (2022). https://pasteur.hal.science/pasteur-03587216

Dedieu, L., *Revues publiant des data papers* (2022). https://collaboratif.cirad.fr/alfresco/s/d/workspace/SpacesStore/4fdec919-30a2-46f1-8acb-2f0fa28fe5f8?attach=true

Données de la recherche – contexte juridique : qui a les droits, quelles obligations ? Doi : 10.13143/8dh5-d615
https://doranum.fr/aspects-juridiques-ethiques/qui-a-les-droits-quelles-obligations_10_13143_8dh5-d615/

Gaillard, R., *De l'open data à l'open research data : quelle(s) politique(s) pour les données de la recherche* (2014). https://www.enssib.fr/bibliotheque-numerique/documents/64131-de-l-open-data-a-l-open-research-data-quelles-politiques-pour-les-donnees-de-recherche.pdf

Hadrossek, C., Janik, J., Libes, M., Louvet, V., Quidoz, M-C., *et al.*, *Guide de bonnes pratiques sur la gestion des données de la Recherche* (2023). https://hal.science/hal-03152732

Inist-CNRS, DoRANum. *Métadonnées, standards, formats : fiche synthétique* (2017). https://doranum.fr/metadonnees-standards-formats/metadonnees-standards-formats-fiche-synthetique_10_13143_vbjs-6288/

Inist-CNRS, DoRANum. *Les principes FAIR* (2019). https://doranum.fr/enjeux-benefices/principes-fair_10_13143_z7s6-ed26/

Ministère de l'Enseignement supérieur et de la Recherche. *Feuille de route 2021-2024 sur la politique des données, des algorithmes et des codes sources*. https://www.enseignementsup-recherche.gouv.fr/fr/la-feuille-de-route-2021-2024-du-mesri-sur-la-politique-des-donnees-des-algorithmes-et-des-codes-50534

Ministère de l'Enseignement supérieur et de la Recherche. *Guide pratique pour une harmonisation internationale de la gestion des données de recherche* (2019). https://www.ouvrirlascience.fr/science-europe-guide-pratique-pour-une-harmonisation-internationale-de-la-gestion-des-donnees-de-recherche/

Organisation de coopération et de développement économique (OCDE), *Principes et lignes directrices de l'OCDE pour l'accès aux données de la recherche financée sur fonds publics*, Paris, 2007, http://www.oecd.org/fr/science/sci-tech/38500823.pdf

Philippe, O., Rennes, S., Szabo, D., Martel, A-S., *Ouverture des données : … Aussi ouvert que possible … aussi fermé que nécessaire* (2022). https://hal.inrae.fr/hal-03659484

Science Europe. « *Criteria for the Selection of Trustworthy Repositories* ». [consulté le 10/07/203]. https://view.officeapps.live.com/op/view.aspx?src=https%3A%2F%2Fwww.scienceeurope.org%2Fmedia%2Ffffkb51ei%2Fse-rdm-template-2-criteria-for-the-selection-of-trustworthy-repositories.docx&wdOrigin=BROWSELINK

Scripps CO2 Program: Carbon Dioxide Measurements. [consulté le 10/07/2023] https://scrippsco2.ucsd.edu/

Soulimane, G., Bouchiha, D., Benslimane, M.S., *La ré-ingénierie des ontologies : État de l'art*. Conférence internationale sur l'informatique et ses applications, CIIA'2006, May 2006, Saida, Algérie. https://hal.science/hal-01585047

Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.*, *The FAIR Guiding Principles for scientific data management and stewardship*. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

GLOSSAIRE

Catalogue de (méta)données : inventaire des données ou métadonnées destiné à les retrouver.
Curation (dans le cas du dépôt d'un jeu de données) : la curation scientifique consiste à nettoyer, éditorialiser, transformer dans l'objectif d'obtenir des jeux de données « propres », lisibles et plus faciles à traiter. Il existe aussi la curation documentaire et technique qui consiste à vérifier des métadonnées de fichiers de données à déposer dans un entrepôt, dans le but de proposer des modifications et d'améliorer la qualité de description des jeux de données.

Données de la recherche : enregistrements factuels (chiffres, textes, images et sons) qui sont utilisés comme sources principales pour la recherche scientifique et sont reconnus par la communauté scientifique comme nécessaires pour la validation des résultats.

***Data paper* :** publication qui décrit un jeu de données scientifiques, notamment à l'aide d'informations structurées appelées métadonnées.

***Data Protection Officer (DPO)* :** personne chargée de la protection des données à caractère personnel au sein d'une organisation.

Données à caractère personnel : données concernant une personne physique qui est identifiée ou identifiable, par exemple par corrélation avec d'autres jeux de données.

Embargo : période pendant laquelle les articles et les données de la recherche déposés sur une plateforme ne sont pas accessibles librement.

Entrepôt de données : service en ligne permettant le dépôt, la description, la recherche et la diffusion des jeux de données. Ils peuvent être pluridisciplinaires ou disciplinaires. Lorsqu'ils respectent une série de critères définis par le guide ▼ ***Criteria for the Selection of Trustworthy Repositories***, ils reçoivent le label de certification qui vise à promouvoir des entrepôts de données fiables et durables.

