Parcours interactif sur la gestion des données de recherche adapté aux domaines de l'environnement

Pour accéder à la ressource : https://doranum.fr/environnement/parcours-interactif-sur-la-gestion-des-donnees-de-recherche-adapte-aux-domaines-de-l-environnement_10_13143_53dt-h558/

Date de publication : 15/06/2021
Date de dernière mise à jour : 19/03/2024

Sommaire

BIENVENUE

Bienvenue dans ce parcours pédagogique adapté aux domaines de l'Environnement et composé de trois grandes parties :

- Les notions importantes à connaître
- Le contenu du Plan de gestion de données (PGD) ou Data management plan (DMP)
- Des exercices pour tester et valider vos connaissances.

Une quatrième partie vous propose un récapitulatif de ce qu'il faut retenir et une webographie.

Ce parcours a été pensé et conçu pour être suivi de façon linéaire et progressive, mais aussi de manière fragmentée. Vous pouvez consulter uniquement les parties qui vous intéressent sachant que des passerelles sont faites entre certaines parties du cours.

Un sommaire s'affiche à la gauche de votre écran et vous permet de naviguer à votre convenance. Il est accompagné d'un indicateur qui passe au vert au fur et à mesure que vous progressez dans les différentes parties.

NOTIONS IMPORTANTES

1. Données de la recherche et cycle de vie

1.1. Données de la recherche

1.1.1. Définitions

Il existe de nombreuses définitions des données de la recherche.

La définition de l'OCDE (Organisation de coopération et de développement économiques) est la plus communément retenue :

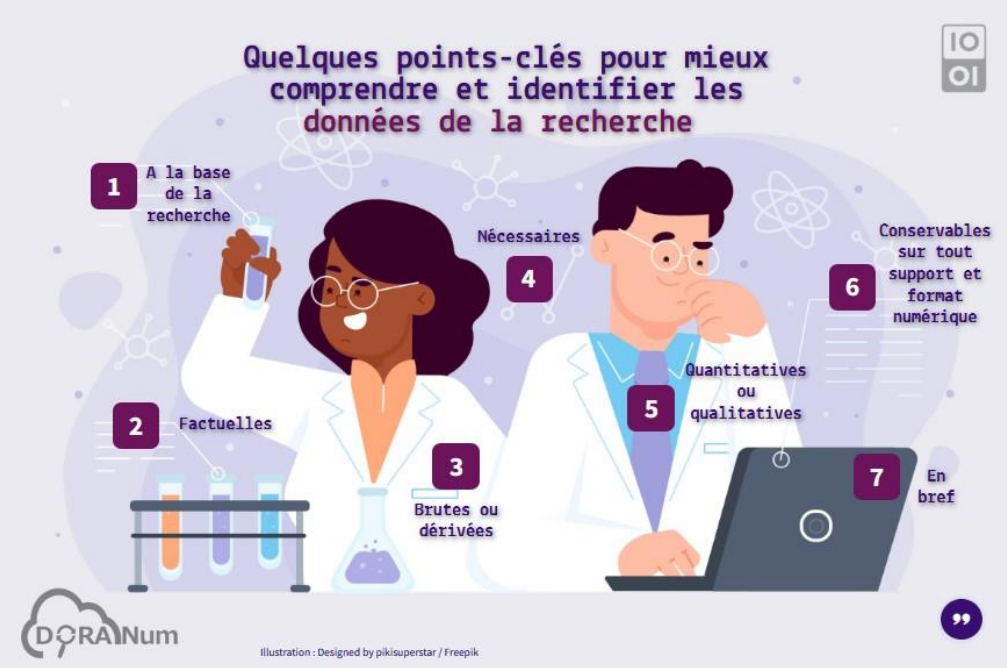
« Les données de la recherche sont définies comme des enregistrements factuels (chiffres, textes, images et sons), qui sont utilisés comme sources principales pour la recherche scientifique et sont généralement reconnus par la communauté scientifique comme nécessaires pour valider les résultats de recherche. »
OCDE, Organisation de Coopération et de Développement Économiques. Principes et lignes directrices de l'OCDE pour l'accès aux données de la recherche financée sur fonds publics. 2007. https://doi.org/10.1787/9789264034020-en-fr

À titre d'exemple, on peut également citer la définition issue de l'ARDC - Australian Research Data Commons (traduction Inist-CNRS) :

« Le terme de données de la recherche désigne les données sous forme de faits, d'observations, d'images, de résultats de programmes informatiques, d'enregistrements, de mesures ou d'expériences sur lesquels un argument, une théorie, un test ou une hypothèse, ou un autre produit de la recherche est basé [...]. Les données peuvent être numériques, descriptives, visuelles ou tactiles. Elles peuvent être brutes, nettoyées ou traitées, et peuvent être conservées dans tout format ou support [...]. »
ARDC, Australian Research Data Commons. What Is Research Data. 2019.

1.1.2. Diversité des données de la recherche

Cette ressource interactive vous permettra de mieux comprendre et identifier les données de la recherche :



https://doranum.fr/plan-gestion-donnees-dmp/definitions-des-donnees-de-la-recherche_10_13143_b4zc-8e79/

1) **À la base de la recherche**

   Pour faire son travail, un chercheur a besoin de **données de recherche**. Elles sont **à la base de son travail**. Sans données, aucun résultat de recherche ne peut être testé, ni vérifié.

2) **Factuelles**

   Les données de la recherche sont des **éléments factuels**. Par exemple, il peut s'agir de chiffres, de textes, d'images, de sons, de mesures, de résultats d'enregistrements, de résultats de programmes informatiques...

3) **Brutes ou dérivées**

   Les données de la recherche peuvent être **brutes**, c'est-à-dire qu'elles n'ont pas encore été traitées, interprétées ou modifiées.

Elles peuvent aussi être **dérivées de sources brutes**, c'est-à-dire qu'elles correspondent à des données brutes transformées (traitées, organisées, nettoyées...).

4) Nécessaires

Les données de la recherche peuvent être de **nature** très **différente selon la discipline scientifique**. En général, un chercheur peut identifier ses données de recherche en se basant sur celles qui sont reconnues dans sa discipline comme **nécessaires pour valider les résultats de la recherche**.

Par exemple :

- En archéologie, les données de la recherche peuvent être des photographies, des relevés topographiques, des mesures dendrochronologiques, un SIG, une datation radiocarbone.....
- En sociologie, il peut s'agir de séries statistiques, de résultats d'enquêtes, de sondages, d'images...
- En sciences de la Terre et de l'environnement, on peut utiliser des cartes géologiques, des photographies, des relevés de température, des calculs du taux de carbone présent dans la mer, des coordonnées GPS...

5) Quantitatives ou qualitatives

Les données de recherche peuvent être **quantitatives** (données quantifiables) et/ou **qualitatives** (faisant référence à une caractéristique non quantifiable).

6) Conservables sur tout support et format numérique

En théorie, les données de la recherche peuvent être conservées sur tout support (papier ou numérique) et tout format (.pdf, .txt, .png, .xml...).

Dans les faits, lorsque l'on se place dans le contexte de l'open science et de l'open data, **les données de recherche sont souvent numérisées ou numérisables** afin de rendre leur stockage et leur partage possible et plus efficace.

Au moment de choisir le format d'enregistrement de ses données, il est vivement recommandé de respecter le principe : "Aussi ouvert que possible, aussi fermé que nécessaire" et de **choisir des formats dits "ouverts"**, non-propriétaires.

Par exemple, pour un fichier audio, mieux vaut privilégier le format .wav au format .mp3.

En effet, un fichier audio conservé :

- Au format .MP3 ne pourra pas être lu par tous les logiciels capables de lire des fichiers audios : il s'agit d'un format fermé.
- Au format .WAV pourra être créé, lu ou modifié par tous les logiciels destinés à lire des fichiers de type audio : il s'agit d'un format ouvert.

7) **En bref**

Les données de la recherche sont donc tous les éléments factuels qu'un chercheur peut **produire**, **collecter** ou **réutiliser** et dont il a besoin pour **valider les résultats de sa recherche**.

Il est important de noter qu'il existe plusieurs définitions des données de la recherche. Elles restent une notion **difficile à définir** car elles englobent un panel diversifié de réalités. D'une discipline à une autre, ou d'un point de vue institutionnel à un autre, la définition des "données de la recherche" est plus ou moins restrictive.

Pour résumer, selon le projet, des données de la recherche peuvent être :

- **Produites ou recueillies** : ce sont les données créées, élaborées, générées lors d'activités de recherche (collectes sur le terrain, observations, mesures…).
- **Préexistantes** : ce sont des données déjà existantes (provenant de corpus, d'archives…) qui sont utilisées pour le projet. Les données utilisées peuvent avoir été recueillies initialement dans un autre contexte que celui de la recherche mais elles sont utilisées comme données de recherche dans le cadre du projet.

Selon leur contexte de création (capture ou production), leur exploitation, leur analyse et les traitements qu'elles subissent, les données de recherche peuvent être :

- **De différentes natures...**

  Brutes, dérivées, formatées, nettoyées, primaires, secondaires, traitées….

- **De tous types…**
  Archives, audios, vidéos, bases de données, codes sources, données géospatiales, images, photographies, langages de programmation, matérielles et physiques, modèles, visualisations, 3D, numériques, textuelles, numérisations, scans, qualitatives, quantitatives, statistiques…

- **Contenues dans divers supports…**
  Carnets de laboratoire, documents électroniques, logiciels, supports papier, programmes informatiques…

1.1.3. Des exemples dans les domaines de l'environnement

En fonction de leur provenance, les données peuvent être des :

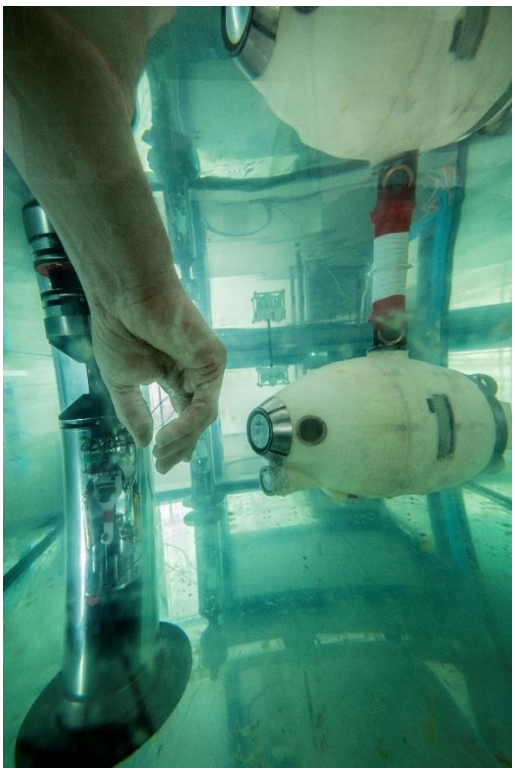
- **Données d'observation**

Les données d'observation sont généralement capturées en temps réel. Elles sont habituellement uniques et donc impossible à reproduire.
Exemples :
- " Les données brutes de biodiversité sont les données d'observation de taxons, d'habitats d'espèces ou d'habitats naturels, recueillies par observation directe, par bibliographie ou par acquisition de données auprès d'organismes détenant des données existantes ".
Code de l'environnement. Article L411-1 A. 25 février 2022.
https://www.legifrance.gouv.fr/codes/article_lc/LEGIARTI000033019166/

" Alouette des champs, dans la Zone Atelier Plaine et Val de Sèvre, au sud de Niort, en Nouvelle-Aquitaine.
Cette espèce a vu ses effectifs diminuer d'un tiers en moins de vingt ans dans cette zone. Cette observation s'inscrit dans le constat tiré de deux études de suivi des oiseaux, l'une menée à l'échelle nationale, l'autre plus localement : les oiseaux des campagnes françaises disparaissent de manière très rapide. En moyenne leurs populations se sont réduites d'un tiers en quinze ans. Cette observation massive réalisée à différentes échelles est concomitante à l'intensification des pratiques agricoles ces 25 dernières années, plus particulièrement depuis 2008-2009. "
Date de production : 25/04/2013
© Vincent BRETAGNOLLE / CEBC / CNRS Photothèque / Référence N° 20180034_0001

- **Données expérimentales**

Les données expérimentales sont obtenues à partir d'équipements de laboratoire. Elles sont souvent reproductibles mais cela peut s'avérer coûteux.
Exemple :



" Test d'un poisson-robot et de deux moules artificielles dans un bassin.
Ils appartiennent à une flotte créée pour observer la lagune de Venise, un milieu écologique extrêmement fragile, très sensible à l'activité humaine et aux changements climatiques. Fruits de la biorobotique, ces dispositifs sont inspirés de la nature. Ils

sont capables d'évoluer en autonomie et d'interagir entre eux grâce à une technologie inspirée du sens électrique de certains poissons, qui produisent et analysent un champ électrique pour s'orienter et communiquer en eaux troubles. Placées dans les fonds vaseux de la lagune, les moules artificielles collectent des données sur leur milieu (oxygène, température, turbidité, courants, etc.) qui seront véhiculées par les poissons-robots. L'objectif est de construire des cartes évolutives de la lagune, pour surveiller son évolution et agir en cas de détérioration. "

Date de production : 10/03/2020

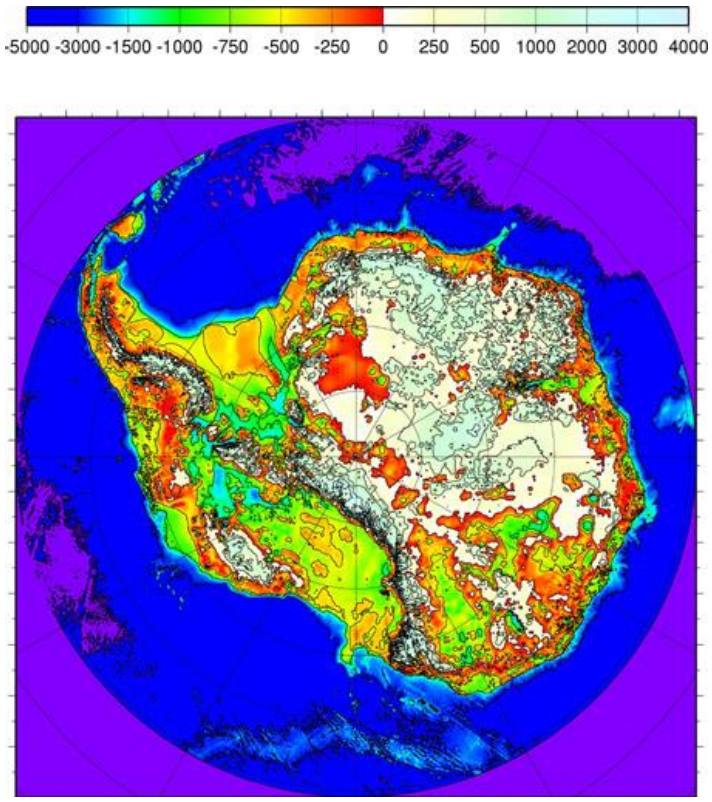
© Jean-Claude MOSCHETTI / LS2N / CNRS Photothèque / Référence N° 20200042_0010

- **Données computationnelles ou de simulation**

Les données computationnelles ou de simulation sont générées par des modèles informatiques ou de simulation. Elles sont souvent reproductibles à condition que le modèle soit correctement documenté.