Identifiant pérenne ou Persistant Identifier (PID) : référence unique et stable pour un objet ou un sujet numérique (un jeu de données, un article, un auteur…). Exemple : *Digital Object Identifier* (DOI) ou l'identifiant auteur - *Open Researcher and Contributor* ID (ORCID).

Indexation : attribution à un document de termes distinctifs (des mots-clés par exemple) renseignant sur son contenu et permettant de le retrouver.

Interopérabilité : capacité de différents systèmes informatiques à dialoguer entre eux, à échanger des données, à communiquer sans ambiguïté et ainsi à interpréter des informations correctement.

Jeu de données ou **dataset :** agrégation, sous une forme lisible, de données brutes ou dérivées présentant une certaine unité, rassemblées pour former un ensemble cohérent.

Licence : texte juridique définissant les conditions de diffusion et de réutilisation d'une production (par exemple : licences logiciels libres, Creative Commons).

Loi pour une République numérique : loi de 2016 offrant un cadre juridique permettant de déposer en accès libre certaines versions des articles des revues, si les recherches sont financées pour moitié au moins sur fonds publics. De même, cette loi assimile les données de la recherche à des données publiques lorsque les travaux sont financés pour plus de la moitié par des fonds publics et traite le cas particulier de la recherche partenariale.

Métadonnées : ensemble d'informations structurées qui décrit, explicite, localise une ressource, dans le but d'en faciliter la recherche, l'usage et la gestion.

Principes FAIR : visent à rendre les données faciles à trouver, accessibles, interopérables et réutilisables.

Provenance : information sur les entités, les activités et les personnes impliquées dans la production d'une donnée ou d'un objet. Elle permet d'apprécier sa qualité, sa fiabilité ou sa crédibilité. (source : https://www.w3.org/TR/prov-overview/).

Réutilisation : utilisation par toute personne physique ou morale de données publiées à des fins autres que celles pour lesquelles elles ont été produites ou reçues.

Règlement général sur la protection des données (RGPD) : cadre juridique défini par l'Union européenne pour la gestion des données personnelles. Voir : https://www.cnil.fr/fr/comprendre-le-rgpd

Vocabulaire contrôlé : lexique raisonné et normalisé facilitant la recherche documentaire et l'analyse comparative de données (liste de mots-clés, glossaire, thésaurus, taxonomie, ontologie).

Web de données ou ***linked data* :** initiative du World Wide Web Consortium (W3C) visant à favoriser la publication de données structurées sur le web, non pas sous la forme de silos de données isolés les uns des autres, mais en les reliant entre elles pour constituer un réseau global d'informations.

Crédits

Direction de la publication
Ministère de l'Enseignement supérieur et de la Recherche

Coordination éditoriale
Université de Lille

Conseil scientifique
Collège Compétences et formation et collège Données de la recherche du Comité pour la science ouverte

Cheffe de projet
Mónica Michel Rodríguez

Rédacteurs
Florence Bouchet Moneret, Romane Coutanson, Amélie Fiocca, Céline Hernandez, Alicia León y Barella, Émilie Lerigoleur, Mónica Michel Rodríguez, Pierrette Paillassard, Agnès Robin, Sara Tandar, Laura Tomasso

Pour créer ce guide, le groupe de travail s'est appuyé sur les contenus pédagogiques des plateformes DoRANum, CoopIST, URFIST et COUPERIN.

Relecture éditoriale
Céline Barthonnat

Design graphique
Studio 4 minutes 34
Studio Lendroit.com

Impression
L'Artésienne, Liévin

Achevé d'imprimer :
Février 2024 à 10 000 exemplaires

Remerciements

Les jeunes chercheurs qui ont partagé leurs expériences de la science ouverte
Chedid Saade, Juliette Godet, Naomi Truan et Joshua Gobe

Les jeunes chercheurs qui ont participé aux échanges sur la première version
Marion Duthoit, Mathis Bachelot, Céline Barzun, Maxime Bedez, Paul Bellevile, Joan Bienaimé, Mélanie Bossu, Fabien Clouse, Violaine Courrier, Violette Delforge, Benjamin Demassieux, Meriam Meziani, Alexandre Van Outryve, Perrine Seguin

Experts consultés
Cécile Arènes, Alexis Arnaud, Flora Badin, Anne Baillot, Nathalie Barré-Lemaire, Laetitia Bracco, Marie Cros, Alina Danciu, Romain David, Delphine Du Pasquier, Emmanuelle Frenoux, Gaëlle Gauvrit, Candice Hector, Frédéric de Lamotte, Sylvie Le Bras, Gaëlle Leroux, Li Ling, Didier Mallarino, Gilles Mathieu, Lionel Maurel, Christelle Pierkot, Céline Rousselot.

Ce guide fait partie de la collection *Passeport pour la science ouverte*.
La version numérique de ce guide est disponible sur www.ouvrirlascience.fr.

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE
Liberté Égalité Fraternité

Ouvrir la science !

Université de Lille