On peut citer par exemple les modèles météorologiques, ceux de simulation sismique...

Exemple :



"Carte de l'altitude du socle antarctique.

La carte numérique provient de BEDMAP (https://www.bas.ac.uk/project/bedmap-2/).

Sur cette représentation, les couleurs ont été choisies pour mettre l'accent sur la

partie en dessous du niveau de la mer. En effet, la topographie de cette zone joue un grand rôle dans l'évolution de la calotte antarctique à l'échelle de temps du cycle glaciaire-interglaciaire. La carte numérique est utilisée comme donnée d'entrée du modèle d'évolution de l'Antarctique développé au LGCE. Elle permet de comprendre les conséquences des changements climatiques sur l'évolution des calottes glaciaires."

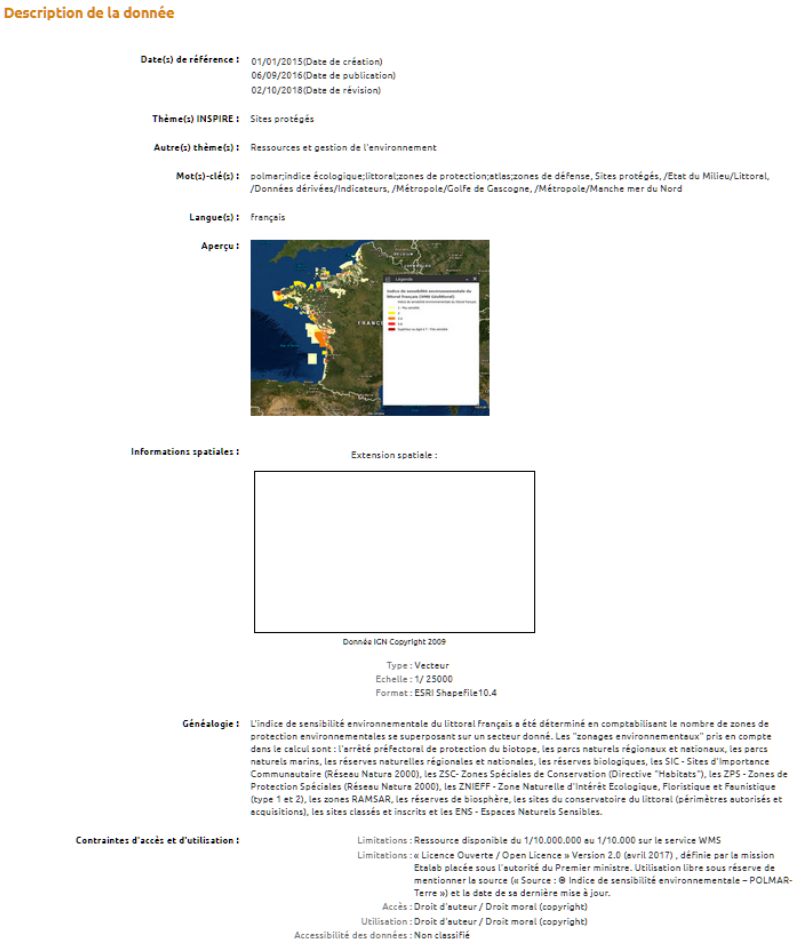
© CNRS photothèque / Référence N° 20020001_1486

- **Données dérivées ou compilées**

Les données dérivées ou compilées sont issues du traitement ou de la combinaison de données brutes. Elles sont souvent reproductibles mais coûteuses.

C'est le cas pour la fouille de texte ou les bases de données compilées.

Exemple :



Description de la donnée

Date(s) de référence : 01/01/2015(Date de création)
06/09/2016(Date de publication)
02/10/2018(Date de révision)

Thème(s) INSPIRE : Sites protégés

Autre(s) thème(s) : Ressources et gestion de l'environnement

Mot(s)-clé(s) : polmar;indice écologique;littoral;zones de protection;atlas;zones de défense, Sites protégés, /Etat du Milieu/Littoral, /Données dérivées/Indicateurs, /Métropole/Golfe de Gascogne, /Métropole/Manche mer du Nord

Langue(s) : français

Aperçu :

Informations spatiales : Extension spatiale :

Donnée IGN Copyright 2009

Type : Vecteur
Echelle : 1/ 25000
Format : ESRI Shapefile10.4

Généalogie : L'indice de sensibilité environnementale du littoral français a été déterminé en comptabilisant le nombre de zones de protection environnementales se superposant sur un secteur donné. Les "zonages environnementaux" pris en compte dans le calcul sont : l'arrêté préfectoral de protection du biotope, les parcs naturels régionaux et nationaux, les parcs naturels marins, les réserves naturelles régionales et nationales, les réserves biologiques, les SIC - Sites d'Importance Communautaire (Réseau Natura 2000), les ZSC- Zones Spéciales de Conservation (Directive "Habitats"), les ZPS - Zones de Protection Spéciales (Réseau Natura 2000), les ZNIEFF - Zone Naturelle d'Intérêt Ecologique, Floristique et Faunistique (type 1 et 2), les zones RAMSAR, les réserves de biosphère, les sites du conservatoire du littoral (périmètres autorisés et acquisitions), les sites classés et inscrits et les ENS - Espaces Naturels Sensibles.

Contraintes d'accès et d'utilization : Limitations : Ressource disponible du 1/10.000.000 au 1/10.000 sur le service WMS
Limitations : « Licence Ouverte / Open Licence » Version 2.0 (avril 2017) , définie par la mission Etalab placée sous l'autorité du Premier ministre. Utilisation libre sous réserve de mentionner la source (« Source : ® Indice de sensibilité environnementale – POLMAR-Terre ») et la date de sa dernière mise à jour.
Accès : Droit d'auteur / Droit moral (copyright)
Utilisation : Droit d'auteur / Droit moral (copyright)
Accessibilité des données : Non classifié

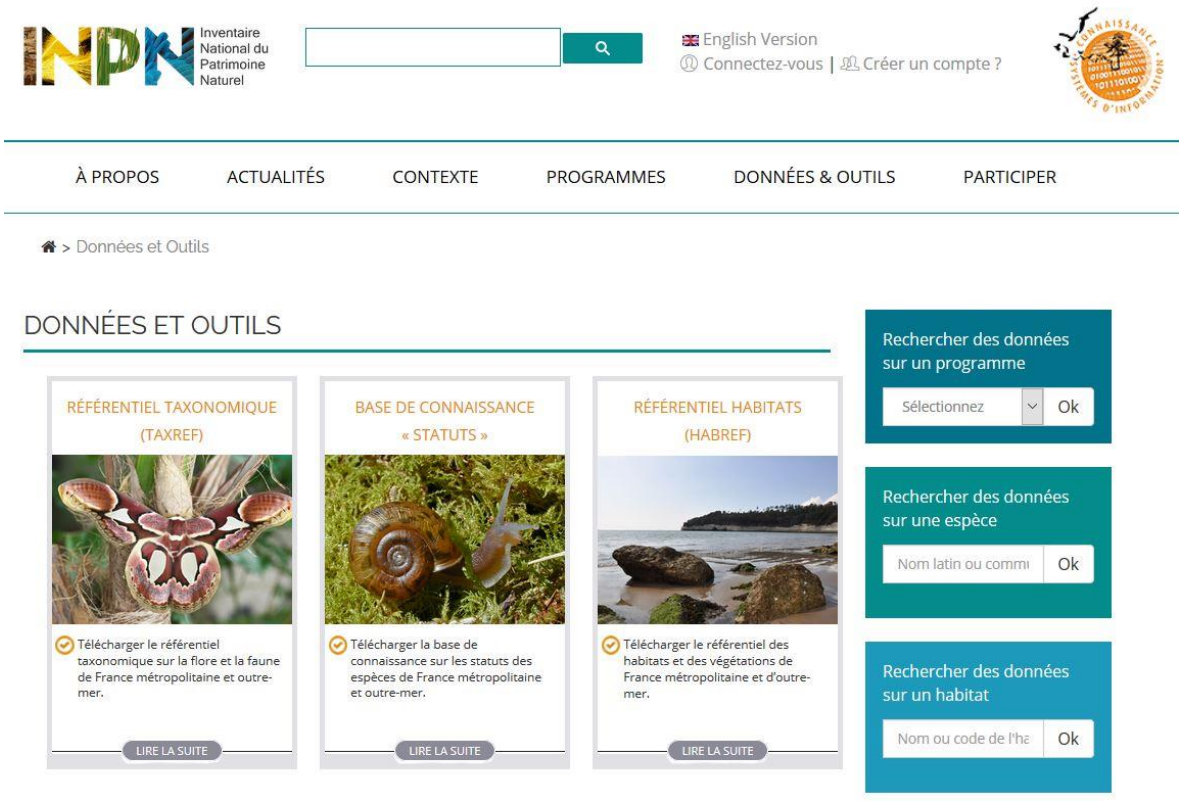
" Indice de sensibilité environnementale du littoral français.

L'Indice de sensibilité environnementale du littoral français a été approché comme étant une approximation de la "valeur patrimoniale de l'environnement". La méthode retenue pour déterminer cet indice a consisté à comptabiliser sur un secteur donné le

nombre de zones de protection environnementales se superposant, en partant du principe que plus un site va être couvert par des zonages différents plus il répond à un nombre d'enjeux importants et donc plus sa valeur patrimoniale est élevée et sa vulnérabilité à la pollution aux hydrocarbures est importante. Cette couche représente l'indice de sensibilité environnementale du littoral français réparti selon 5 classes d'indice allant du peu sensible au très sensible : 1- peu sensible, 2, 3-4, 5-6 et supérieure ou égale à 7 - très sensible. "
http://www.geocatalogue.fr/Detail.do?fileIdentifier=399909ce-8ffd-4938-bb7c-602a3273caa3

- **Données de référence**

Collection ou accumulation de petits jeux de données qui ont été revus par les pairs, annotés et mis à disposition.
Exemple :



« L'Inventaire National du Patrimoine Naturel est le portail de la biodiversité française, de métropole et d'outre-mer. Il diffuse la connaissance sur les espèces animales et végétales, les milieux naturels, les espaces protégés et le patrimoine géologique. L'ensemble de ces données de référence, validées par des réseaux d'experts, sont mises à la disposition de tous, professionnels, amateurs et citoyens. »
https://inpn.mnhn.fr/accueil/donnees-referentiels

Il peut s'agir de :

- **Données textuelles**

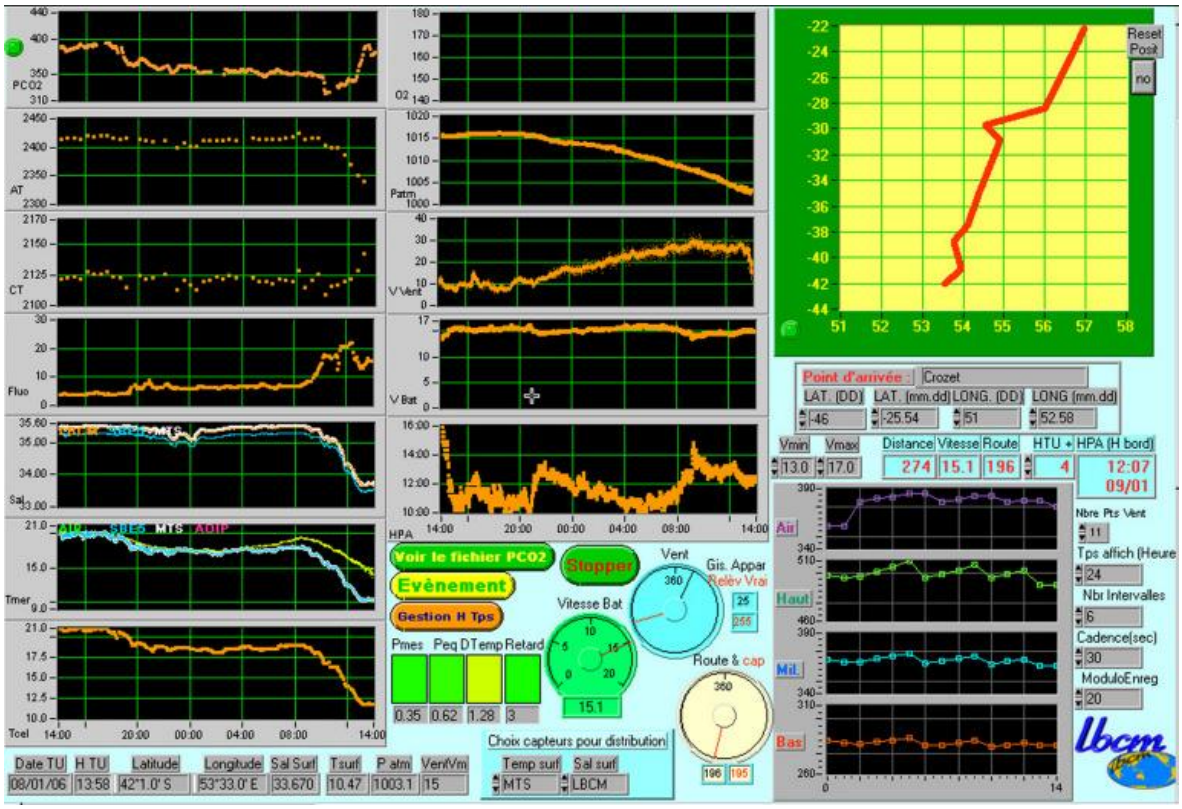
Notes de terrain ou de laboratoires, réponses d'enquête, corpus de textes…
Exemple :

" Prélèvement d'eau dans un des affluents du lac Titicaca, en Bolivie, pour tenter de retracer l'origine de la pollution au mercure des cours d'eau.
Une zone d'orpaillage importante se situe à l'amont de cet affluent et les scientifiques souhaitent voir si cette rivière est une source de mercure au lac. Berceau de la civilisation Inca, le lac Titicaca, plus grand lac d'eau douce d'Amérique latine, est aujourd'hui menacé d'asphyxie. Dans le cadre d'un projet ANR original "La Pachamama", une équipe internationale de scientifiques réalise le bilan écologique du lac et étudie, pour la première fois in situ, le transport et le devenir biochimiques du mercure et de ses isotopes. "
Date de production : 12/11/2014
© Erwan AMICE / LEMAR / IRD / CNRS Photothèque / Référence N° 20140001_2171

- **Données numériques**

Tableaux, comptes, mesures, latitude, longitude…

Exemple :



" Tableau de bord affichant les divers paramètres mesurés à bord du Marion Dufresne au cours des campagnes OISO (Océan Indien Service d'Observation).

Le but des campagnes de mesure OISO, menées à bord du Marion Dufresne dans le sud de l'océan Indien, est de mieux connaître la répartition des puits et des sources de gaz carbonique à la surface des océans. "

© Christian BRUNET/CNRS Photothèque / Référence N° : 20060001_0895

- **Données audiovisuelles**

Images, sons, vidéos…

Exemple :

https://phenocam.nau.edu/webcam/sites/arsltarmdcrresw/

Le réseau PhenoCam permet de fournir une télédétection automatisée et proche de la surface de la phénologie du couvert végétal dans le nord-est des États-Unis et le Canada voisin. Des caméras numériques à haute résolution ("webcams") ont été installées sur plus d'une douzaine de sites de recherche établis et répartis dans cette région. Le réseau comprend actuellement des caméras situées dans toute l'Amérique du Nord, de l'Alaska à la Floride et d'Hawaï au Maine. Plus de 400 caméras, dont la plupart ont été déployées selon un protocole normalisé, téléchargent des images toutes les demi-heures sur le serveur PhenoCam. Des techniques d'analyse simples sont ensuite utilisées pour extraire des informations quantitatives sur les couleurs de chaque image. Les indices de verdissement du couvert végétal fournissent ainsi des informations sur la quantité de feuillage présent, et sa couleur. Sur de nombreux sites du réseau PhenoCam, les chercheurs effectuent des mesures continues de l'échange surface-atmosphère de carbone et d'eau en utilisant la méthode de covariance des tourbillons. Ces données de flux sont utilisées pour évaluer les implications des changements saisonniers de l'état de la canopée sur le fonctionnement des écosystèmes.

Nom du site : arsltarmdcrresw

Localisation : Restored Wetland, Caroline County, Maryland

Lat: 39.0549 Lon: -75.7532 Elev(m): 16

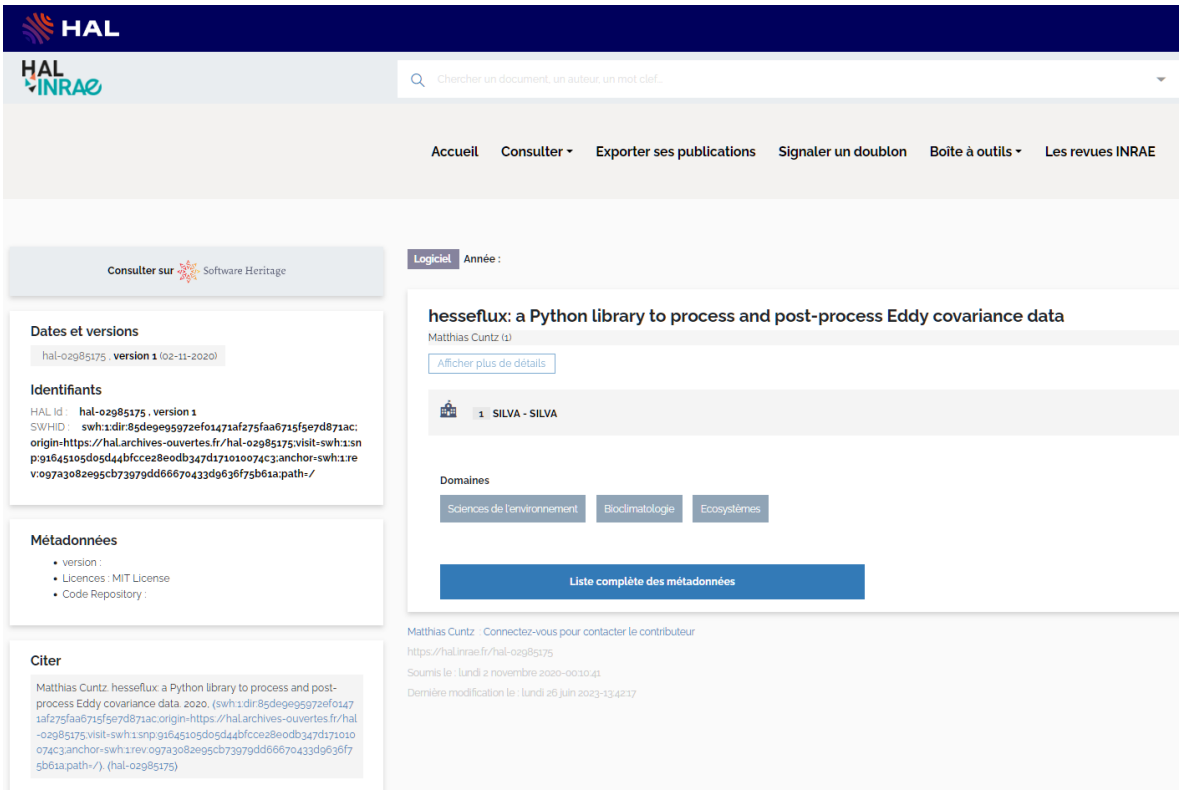
Image Count: 55757 Start Date: 2019-03-19 Last Date: 2023-07-17

Seyednasrollah Bijan, Young Adam M., Hufkens Koen, Milliman Tom, Friedl Mark A., Frolking Steve, Richardson Andrew D. Tracking vegetation phenology across diverse biomes using version 2.0 of the phenocam dataset. Scientific Data, 6(1):222. 2019.

https://doi.org/10.1038/s41597-019-0229-9

- **Logiciels (scripts…)**

Exemple :



Code source déposé dans HAL et dans Software Heritage. Il fournit des fonctions utilisées dans le traitement et le post-traitement des données de flux de covariance Eddy (technique utilisée pour mesurer et calculer les flux de CO2/H2O et d'autres flux de traces de gaz dans la couche limite atmosphérique) du site de l'écosystème ICOS FR-Hes.

Matthias Cuntz. hesseflux: a Python library to process and post-process Eddy covariance data. 2020. https://hal.inrae.fr/hal-02985175

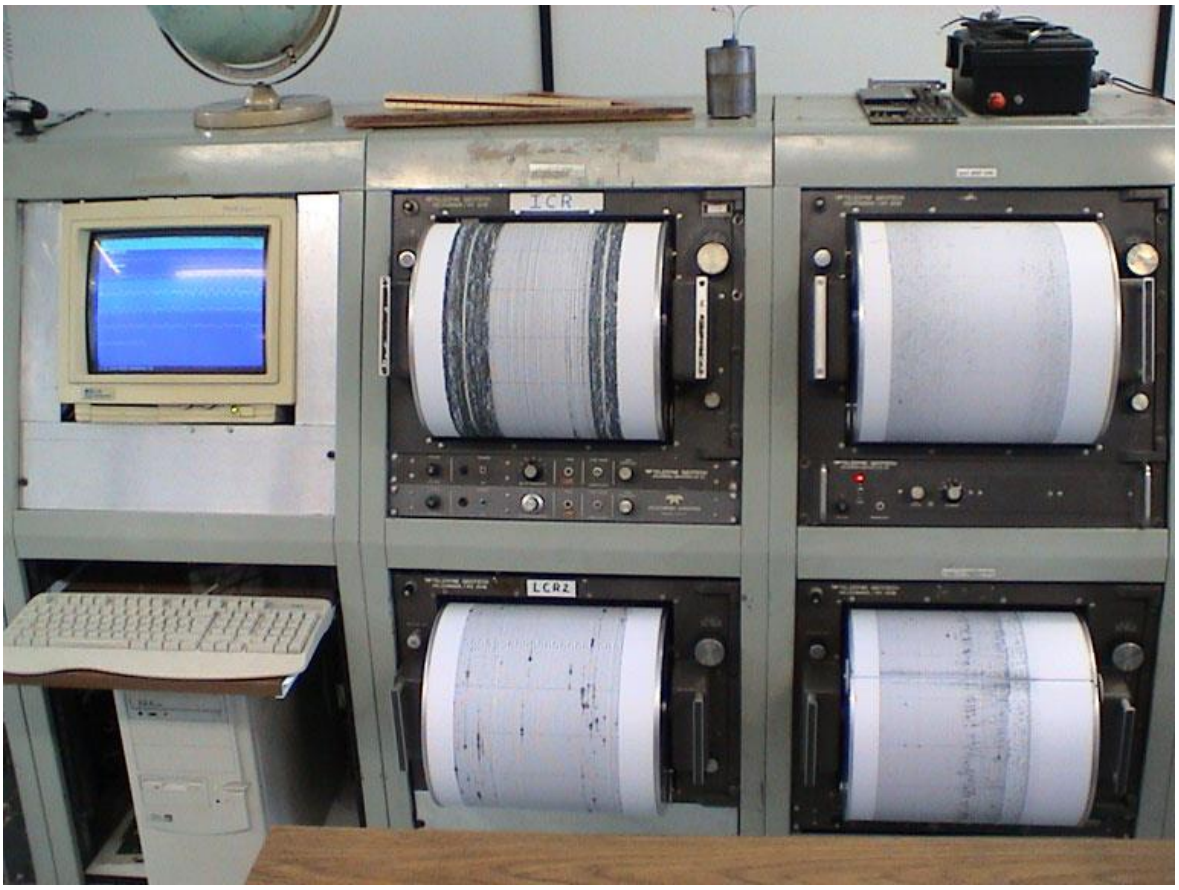
- **Données spécifiques liées à la discipline**

Exemple :

" Sonde multi-paramètres installée dans un cours d'eau de la Zone atelier Armorique, sur le site de Pleine-Fougères, en Ille-et-Vilaine.
Elle permet de suivre la température, la hauteur et la qualité de l'eau (nitrate, ammonium, pH, conductivité électrique). Six exutoires sont équipés de ce type de sonde, ce qui permet de suivre la qualité de l'eau en temps réel, sur six sous bassins versants. L'objectif est d'acquérir des données à haute fréquence temporelle, pour caractériser les processus impliqués dans le transfert des solutés en crue et hors crue. La ZA Armorique permet de mener des recherches sur les relations entre dynamiques sociétales et dynamiques environnementales, le long de gradients allant de l'urbain au rural, avec le paysage comme objet d'articulation. "
Date de production : 01/05/2013
© Cyril FRESILLON/CNRS Photothèque / Référence N° 20130001_0993

- **Données spécifiques produites par certains instruments**

Exemple :



" Sismogrammes (tracés des ondes sismiques obtenu grâce au sismographe) du Réseau National de Sismologie du Costa Rica pour la surveillance des volcans. "
© Franck DONNADIEU / CNRS Photothèque / Référence N° 20040001_1050

1.2. Jeux de données (dataset)

Les données de la recherche sont le plus souvent organisées, analysées et traitées par lots : c'est ce qu'on appelle des " jeux de données ".

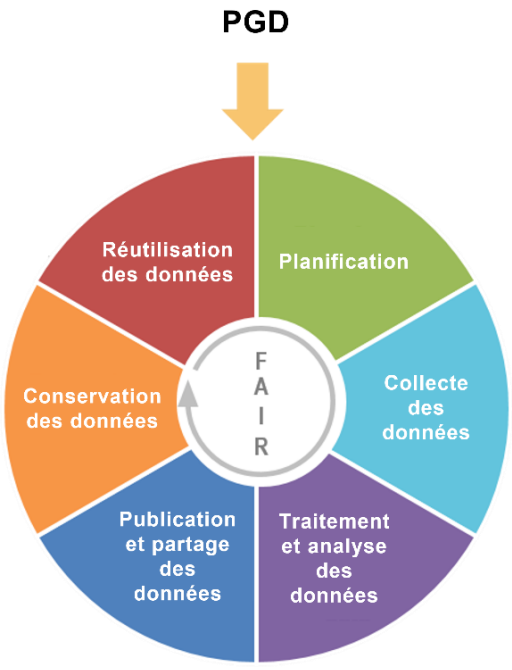
« La notion de « **jeu de données** » (dataset) peut être définie comme l'agrégation, sous une forme lisible, de données brutes ou dérivées présentant une certaine «unité», rassemblées pour former un ensemble cohérent.
Toutefois, l'échelle à laquelle ces données assemblées acquièrent leur unité pour former un « jeu » varie selon les disciplines, les types de données, les projets, les raisons pour lesquelles ces données sont agrégées. Sous l'angle spécifique de

« l'ouverture » des données de recherche, on peut définir le jeu de données comme un enregistrement de données sous la forme d'un ou plusieurs fichiers électroniques, téléchargeables, citables (notamment par l'intermédiaire d'un DOI) et intelligibles – ce jeu étant accompagné des métadonnées descriptives suffisantes. »
Gaillard Rémi. De l'open data à l'open research data : quelle(s) politique(s) pour les données de la recherche. Janvier 2014. http://www.enssib.fr/bibliotheque-numerique/documents/64131-de-l-open-data-a-l-open-research-data-quelles-politiques-pour-les-donnees-de-recherche.pdf

La rédaction d'un PGD (Plan de gestion de données) ou DMP (Data management plan) se fait sous forme de rubriques pour chaque jeu de données. Il peut également concerner les logiciels, les modèles de simulation...

1.3. Cycle de vie des données de la recherche

« Le cycle de vie des données de la recherche est l'ensemble des étapes de gestion, conservation, diffusion et réutilisation des données scientifiques, associées aux activités de recherche. »
Deboin Marie-Claude. Découvrir de nouveaux métiers liés aux données de la recherche. CIRAD. 5 p. 5 octobre 2018. https://doi.org/10.18167/coopist/0061



Le Data management plan (DMP) s'inscrit pleinement dans le cycle de vie des données d'un projet

Le PGD (Plan de gestion de données) ou DMP (Data management plan) prend en compte toutes les étapes du cycle de vie des données, durant le temps du projet de recherche et au-delà : la planification et la collecte des données, le traitement, l'analyse, le partage et la publication des données, la conservation à long terme et la réutilisation.

Un PGD peut être exigé si le projet est accepté pour financement et figurer parmi les livrables.

Il est initié très tôt car il aide à planifier et anticiper la gestion des données d'un projet. En général, la version initiale du PGD est à livrer dans les 6 mois qui suivent le début du projet.

1.4. Pourquoi gérer et partager ses données

- **Quantité** : face à l'accroissement de la quantité de données, la mise en place d'une bonne gestion est nécessaire, notamment pour éviter la perte de données et pour anticiper les moyens de stockage et de sauvegarde adaptés au volume de données.
- **Qualité** : partager ses données nécessite d'adopter de bonnes pratiques de gestion de données, ce qui améliore la qualité du travail de recherche.
- **Validation des résultats de recherche** : partager ses données permet de valider les résultats de recherche. De plus en plus d'éditeurs exigent que les chercheurs rendent accessibles toutes les données sous-jacentes aux résultats rapportés dans l'article soumis.
- **Intégrité** : rendre ses données disponibles offre une meilleure garantie contre la fraude scientifique.
- **Valorisation** : le partage des données permet au chercheur de valoriser ses données et d'accroître sa visibilité (citation).
- **Financement** : le partage des données (selon le principe " aussi ouvert que possible, aussi fermé que nécessaire ") peut être une condition pour l'obtention du financement du projet.
- **Reproductibilité et réutilisation** : le coût engendré par la création, la collecte et le traitement des données peut être très élevé. Réutiliser des données existantes

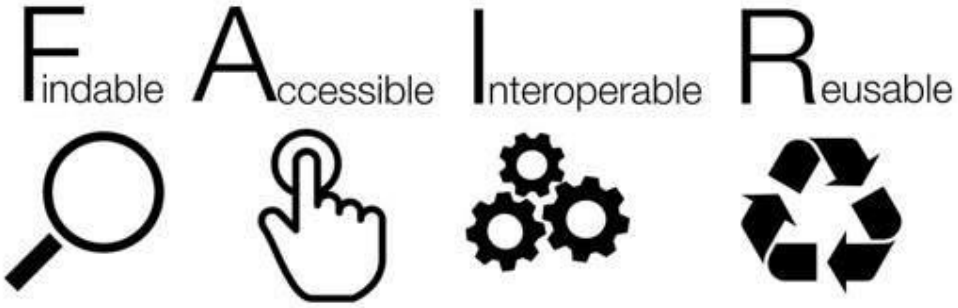
plutôt que les recréer permet de rentabiliser la recherche. C'est un gain de temps et d'argent.

- **Interdisciplinarité** : la constitution de bases de données permet la fouille de données, mais aussi d'extraire celles-ci, de les recouper et d'en construire des visualisations. La gestion et le partage des données facilitent de nouvelles recherches et le croisement de données provenant de différentes disciplines.
- **Data driven science** : c'est une démarche qui automatise les découvertes en exploitant la puissance de calcul des ordinateurs pour trouver des corrélations parmi des grandes quantités de données.
- **Exhumation de données « fossilisées »** : les publications permettent d'accéder à environ 10 % des données, le reste demeurant disponible mais non utilisé sur les disques durs d'ordinateurs. On parle de "données fossilisées". Une bonne gestion et un partage de ces données éviteraient la perte de données uniques.
- **Valeur patrimoniale** : certaines données de recherche présentent une valeur patrimoniale. Une bonne gestion et un partage de ces données de recherche sont nécessaires à leur valorisation.

2. Principes FAIR

2.1. Définition

« Les principes FAIR sont 4 principes à respecter pour garantir une utilisation optimale des données de recherche et des métadonnées associées, à la fois par les hommes et par les machines. »



Source image : Par SangyaPundir — Travail personnel, CC BY-SA 4.0.
https://commons.wikimedia.org/wiki/File:FAIR_data_principles.jpg

2.2. Les principes FAIR en quelques mots

2.2.1. Rendre ses données de recherche FAIR

F- Facile à trouver

Le principe F est mis en œuvre par l'utilisation d'**identifiants pérennes** (par ex. : DOI), de **métadonnées** riches, par le **signalement dans des catalogues**, des **entrepôts**...

A- Accessible

Le principe A signifie la mise en place d'un **stockage durable** des données et des **métadonnées**, avec accès et/ou téléchargement facilités (**protocoles** de communication **standardisés** et **ouverts**), et spécification des **conditions d'accès** et **d'utilisation**.

I – Interopérable

Le principe I signifie que la donnée est **téléchargeable**, **utilisable**, **intelligible** et **combinable** avec d'autres données, par des humains et des machines, grâce à l'utilisation de **formats standards**, de **vocabulaires** et **d'ontologies**.

R – Réutilisable

Le principe R s'appuie sur les caractéristiques qui rendent les données réutilisables pour de **futures recherches ou d'autres finalités** (enseignement, innovation, reproduction/transparence de la science). Cela est rendu possible par une **description riche** qui précise la **provenance** des données, l'utilisation de **standards communautaires** et l'ajout de **licences**.

INRAE, Institut National de Recherche pour l'agriculture, l'alimentation et l'environnement. Produire des données FAIR. https://science-ouverte.inrae.fr/les-donnees-et-le-numerique-scientifiques/produire-des-donnees-fair

L'adoption progressive de ces principes FAIR va rendre les données plus faciles à partager et réutilisables aussi bien par les hommes que par les systèmes informatiques.

2.3. Exemples de mise en œuvre des principes FAIR

De nombreuses actions recommandées pour la gestion et le partage des données de recherche répondent en partie ou en totalité aux principes FAIR.

En voici quelques exemples :

Je suis chercheur en Écologie marine : je sauvegarde et partage mes données de manière sécurisée tout au long du projet grâce aux services proposés par SISMER.	J'organise et nomme mes fichiers de la même manière que l'ensemble des partenaires du projet...
Je travaille dans le domaine de l'écologie : je renseigne les métadonnées associées à mes données selon le standard EML (Ecological Metadata Language)...	Mon domaine est l'aménagement du territoire : j'utilise un vocabulaire contrôlé disciplinaire, le thésaurus GEMET...
J'applique une licence Etalab ou Creative Commons à mes jeux de données...	Mes jeux de données sont identifiés de manière unique et pérenne par un DOI...
Je dépose mes jeux de données dans l'entrepôt de données DRYAD...	Je communique mes codes sources.
Je mets à disposition mes fichiers en .csv plutôt qu'en .xls, c'est-à-dire dans un format ouvert et non propriétaire...	Je suis ethnobotaniste : durant mon projet de recherche, j'ai réalisé des interviews qui présentent une grande valeur patrimoniale. Je dépose mes jeux de données dans la plateforme d'archivage pérenne du CINES ...

Ces différentes actions contribuent à rendre mes données FAIR !

Ce cours a été conçu de manière progressive. Vous trouverez toutes les explications, conseils et outils pour savoir comment respecter les principes FAIR en pratique dans

un second temps. Ce sera l'objet de la partie "Contenu du PGD/DMP" où chaque aspect sera détaillé.

2.4. Jeu sur les principes FAIR

D'après vous, qu'apportent les principes FAIR au chercheur et à la communauté scientifique ?

Consigne : Un chercheur a produit et partagé des données dans le cadre d'un projet de recherche, en accord avec les principes FAIR. Cela lui offre des bénéfices immédiats dans le cadre de son projet et de sa carrière, mais cela peut profiter aussi plus tard à la communauté scientifique.
Placez chaque carte sur une des deux zones identifiées "Pour le chercheur" et "Pour la communauté scientifique".

1.Balisage du cycle de vie des données par de bonnes pratiques 2.Accès facile à des données publiques, réutilisables 3.Gain de temps dans la gestion d'un projet de recherche 4.Accès à des corpus utiles pour d'autres domaines 5.Interopérabilité des données 6.Meilleure visibilité du chercheur et de ses travaux de recherche 7.Gain de temps et d'argent : ne pas recréer des données déjà existantes 8.Clarification de l'environnement de travail 9.Reproductibilité de la recherche 10.Favorisent les collaborations	A.Pour le chercheur B.Pour la communauté scientifique

Solution : 1A / 2B / 3A / 4B / 5B / 6A / 7B / 8A / 9B / 10A

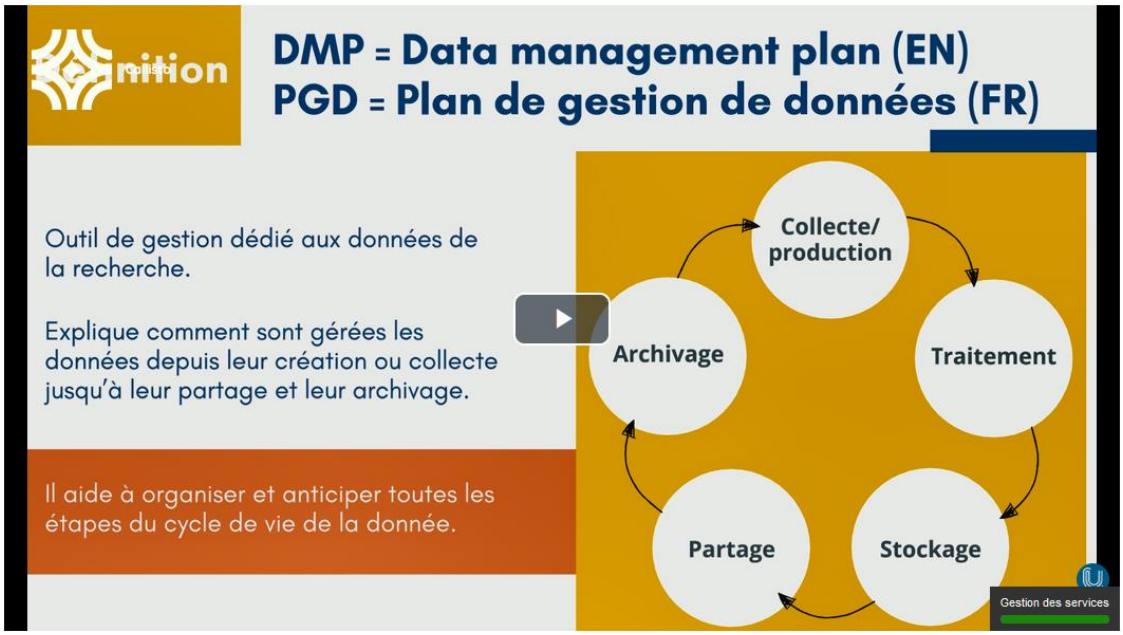
3. Plan de gestion de données (PGD) ou Data management plan (DMP)

3.1. Définition

" Le Plan de Gestion des Données est un outil de gestion. Il se présente sous la forme d'un document évolutif, structuré en rubriques. Il a pour objectif de synthétiser la description et l'évolution des jeux de données de votre projet de recherche. Il prépare le partage, la réutilisation et la pérennisation des données. "

Remarque : nous utilisons indifféremment les sigles PGD ou DMP au fil du parcours.

Découvrez en 3 minutes ce qu'est un Plan de Gestion de Données :



https://www.canal-u.tv/chaines/callisto/les-minutes-doranum/la-minute-plan-gestion-de-donnees

3.2. Le PGD, un document incontournable

3.2.1. À l'échelle mondiale

Le PGD est devenu un outil de pilotage très répandu. Il est de plus en plus recommandé ou exigé, partout dans le monde.

3.2.2. À l'échelle européenne

La rédaction d'un PGD est une exigence de la Commission européenne : la version initiale du PGD est inscrite parmi les livrables à 6 mois après le début du projet (Modèles Horizon Europe, ERC-European Research Council, Science Europe).
Pour favoriser la gestion et le partage des données de la recherche, de nombreuses initiatives d'ampleur européenne ont été déployées, notamment des outils et des infrastructures, par exemple :

- OpenAIRE (Open Access Infrastructure for Research in Europe),
- EOSC (European Open Science Cloud)…

3.2.3. À l'échelle nationale

- L'État français a élaboré une politique nationale avec un premier Plan national pour la science ouverte (juillet 2018) qui recommande la généralisation de la mise en place de plans de gestion des données dans les appels à projets de recherche. Après un bilan positif, cette politique nationale a été précisée et renforcée dans un deuxième Plan national pour la science ouverte (juillet 2021), notamment par la mise en œuvre de l'obligation de diffusion des données de recherche financées sur fonds publics.

- Dans le cadre des Plans nationaux pour la science ouverte, les agences françaises de financement de la recherche, l'ADEME, l'ANR, l'ANRS-MIE, l'Anses, et l'INCa, ont constitué un réseau d'échanges pour définir une approche concertée en faveur de la science ouverte.

Parmi les **engagements pris par les institutions et les agences de financem**ent on retrouve :

- La promotion du partage et de l'ouverture des données de la recherche
- La rédaction d'un Plan de Gestion des Données, dès le démarrage du projet de recherche, afin de préparer les données à leur partage et à leur diffusion éventuelle dans le respect du principe « aussi ouvert que possible, aussi fermé que nécessaire ».
- Un modèle de PGD commun, celui développé par Science Europe (mis à la disposition des chercheurs dans l'outil DMP OPIDoR).

ANR, Agence nationale de la recherche. Science ouverte : point d'étape sur la politique commune du réseau des agences de financement françaises. 11 mars 2022. *https://anr.fr/fr/actualites-de-lanr/details/news/science-ouverte-point-detape-sur-la-politique-commune-du-reseau-des-agences-de-financement-franca/*

3.2.4. À l'échelle des organismes

- De nombreux instituts, organismes, établissements mettent à disposition de leur communauté des modèles de PGD institutionnels (modèles de PGD du CNRS, du CIRAD, de l'INRAE, d'AgroParisTech, des universités de Lille, Montpellier, Paris, Toulouse…).
- Certains établissements comme l'INRAE, le CNRS et AgroParisTech ont mis en place une politique d'établissement pour la gestion et le partage des données de recherche.
- De plus en plus d'établissements émettent des recommandations institutionnelles.

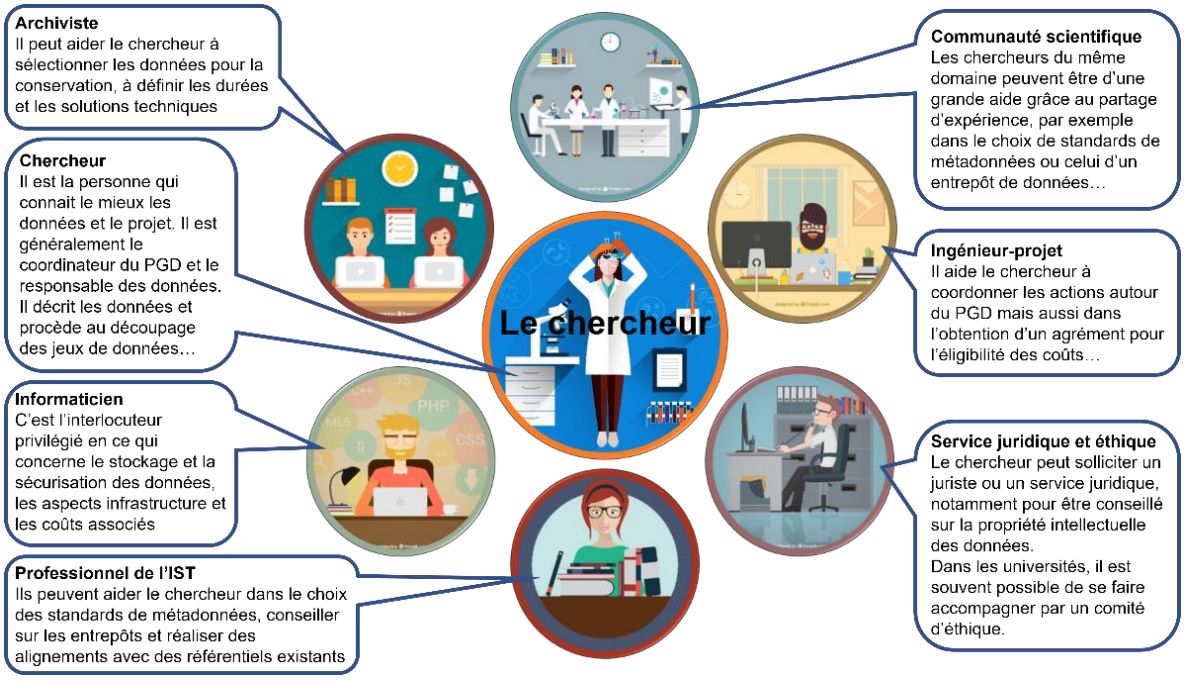
3.2.5. À l'échelle disciplinaire

Des modèles de PGD disciplinaires commencent à être mis en place. C'est le cas en géographie (modèle PRODIG, disponible en français dans l'outil DMP OPIDoR). Ce modèle comporte des recommandations pouvant s'appliquer au domaine de l'environnement.

3.3. Acteurs et contributeurs

Le chercheur n'est pas seul face à la rédaction du PGD.
Le PGD est une opportunité de dialogue entre les différents acteurs d'un projet : scientifiques, informaticiens, professionnels de l'IST (Information Scientifique et Technique), chargés de projet, juristes...
La gestion des données demande un effort collectif !

Consigne : cliquer sur les petites croix qui accompagnent les images ci-dessous pour découvrir quelle aide peut apporter chaque contributeur.



Sources : images Freepik

(Designed by Freepik ; Designed by macrovector / Freepik Designed by makyzz / Freepik)

Détail des acteurs et contributeurs :

- **Chercheur**

  Il est la personne qui connait le mieux les données et le projet. Il est généralement le coordinateur du PGD et le responsable des données. Il décrit les données et procède au découpage des jeux de données…

- **Communauté scientifique**

  Les chercheurs du même domaine peuvent être d'une grande aide grâce au partage d'expérience, par exemple dans le choix de standards de métadonnées ou celui d'un entrepôt de données…

- **Ingénieur-projet**

  Il aide le chercheur à coordonner les actions autour du PGD mais aussi dans l'obtention d'un agrément pour l'éligibilité des coûts…

- **Service juridique et éthique**
  Le chercheur peut solliciter un juriste ou un service juridique, notamment pour être conseillé sur la propriété intellectuelle des données.
  Dans les universités, il est souvent possible de se faire accompagner par un comité d'éthique.

- **Professionnel de l'IST**
  Ils peuvent aider le chercheur dans le choix des standards de métadonnées, conseiller sur les entrepôts et réaliser des alignements avec des référentiels existants.

- **Informaticien**
  C'est l'interlocuteur privilégié en ce qui concerne le stockage et la sécurisation des données, les aspects infrastructure et les coûts associés.

- **Archiviste**
  Il peut aider le chercheur à sélectionner les données pour la conservation, à définir les durées et les solutions techniques.

Les universités, infrastructures et organismes de recherche émettent souvent des recommandations à destination de leurs communautés de recherche.

Le répertoire des Services Opérationnels de Soutien à la rédaction des Plans de Gestion des Données (SOS-PGD) recense les services accompagnant la rédaction des plans de gestion des données au sein des établissements d'enseignement supérieur et de la recherche. Il vise à aider les chercheurs à identifier leurs interlocuteurs au sein de leur institution et à faciliter la mise en relation entre les services supports de différentes institutions pour les projets de recherche multi-partenariaux.

Dans le cadre de l'écosystème *Recherche Data Gouv*, les ateliers de la donnée sont en proximité géographique des équipes de recherche pour leur apporter une première expertise dans la gestion raisonnée des données de recherche.

De même les financeurs (comme l'ANR ou la Commission européenne) peuvent proposer des conseils ou donner des consignes précises (par exemple : obligation de rédaction d'un PGD dans les 6 mois suivant le début du projet pour les projets financés par l'ANR).

3.4. Le PGD, un outil de gestion de projet

C’est un document évolutif, dynamique et continuellement mis à jour (introduction d’un nouveau jeu de données, données faisant l’objet d’un dépôt de brevet, changement dans le consortium…).

C'est aussi un outil de gestion de projet qui aide à organiser ses données, à les décrire, à bien définir les responsabilités, les ressources nécessaires et à produire des données fiables.

- **Organisation des données**
  Le PGD aide à bien organiser les données, tout au long du projet.

- **Document évolutif**
  Il faut commencer à rédiger le PGD dès le début du projet, avec les éléments déjà connus ou prévus. Ensuite compléter le PGD au fur et à mesure.
  Prévoir 2 versions au minimum : au début et à la fin du projet. Pour les projets de plus de 30 mois, une version intermédiaire est demandée.

- **Description des données**
  Dans le PGD, il faut décrire la façon dont les données seront obtenues, traitées, organisées, stockées, sécurisées, préservées, partagées… (cycle de vie des données).

- **Responsabilités**
  Dans le PGD, désigner la ou les personne(s) responsable(s) de la gestion des données pour toutes les étapes du projet et au sein du partenariat s’il y a lieu : saisie des données ; production des métadonnées ; contrôle de la qualité des données ; stockage, partage et archivage des données ; mise à jour du PGD.

On peut désigner des personnes nominativement ou indiquer une fonction si la personne qui l'occupe peut être amenée à changer au cours du projet.

- **Ressources**
  Il est demandé d'évaluer les ressources nécessaires (budget, temps alloué, personnels) permettant la mise en œuvre des actions décrites dans le PGD : temps nécessaire à la préparation des données pour le stockage, le partage et l'archivage des données ; coûts de matériel et rémunération des personnels ; frais de stockage (serveurs dédiés, traitement, maintenance, sécurité, accès…), frais de partage (site web, publication…) et frais d'archivage des données.

- **Données fiables**
  Le PGD permet aux producteurs de données de se poser les bonnes questions et donc d'améliorer la fiabilité de leurs données.

La rédaction d'un plan de gestion permet ainsi d'initier très tôt une réflexion collective sur les bonnes pratiques et d'anticiper les questions relatives à la gestion des données (comme le choix de l'entrepôt, la documentation à associer…).

3.5. Différents modèles

3.5.1. Éléments communs à tous les modèles

Il n'existe pas de trame unique. Toutefois de nombreux modèles de DMP ont été établis par des organismes, instituts, financeurs à destination de leurs utilisateurs, afin de répondre aux spécificités propres à certains organismes de recherche, pour correspondre au contexte local des établissements, etc.

On y retrouve néanmoins les mêmes éléments, à savoir :

- Informations administratives
- Description des données
- Documentation, métadonnées, standards
- Aspects juridiques
- Sécurité des données

- Stockage des données durant le projet
- Partage des données après le dépôt dans un entrepôt
- Archivage pérenne
- Coûts
- Responsabilités

3.5.2. Exemple de modèle de PGD : le modèle ANR structuré

Le modèle de plan de gestion de données ANR structuré est composé de 6 grandes thématiques illustrant les bonnes pratiques de gestion et de partage. Pour chaque thématique, le chercheur est invité à répondre à plusieurs questions.

1. Description des données et collecte ou réutilisation des données existantes

1.1. Description générale du produit de recherche

1.2. Est-ce que des données existantes seront réutilisées ?

1.3. Comment seront produites/collectées les nouvelles données ?

2. Documentation et qualité des données

2.1. Quelles métadonnées et quelle documentation (par exemple mode d'organisation des données) accompagneront les données ?

2.2. Quelles seront les méthodes utilisées pour assurer la qualité scientifique des données ?

3. Exigences légales et éthiques, codes de conduite

3.1. Quelles seront les mesures appliquées pour assurer la protection des données à caractère personnel ?

3.2. Quelles sont les contraintes juridiques (sensibilité des données autres qu'à caractère personnel, confidentialité…) à prendre en compte pour le partage et le stockage des données ?

3.3. Comment les éventuelles questions éthiques seront-elles prises en compte, les codes déontologiques respectés ?

4. Traitement et analyse des données

4.1. Comment et avec quels moyens seront traitées les données ?

5. Stockage et sauvegarde des données pendant le processus de recherche

5.1. Comment les données seront-elles stockées et sauvegardées tout au long du projet ?

6. Partage des données et conservation à long terme

6.1. Comment les données seront-elles partagées ?

6.2. Comment les données seront-elles conservées à long terme ?

3.5.3. Exemples de modèles de PGD plus adaptés au domaine de l'environnement

L'INRAE propose deux types de modèles disponibles dans les modèles de DMP OPIDoR :

- L'un pour les projets de recherche en français ou en anglais. On y trouve des informations sur les services, les outils et les bonnes pratiques recommandées par l'INRAE pour gérer, partager et réutiliser les données de la recherche sur le site « Le numérique pour la science et les données scientifiques ».
- L'autre pour les structures (entités de recherche) qui peut être utilisé pour gérer les données produites et utilisées dans toute type de structure (unité, plateforme, observatoire...), indépendamment d'un projet de recherche.

Le CIRAD propose un modèle spécifique basé sur le modèle H2020, en version française et anglaise.
Le modèle PRODIG (en français) a été élaboré sur la base de celui proposé par l'ANR. Il comporte des recommandations et des exemples de réponses pouvant s'appliquer aux domaines de l'environnement.

Un modèle de DMP se distingue et a été conçu pour un usage spécifique : le modèle de Plan de Gestion de Logiciel.

3.6. **L'outil de rédaction DMP OPIDoR** (Data Management Plan pour une Optimisation du Partage et de l'Interopérabilité des Données de la Recherche)

DMP OPIDoR est un outil de rédaction du DMP.

Cet outil collaboratif en ligne est accessible gratuitement à l'ensemble de la communauté scientifique de l'Enseignement Supérieur et de la Recherche ainsi qu'à ses partenaires français ou étrangers.

Pour créer des PGD, il est nécessaire de se créer un compte.

Attention : DMP OPIDoR est uniquement un outil de rédaction. Il n'a pas été conçu pour servir d'entrepôt de DMPs et n'a donc pas vocation à conserver les DMPs sur le long terme.

Ces tutoriels permettent de prendre en main l'outil en toute autonomie, avec des explications pas à pas :

- L'outil de rédaction DMP OPIDoR
- DMP OPIDoR – Le modèle structuré

3.7. Exemples de PGDs publics concernant des projets dans le domaine de l'environnement

- **SNO KARST**

  Modèle ANR - 21/04/2022 - 10 p.

  En français

  Le SNO KARST vise essentiellement à caractériser l'état qualitatif et quantitatif de la ressource en eau des hydrosystèmes karstiques, et à prévoir son évolution en réponse à des forçages à différentes échelles temporelles. Cette problématique se décline en trois questions scientifiques et challenges :

  - Les mécanismes de transfert et de transport en milieu karstique
  - Les liens entre structure géologique et écoulement
  - L'évolution des flux d'eau et de matière en réponse aux changements globaux.

- **IMPRINT**

  Modèle ANR - 12/03/2020 - 7 p.

  En français

  " Les modèles de niche, couramment utilisés pour prédire la redistribution du vivant en contexte de réchauffement climatique, sont habituellement calibrés sur les

températures synoptiques (i.e. macroclimat) mais ignorent les températures ressenties (i.e. microclimat). Pourtant, le microclimat ressenti par de nombreux organismes peut être très différent du macroclimat régional, surtout au sein des écosystèmes forestiers dont la sylviculture peut servir de médiateur entre microclimat et macroclimat. L'objectif ultime du projet IMPRINT est d'utiliser la reconstruction à long terme du microclimat sous-couvert forestier dans des modèles de niche afin de montrer comment le microclimat affecte les prédictions de redistribution de la biodiversité forestière en contexte de réchauffement et ce par rapport aux modèles de niche actuels basés sur le macroclimat. "

- **RINGO (Readiness of ICOS for Necessities of integrated Global Observations)**
  Modèle H2020
  Ouvrir la rubrique Deliverables, Documents, reports
  Version initiale (D6.4) - 30/06/2017 - 6 p.
  Version finale (D6.12) - 24/02/2020 - 10 p.
  En anglais
  Ce projet de 4 ans compte 43 partenaires dans 19 pays et comprend cinq modules de travail qui mettent l'accent sur le développement de l'infrastructure de recherche ICOS (ICOS RI) afin de favoriser sa durabilité. Les principaux objectifs sont :
  1. La disponibilité scientifique : soutenir la poursuite de la consolidation des réseaux d'observation et améliorer leur qualité.
  2. La préparation géographique : améliorer l'adhésion à ICOS et sa durabilité en aidant les pays intéressés à créer un consortium national, à promouvoir ICOS auprès des parties prenantes nationales, à recevoir des conseils et à recevoir une formation pour améliorer la préparation des scientifiques à travailler au sein d'ICOS.
  3. La préparation technologique : développer et normaliser les technologies d'observation des gaz à effet de serre nécessaires pour répondre aux nouvelles demandes de connaissances et pour prendre en compte et contribuer aux progrès technologiques.
  4. La disponibilité des données : améliorer les flux de données vers les différents groupes d'utilisateurs, en s'adaptant aux normes (web) en développement et dynamiques.
  5. La préparation politique et administrative : approfondir la coopération mondiale des infrastructures d'observation et l'impact sociétal commun.

- **AQUACOSM**

  Modèle H2020 - 30/06/2017 - 18 p.

  En anglais

  Network of Leading European AQUAtic MesoCOSM Facilities - Connecting Mountains to Oceans from the Arctic to the Mediterranean

  AQUACOSM recueillera des données sur le mésocosme aquatique auprès de 37 installations dans toute l'Europe. L'objectif d'AQUACOSM est de faire progresser la science du mésocosme grâce à une expérimentation plus standardisée et synchronisée pour une meilleure compréhension des relations de cause à effet de l'écosystème aquatique.

- **Migration of legacy data to new media formats for long-time storage and maximum visibility: Modern pollen data from the Canadian Arctic**

  25/08/2016 - 4 p.

  En anglais

  Ce plan de gestion des données décrit les données modernes sur le pollen collectées le long d'un transect de 2500 miles (~4000 km) dans l'Arctique canadien en 1972/73 dans le cadre d'un projet de recherche financé par la NSF. Le projet a été entrepris à l'Institut de recherche arctique et alpine de l'Université du Colorado à Boulder. Cet ancien ensemble de données, initialement stocké sur papier et sur film 35 mm, sera transféré dans des formats numériques qui permettront de le télécharger dans une bibliothèque internationale en libre accès pour un stockage pérenne et une meilleure visibilité.

  L'étude fournit un jeu de données de pollens pour l'interprétation des diagrammes polliniques holocènes de cette région et pour la comparaison avec les échantillons modernes de pollens de surface, ce qui permet d'évaluer les effets du changement climatique moderne sur les écosystèmes arctiques et subarctiques.

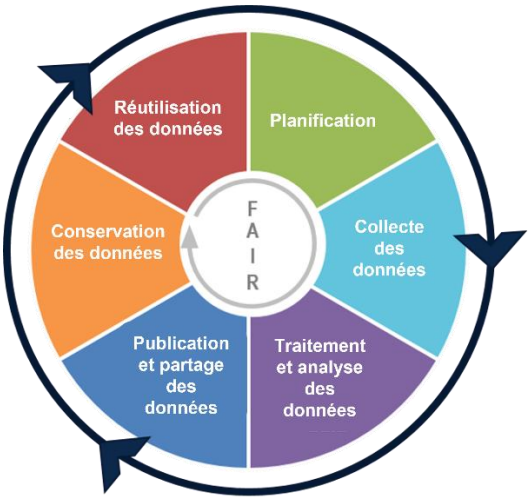
LE CONTENU DU PGD/DMP

4. Aspects juridiques et éthiques

4.1. Les aspects juridiques et éthiques dans le cycle de vie des données

Les aspects juridiques et éthiques accompagnent tout le cycle de vie des données. Dès le début du projet, il faut s'y intéresser et se poser les bonnes questions.

En cas de doute, ne pas hésiter à prendre conseil auprès des juristes de votre institution.



4.2. Droits et obligations du chercheur

Dès le début du projet, au moment de la collecte et de la production des données, le chercheur doit être vigilant concernant ses droits et obligations.

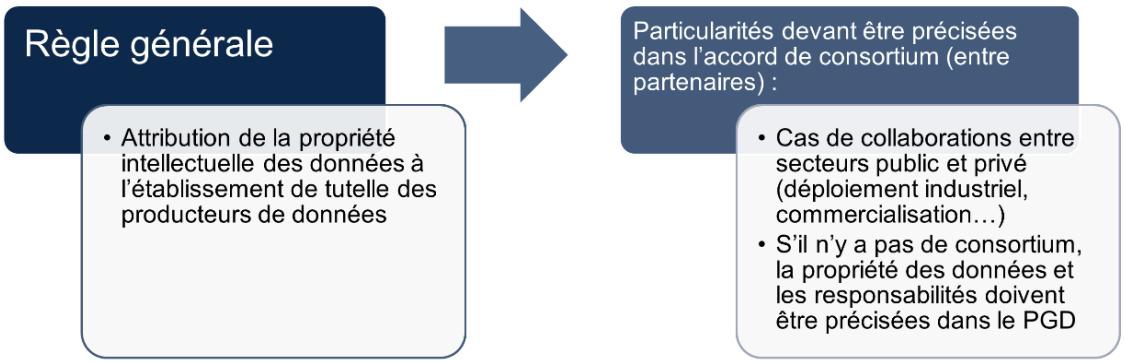
Ce point est crucial car il détermine la latitude dont le chercheur disposera ensuite pour publier, diffuser et communiquer ses données et les résultats de ses recherches. Exemples :

- Dans le cas d'une interview ou de prises de son/de vue, le chercheur doit recueillir le consentement écrit des personnes concernées, car il est question de données personnelles (RGPD) et de droit à l'image.
- Dans le cas d'une consultation de données d'archives, quels sont les droits afférents ?

- Dans le cas d'une collecte d'échantillons biologiques ou hydrologiques, quels sont les droits liés au pays de collecte ou aux mesures de protection internationales ? Quelle est la réglementation locale en matière de respect de l'environnement et des populations... ? (Voir par exemple les recommandations de l'IRD).
- Dans le cas de données géographiques, la directive INSPIRE s'applique (élaborée par la Direction générale de l'environnement de la Commission européenne en 2007, elle vise à établir en Europe une infrastructure de données géographiques pour favoriser la protection de l'environnement, assurer l'interopérabilité entre bases de données et faciliter la diffusion, la disponibilité, l'utilisation et la réutilisation de l'information géographique en Europe. INSPIRE vise ainsi à mieux partager les données de la recherche).

Hadrossek Christine, Janik Joanna, Libes Maurice, Louvet Violaine, Quidoz Marie-Claude, Rivet Alain, Romier Geneviève. Atelier Données. Guide de bonnes pratiques sur la gestion des données de la recherche. Version 2.0. 23 août 2023. https://mi-gt-donnees.pages.math.unistra.fr/guide/

4.3. Propriété intellectuelle des données de recherche



" […] Il existe une différence importante de régime juridique applicable aux œuvres de type écrits scientifiques et aux données. Si l'on entend le terme « données de recherche » au sens strict (c'est-à-dire d'informations générées dans le cadre d'un processus de recherche), alors le chercheur ne sera pas considéré comme un « auteur », car les données ne sont généralement pas en tant que telles des « œuvres » protégeables par le droit d'auteur. Un droit de propriété intellectuelle spécifique existe néanmoins pour la protection des bases de données, mais son fonctionnement est

différent de celui du droit d'auteur. Ce droit dit de « producteur de base de données » n'appartient pas à l'origine aux personnes physiques qui réalisent la base, mais à l'entité qui a effectué des « investissements substantiels » pour rendre cette opération possible. Dans la plupart des hypothèses, ce seront donc les établissements de tutelle des chercheurs qui auront la qualité de « producteurs » et possèderont les droits attachés aux bases de données de recherche.

Or la loi République numérique a explicitement « neutralisé » le droit des bases de données des administrations pour faire primer le principe de libre réutilisation. Le nouvel article 11 du texte indique ainsi :

« Sous réserve de droits de propriété intellectuelle détenus par des tiers, les droits des administrations mentionnées au premier alinéa de l'article L. 300-2 du présent code, au titre des articles L. 342-1 et L. 342-2 du code de la propriété intellectuelle [c'est-à-dire le droit de producteur de bases de données], ne peuvent faire obstacle à la réutilisation du contenu des bases de données que ces administrations publient en application du 3° de l'article L. 312-1-1 du présent code. »

Il en résulte que les données produites par les chercheurs sont bien comprises dans le principe d'ouverture par défaut. La situation sera donc très différente de celles des écrits scientifiques et autres créations produites par les chercheurs dans le cadre de leurs activités. "

Maurel Lionel. La réutilisation des données de la recherche après la loi pour une République numérique. La diffusion numérique des données en SHS - Guide de bonnes pratiques éthiques et juridiques. 2018. https://hal.science/hal-01908766

4.4. Recommandations et obligations

Il est important de prendre en compte, dès le début du projet, les aspects de diffusion et de partage des données qui interviendront plus tard dans le cycle de vie de la donnée. Les institutions émettent souvent des recommandations dans ce sens. Dans certains cas, il peut aussi y avoir une obligation de la part des financeurs.

Recommandations	Obligations
Certaines institutions peuvent émettre des recommandations précises en matière de diffusion et de partage des données produites. Exemple de recommandation de l'INRAE : Répertorier les jeux de données susceptibles d'échapper au principe de diffusion : données scientifiques protégées ou à risques (sécurité état, sécurité des populations, etc.), données personnelles et données de santé, données liées à l'intelligence économique (secret industriel et commercial), données soumises au secret statistique, etc. Le guide « Ouverture des données de recherche. Guide d'analyse du cadre juridique en France » (auquel a participé l'INRAE) précise les modalités de communication des données qui, selon leur nature, peut être rendue obligatoire, interdite, ou soumise à conditions. Ce document de référence explicite par ailleurs les principes à respecter en matière de diffusion des données. Il rappelle les critères techniques à satisfaire pour atteindre la qualification de "données ouvertes" et oriente sur le choix délicat de la licence de diffusion, et les modalités de diffusion. Il fournit enfin un logigramme d'aide à la décision et une série de fiches pratiques. *INRAE. Les aspects éthiques et juridiques, et la propriété intellectuelle. https://science-ouverte.inrae.fr/les-donnees-et-le-numerique-scientifiques/partager-publier-des-donnees-et-des-codes* ○ *Becard Nicolas, Castets-Renard Céline, Chassang Gauthier, Dantant Martin, Freyt-Caffin Laurence, Gandon Nathalie, Martin Caroline, Martelletti Andrea, Mendoza-Caminade Alexandra, Morcrette Nathalie, Neirac Claire. Ouverture des données de la recherche. Guide d'analyse du cadre juridique en France. Décembre 2017. https://hal.science/hal-02791224*	Suivant le financeur du projet, il peut être obligatoire de diffuser ses données et de rédiger un Plan de Gestion des Données (PGD), toujours en respectant le principe " **aussi ouvert que possible, aussi fermé que nécessaire** ". Ainsi les données relevant du secret médical ne seront pas diffusées et les données liées à un brevet ne pourront être accessibles qu'après le dépôt de celui-ci. Exemples d'obligations concernant le PGD : - **ANR** : élaboration obligatoire d'un PGD pour tous les projets financés, dans les 6 mois qui suivent le démarrage du projet - **Horizon Europe** : obligation de rendre librement accessibles les articles scientifiques et les données (dont celles liées aux publications) et rédaction obligatoire d'un PGD selon le principe "aussi ouvert que possible, aussi fermé que nécessaire".

4.5. Communicabilité des données

La communicabilité des jeux de données peut être conditionnée par :

- La nature ou le type des données
- L'origine des données
- Leur(s) utilisation(s).

Elle peut être empêchée temporairement ou définitivement.

Toute restriction doit être mentionnée et expliquée dans le PGD.

Communication obligatoire pour certaines disciplines :
- Données géographiques
- Données environnementales...

Communication sous conditions :
- Données protégées par le droit d'auteur ou par contrat
- Données personnelles
- Statistiques...

Communication interdite par principe :
- Secret professionnel
- Secret défense
- Sécurité de l'établissement...

Outil " Aide à la décision sur la diffusion des données de recherche " du CIRAD.

4.6. Accès, sécurité et licences

4.6.1. Accès

En fonction de l'avancée du projet de recherche, les modalités d'accès pourront être différentes.

Durant le projet, il peut être nécessaire, voire crucial, de limiter l'accès aux données aux seuls membres de l'équipe de recherche.
Une fois le projet achevé, il peut être tout aussi important de limiter l'accès aux données.

On distingue 4 modes d'accès aux données : ouvert, avec embargo, restreint et fermé.

En fonction du mode d'accès souhaité, différentes modalités peuvent être mises en place :

- **Accès limité par un mot de passe**
  Il est nécessaire de s'authentifier pour accéder aux données.
- **Accès limité à certaines personnes**
  Il peut être pertinent selon les cas de limiter l'accès de vos données uniquement aux membres du consortium ou à une communauté scientifique par exemple.
- **Accès limité dans le temps avec embargo**
  L'accès à vos données avec embargo peut
  - dépendre de la discipline
  - permettre de disposer du temps nécessaire au dépôt de brevets
  - être déterminé par les éditeurs.

4.6.2. Sécurité

On peut procéder à une sécurisation des données elles-mêmes grâce :
- Au chiffrement
- À la pseudonymisation
- À l'anonymisation.

Une attention toute particulière doit être portée aux données sensibles et aux données à caractère personnel !

Attention, pseudonymisé ne veut pas dire anonymisé !
Contrairement à l'anonymisation qui est une action irréversible, la pseudonymisation est réversible, elle peut donc permettre d'identifier une personne physique.

Pseudonymisation	Anonymisation
Pour connaître les techniques de pseudonymisation, vous pouvez consulter les guides d'Etalab : https://guides.etalab.gouv.fr/pseudonymisation/pourquoi-comment/#quelles-sont-les-differentes-methodes-de-pseudonymisation	Pour en savoir plus, vous pouvez consulter le site de la CNIL : https://www.cnil.fr/fr/lanonymisation-de-donnees-personnelles OpenAIRE a développé l'outil spécifique Amnesia qui modifie les informations personnelles et sensibles, élimine toute violation de la confidentialité des données et toute exposition d'informations sensibles.

4.6.3. Licences

Attribuer une licence à vos données est très important car cela permet de bien définir les modalités de réutilisation et de les afficher clairement.

Si besoin, vous pouvez attribuer une licence différente à chacun de vos jeux de données.

Il existe de nombreuses licences, modulables en fonction de vos besoins pour certaines, très spécifiques pour d'autres. Voici quelques exemples :

Licence Ouverte (Etalab)
En France, la réglementation stipule que la Licence Ouverte (Etalab) doit être attribuée aux données publiques.
Cette licence ouverte, libre et gratuite est compatible avec la CC-BY 2.0, ce qui veut dire que la paternité des données devra être mentionnée en cas de réutilisation des données.
Consulter le décret n° 2017-638

Creative Commons CC0

C'est une licence ouverte conçue pour les données dédiées au domaine public.

Voir le site Creative Commons

Licences CC - Creative Commons

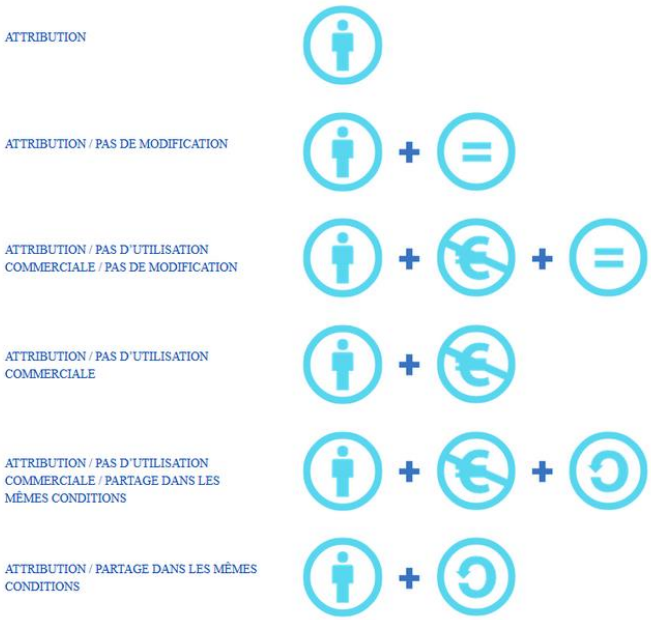
Il existe 6 licences gratuites Creative Commons combinant quatre éléments :

- BY = attribution
- NC = pas d'utilisation commerciale
- SA = partage dans les mêmes conditions
- ND = pas de modification

Voici les 6 licences Creative Commons et leurs icônes correspondantes (illustration ci-dessous) :

- CC BY
- CC BY-ND
- CC BY-NC-ND
- CC BY-NC
- CC BY-NC-SA
- CC BY-SA

La plus permissive est la CC BY et la plus restrictive est la CC BY-NC-ND.



Source : site Creative Commons

Licences pour logiciels

Certaines licences sont dédiées aux logiciels comme la licence de logiciel libre GNU GPL (GNU General Public License ou licence publique générale GNU) ou la licence de logiciel libre CeCILL-B. Cette dernière a été créée conjointement par le CEA, le CNRS et l'INRIA (en savoir plus).

Licences pour bases de données

Il existe des licences spécifiques aux bases de données comme la licence libre Open Database License (ODbL).

En France, il est recommandé de **citer à minima les auteurs** (donc d'utiliser a minima la Licence Ouverte ou la licence CC-By).
Notez que dans le cas de données publiques ayant une visibilité internationale, **il est tout à fait possible d'attribuer à la fois une Licence ouverte Etalab et une licence CC-By**.

4.6.4. Outils et liens utiles

Voici quelques outils et liens utiles pour choisir une licence :

- **data.gouv.fr**
  Ce site recense les licences de réutilisation autorisées dans le cadre de la loi pour une République numérique pour les "informations publiques (données, documents...)" et codes sources, ainsi que les licences spéciales homologuées.
- **Licentia by inria**
  Cet outil permet de choisir quelle licence attribuer à ses données en utilisant quelques critères (permissions / obligations / interdictions), de déterminer si une licence est compatible avec ses besoins, de visualiser et télécharger une licence, de la convertir en RDF.
- **License selector**
  Cet outil permet de choisir une licence en répondant à des questions.
- **Choose an open source license**
  Ce site permet de choisir une licence en fonction de ses besoins.

4.7. Intégrité scientifique et éthique des données de recherche

4.7.1. Définitions et principes fondamentaux

« L'**éthique** nous invite à réfléchir aux valeurs qui motivent nos actes et à leurs conséquences et fait appel à notre sens moral et à celui de notre responsabilité. La **déontologie** réunit les devoirs et obligations imposés à une profession, une fonction ou une responsabilité. L'**intégrité scientifique** concerne, quant à elle, la « bonne » conduite des pratiques de recherche. »

CNRS, Centre national de la recherche scientifique. Responsabilité de recherche. *https://www.cnrs.fr/fr/le-cnrs/responsabilites/responsabilite-de-recherche*

Il existe des enjeux pratiques, éthiques et intellectuels inhérents à la recherche. Les bonnes pratiques reposent sur des principes fondamentaux qui orientent les chercheurs dans leurs travaux et dans leur engagement.

Ces principes sont les suivants :

- **Fiabilité** : garantir la qualité de la recherche, qui transparaît dans la conception, la méthodologie, l'analyse et l'utilisation des ressources.
- **Honnêteté** : élaborer, entreprendre, évaluer, déclarer et faire connaître la recherche d'une manière transparente, juste, complète et objective.
- **Respect** envers les collègues, les participants à la recherche, la société, les écosystèmes, l'héritage culturel et l'environnement.
- **Responsabilité** assumée pour les activités de recherche, de l'idée à la publication, leur gestion et leur organisation, pour la formation, la supervision et le tutorat, et pour les implications plus générales de la recherche.

ALLEA, All European Academies. The European Code of Conduct for Research Integrity. Revised Edition 2023. Berlin (DE). Juin 2023. *https://allea.org/wp-content/uploads/2023/06/European-Code-of-Conduct-Revised-Edition-2023.pdf*

Une partie du PGD concerne les aspects éthiques :

- Dans le cas de données devant respecter des règles d'éthique particulières, préciser les normes, déclarations, codes, politiques auxquels on se réfère.
- En signant la charte d'éthique, le chercheur prend un engagement important.
- En cas de recours à un comité d'éthique, expliquer le processus de recrutement et d'évaluation.

4.7.2. Exemple de démarche éthique à observer pour la préservation de la biodiversité et des espèces sensibles

Le GBIF (Global Biodiversity Information Facility) a publié en novembre 2020 une mise à jour du guide pour la publication des données sur les espèces sensibles, conçu pour offrir des conseils sur la manière de rendre les données d'occurrence des espèces rares, menacées et ayant une valeur commerciale aussi disponibles que possible et aussi protégées que nécessaire : Current Best Practices for Generalizing Sensitive Species Occurrence Data.

Ce guide offre un aperçu actuel important de la manière dont on peut partager les données sur les taxons sensibles de la manière la plus responsable possible et accroître la compréhension de ces taxons tout en les protégeant des dangers.

4.7.3. Ressources complémentaires

- *ALLEA, All European Academies. The European Code of Conduct for Research Integrity. Revised Edition 2023. Berlin (DE). Juin 2023. https://allea.org/wp-content/uploads/2023/06/European-Code-of-Conduct-Revised-Edition-2023.pdf*
- *CIRAD, INRAE. Avis 8 sur les enjeux éthiques et déontologiques du partage et de la gestion des données issues de la recherche. Février 2016. https://hal.inrae.fr/hal-02796585*
- *COMETS, Comité d'éthique du CNRS. Guide pratique. Pratiquer une recherche intègre et responsable. Mars 2017. https://comite-ethique.cnrs.fr/guide-pratique/*
- *COMETS, Comité d'éthique du CNRS. Charte nationale de déontologie des métiers de la recherche. 26 janvier 2015. https://comite-ethique.cnrs.fr/charte/*
- *European Commission. Ethics and data protection. 14 novembre 2018. https://cache.media.education.gouv.fr/file/2018/54/9/h2020_hi_ethics-data-protection_en_1046549.pdf*

4.8. En résumé

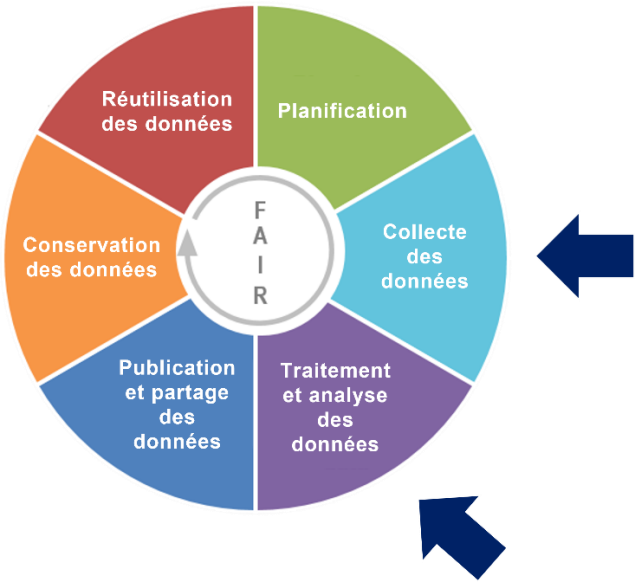
Voici une vidéo (2 min) qui résume les aspects juridiques et éthiques à respecter pour diffuser ses données de recherches dans de bonnes conditions :

https://www.canal-u.tv/chaines/callisto/les-minutes-doranum/la-minute-aspects-juridiques-et-ethiques

5. Création, collecte, traitement et description des données

5.1. La création/collecte des données et leur traitement dans le cycle de vie des données

Ces deux phases successives interviennent au début du projet de recherche et au début du cycle de vie des données.



5.2. Rappel sur la diversité des données de recherche

Cette infographie permet de retrouver les éléments essentiels qui composent les définitions institutionnelles des données de la recherche au niveau européen et international :

https://view.genial.ly/620e2be536a14900172bc0db

Il est difficile de détailler plus avant les étapes de création/collecte et traitement des données car elles sont très liées à la discipline et propres à chaque projet.

5.3. Préparation et documentation des données

Il est impératif de bien préparer et documenter ses données afin d'optimiser le stockage, le partage, l'archivage et la réutilisation.
Il est recommandé de réaliser ce travail de documentation dès l'étape de collecte/ création des données.

Dans le PGD, il faudra indiquer de manière précise quelles méthodes sont utilisées pour recueillir ou produire les données.

Dans le cas de données préexistantes
Indiquer :

- Leur provenance (corpus, archives…)
- Sur quels critères elles ont été sélectionnées
- Les conditions de réutilisations préexistantes de ces données.

  - **Exemples de données réutilisables** :

- Données cartographiques IGN, Géoportail, Open Street Map…
- Données du Réseau de Mesure de la Qualité des Sols
- Données scientifiques partagées et disponibles dans des entrepôts.

Arnould Pierre-Yves, Jacquemot-Perbal Marie-Christine. Guide de bonnes pratiques. Gestion et valorisation des données de la recherche. 23 février 2016. https://hal.science/hal-01275841/

Dans le cas de données produites ou recueillies
Indiquer :

- Le contexte de création
- Les méthodes utilisées
- Les protocoles suivis ou établis
- Les contrôles qualité mis en place.

Exemple de méthodologies/protocoles décrits dans un PGD :

" Les données microclimatiques (p.ex. T°C et humidité relative) seront recueillies à l'aide de capteurs de données environnementales posés directement in-situ par les scientifiques impliqués dans le projet ou bien posés à distance par l'intermédiaire des agents du Réseau National de suivi à long terme des ECOsystèmes FORestiers (RENECOFOR) de l'Office National des Forêts (ONF), auxquels les capteurs seront envoyés par courrier accompagné d'un protocole d'installation in-situ. [...] Les capteurs seront posés in-situ autant que faire se peut dès la première année du projet (2020). Plusieurs types de capteurs seront utilisés : (i) des capteurs HOBO UA-001-64 pour les données de T°C de l'air à 1 m au-dessus du sol (une fréquence d'un enregistrement toute les heures est visée) ; (ii) des capteurs HOBO UA-001-08 pour les données de T°C à 8 cm sous la surface du sol (une fréquence d'un enregistrement toute les 2 heures est visée) ; et (iii) des capteurs TMS4 qui permettent de prendre la T°C à 10 cm au-dessus du sol, au niveau du sol et à 8 cm sous la surface du sol ainsi que l'humidité relative à 8 cm sous la surface du sol (une fréquence d'un enregistrement toute les 15 mins est visée). Les données microclimatiques seront déchargées au moins une fois par an (2021 et 2022), lors des campagnes de terrain, à l'aide des logiciels dédiés HOBOware et Lolly Manager puis analysées sous le logiciel libre R de traitements statistiques. [...]

Concernant la qualité et la conformité de la collecte des données microclimatiques, il est prévu une phase d'intercalibration des capteurs HOBO UA-001-08, HOBO UA-001-64 et TMS4 en conditions contrôlées. L'installation des capteurs de T°C et d'humidité relative du sol, in-situ, suivra un protocole standardisé pour l'ensemble des sites étudiés.

Lenoir Jonathan. DMP du projet "IMPRINT". 12 mars 2020.

https://dmp.opidor.fr/plans/5082/export.pdf

Attention aux données sensibles, personnelles ou confidentielles : prendre les précautions nécessaires afin de respecter les règles juridiques et éthiques en vigueur.

De même, il est recommandé de réaliser ce travail de documentation au fur et à mesure de l'étape de traitement des données.

5.3.1. Exemple de modèle de documentation de données produites

Modèle

Nom de l'échantillon	Selon la nomenclature établie
Méthode de prélèvement	Description, version ou référence bibliographique,
Date de prélèvement	YYYY-MM-DD (Norme 8601)
Personne responsable	Nom|prénom
Géolocalisation du point de collecte/prélèvement	Coordonnées GPS
Système de coordonnées utilisé	Lambert
Type d'échantillon	Liste contrôlée (p. ex. sol, eau)
Autres caractéristiques (champ répétable)	
Conditionnement	
Stockage	Localisation
Commentaire	

Arnould Pierre-Yves, Jacquemot-Perbal Marie-Christine. Guide de bonnes pratiques. Gestion et valorisation des données de la recherche. 23 février 2016. https://hal.science/hal-01275841/

5.4. Les logiciels

Dans de nombreux projets de recherche des logiciels sont utilisés et/ou créés et/ou adaptés. Cette utilisation de logiciels peut intervenir à toutes les étapes de la recherche, dans tous les domaines scientifiques et se révèle essentielle. Pour reproduire une expérience, il est indispensable de connaître avec exactitude la version du logiciel employé.

Le logiciel joue un triple rôle dans la recherche :

- Il sert d'outil dans de nombreux domaines, en traitant efficacement divers types de données pour construire et tester des modèles visant à étayer ou invalider des hypothèses
- Il peut constituer en lui-même un résultat de recherche, en tant que preuve d'existence d'une solution algorithmique efficace à un problème donné

- Il peut être lui-même objet de recherche. En particulier, la communauté scientifique s'intéresse aux modes de développement des logiciels et à la preuve de leurs propriétés, en lien notamment avec la transparence et la confiance dans les traitements informatisés.

En fonction du projet, il s'agira donc de renseigner la version du logicielle utilisée ou de communiquer ses codes sources.

Cette ressource vous permettra de mieux comprendre l'importance et l'utilité des codes sources :

https://doranum.fr/stockage-archivage/les-codes-sources-definition-enjeux-et-preservation_10_13143_7tj2-gw58/

Il est donc fortement recommandé de **documenter soigneusement tout ce qui concerne les logiciels et codes sources** utilisés dans le cadre du projet en parallèle du travail sur les données.

6. Métadonnées

6.1. Définition

Les métadonnées permettent de décrire plus précisément les données.

Ce sont des données sur les données.

Si on s'imagine un jeu de données sous la forme d'une boite de conserve, alors les métadonnées équivalent à l'étiquette qui orne celle-ci et en décrit le contenu (date de fabrication, créateur etc.).

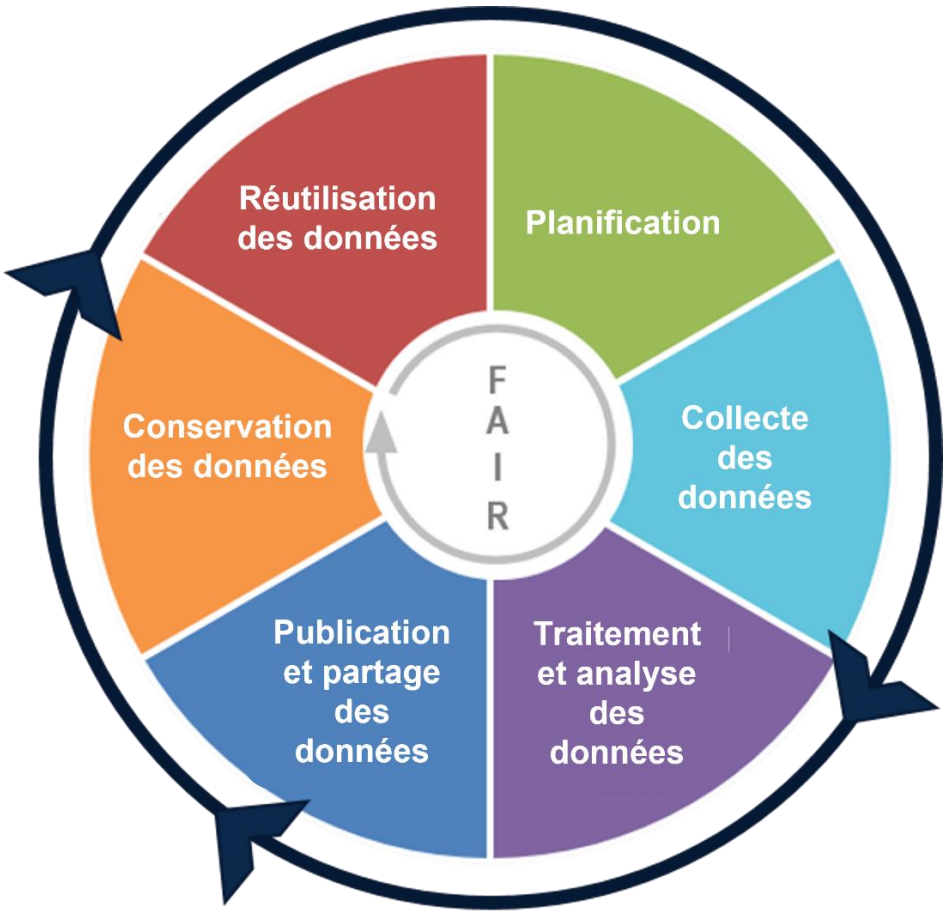
Sans métadonnées

Avec métadonnées

6.2. Les métadonnées dans le cycle de vie des données

Il est recommandé de renseigner les métadonnées au fur et à mesure de l'avancée du projet, avec une attention particulière :

- Au moment du partage des données,
- À l'étape de l'archivage pérenne (des métadonnées spécifiques seront à renseigner).



6.3. Métadonnées embarquées et métadonnées enrichies

Il existe deux types de métadonnées : les métadonnées embarquées et enrichies.

Métadonnées embarquées	Métadonnées enrichies
Elles sont **produites automatiquement** par les appareils (de prise de vue ou de son, de mesure…). C'est typiquement le cas pour la photo ou la vidéo du smartphone. Exemples de métadonnées générées : données GPS, type d'appareil, date, calibrage technique, etc.	Elles sont **ajoutées par l'auteur**. Exemples : mots-clés, sujet, auteur, laboratoire ou organisme, nom du projet, licence, etc.

N'oubliez pas de bien compléter les métadonnées embarquées par des métadonnées enrichies.
L'idéal est de renseigner ces métadonnées au fur et à mesure.

6.4. La différence entre schéma et standard

Il est recommandé d'utiliser un **standard de métadonnées** propre à votre discipline. S'il n'en existe pas, il est possible de créer un **schéma de métadonnées** adapté à vos besoins.

Schéma	Standard
Un **schéma de métadonnées** est une **construction organisée d'informations**. C'est une **liste structurée** et composée **d'éléments descriptifs** reliés entre eux. Pour chaque élément, le schéma définit : • Sa **signification** (par exemple ici se trouve le titre, ici l'auteur et là la date de publication), • Le **type de contenu** attendu (texte, nombre...), • Sa **formulation** (texte libre, format précis, norme à respecter...) • Les **valeurs** qu'il est possible d'attribuer (terme issu d'un thésaurus, choix à faire dans une liste fermée...). Le schéma définit aussi **ce qu'il est possible ou non de faire** avec les éléments. On peut distinguer : • Le **niveau d'obligation** (éléments obligatoires, conseillés, facultatifs) • La possibilité d'**ajouter ou non des éléments** • Des **règles plus spécifiques** (par exemple si tel champ est renseigné, celui d'après doit l'être aussi. Un schéma de métadonnées est donc un plan logique, structuré, qui indique les relations entre les éléments qui le composent.	Un **standard de métadonnées** est décrit par un schéma qui a été adopté comme **modèle** par un ensemble d'utilisateurs : il est **reconnu, normalisé et utilisé à grande échelle**. L'utilisation d'un standard de métadonnées, notamment disciplinaire, est un élément clé pour atteindre un haut degré de **respect des principes FAIR** : • ***Facile à trouver*** : une donnée n'est souvent trouvable que par les éléments de métadonnées indexés dans le moteur de recherche consulté, • ***Accessible*** : notamment grâce aux métadonnées, • ***Interopérable*** : grâce au standard commun qui facilite les traitements informatiques, • ***Réutilisable*** : grâce à la provenance décrite dans les métadonnées, la licence attribuée et au recours à un standard disciplinaire.

Dans le domaine de l'environnement, il existe plusieurs standards de métadonnées qui permettent de renseigner les métadonnées de manière précise et de couvrir les besoins spécifiques de multiples communautés.

Pour trouver quels standards sont utilisés dans votre discipline, vous pouvez interroger vos collègues chercheurs mais aussi les informaticiens et professionnels de l'IST (information scientifique et technique) disponibles localement.
Vous pouvez aussi rechercher dans des répertoires et catalogues tels que :

- Section " Disciplinary Metadata " du Digital Curation Center (DCC)
- " RDA Metadata Standards Catalog " de la Research Data Alliance
- Rubrique " Standards " de FAIRsharing

Pensez également à regarder les informations fournies par les entrepôts de données sur les standards de métadonnées.

6.4.1. Quelques exemples de standards de métadonnées

Standard	Description	URL
Dublin Core	Il s'agit d'un standard international et pluridisciplinaire pour la description des ressources numériques. Il comporte 15 éléments qui constituent le minimum exigé avec des éléments relatifs au contenu et à la propriété intellectuelle.	http://dublincore.org/
DataCite Metadata Schema	Standard lié à l'attribution d'identifiants pérennes DOI.	https://schema.datacite.org/
DwC (Darwin Core)	Standard disciplinaire du domaine de la biodiversité.	http://rs.tdwg.org/dwc/
EML (Ecological Metadata Language)	Standard disciplinaire dans le domaine de l'écologie : il a en grande partie été conçu pour décrire des ressources numériques. Il peut également être utilisé pour décrire des ressources non numériques telles que des cartes papier ou d'autres médias.	https://eml.ecoinformatics.org/
ISO 19115	Standard international pour décrire les informations et les services géographiques.	https://www.iso.org/standard/53798.html

Pour certains projets de recherche, il est parfois nécessaire de créer un schéma de métadonnées plus spécifique, basé sur un standard disciplinaire existant.

6.5. L'enrichissement des métadonnées

Il est fortement recommandé d'utiliser des **vocabulaires contrôlés disciplinaires** utilisés couramment par la communauté scientifique du domaine : par exemple des classifications taxonomiques, la nomenclature internationale des formules chimiques, des thésaurus, lexiques....

Les champs de métadonnées sont à renseigner avec ce vocabulaire contrôlé, notamment au moment du dépôt dans un entrepôt pour le partage.

Le recours à des mots-clés, à du vocabulaire connu, reconnu et utilisé par une communauté scientifique augmente ainsi la capacité des données à être combinées avec d'autres données dans le cadre d'une réutilisation.

6.5.1. Exemples de vocabulaires contrôlés dans le domaine de l'environnement

- **Thésaurus de la biodiversité**
  Ce thésaurus bilingue (français-anglais) structure les concepts-clefs de la biodiversité dans ses composantes écologiques fondamentales et appliquées.

- **GEMET**
  General Multilingual Environmental Thesaurus
  " Thésaurus documentaire multilingue développé et publié par l'Agence européenne pour l'environnement.
  C'est la compilation de plusieurs vocabulaires multilingues. [...] Il a pour but de définir une terminologie générale pour le domaine de l'environnement. "
  (https://fr.wikipedia.org/wiki/General_Multilingual_Environmental_Thesaurus).

- **ThesauForm - T-SITA**
  Thésaurus sur les caractéristiques des invertébrés du sol.

Propose une liste non exhaustive, structurée sémantiquement, de caractéristiques et de préférences écologiques.

- **TAXREF**
  Référentiel taxonomique Faune, Flore et Fonge de France métropolitaine et d'Outre-Mer.
- **Catalogue of Life**
  Base de données de taxonomie des espèces. Le catalogue contient des informations sur les noms, les relations et les distributions de plus de 1,6 millions d'espèces.

Cette ressource brosse un panorama des principaux vocabulaires contrôlés dans le domaine de l'Environnement :

https://doranum.fr/environnement/principaux-vocabulaires-controles-dans-le-domaine-de-lenvironnement_10_13143_sd1c-9a43/

Exemple de métadonnées dans un PGD :

DMP du projet "SNO KARST".

Utilisation des standards de métadonnées ISO 19115 et CSW, en respectant la directive INSPIRE.

2. Documentation et qualité des données

2a. Quelles métadonnées et quelle documentation (par exemple méthodologie de collecte et mode d'organisation des données) accompagneront les données ?

Métadonnées et standards

Les métadonnées peuvent être regroupées en plusieurs catégories :

- description de la thématique des données du SNO KARST ;
- description des observatoires ;
- description des personnes ;
- description des stations de mesure ;
- description des paramètres mesurés ;
- description des instruments de mesure ;
- description des procédures de prélèvement pour les données de chimie, hydrobiologie, bactériologie et chimie des précipitations.

On utilise tant que possible des référentiels ou des vocabulaires contrôlés pour renseigner les métadonnées.

Diffusion des métadonnées sur le portail de données SNO KARST

L'intégralité des métadonnées sont associées aux données publiques lors de leur téléchargement depuis le portail des données du SNO KARST. Elles sont exportées dans un fichier CSV associé aux fichiers de données.

Diffusion des métadonnées sur le catalogue de l'OSU OREME

Les métadonnées qui concernent les données produites par les observatoires et partenaires du SNO KARST sont également consultables et téléchargeables sur le catalogue des données de l'OSU OREME . Dans le catalogue, les métadonnées sont agrégées à 2 niveaux de granularité :

- pour **l'ensemble** des séries de données produites par les observatoires et les partenaires du SNO KARST ;
- pour chaque **jeu de données** produit par les observatoires et les partenaires du SNO KARST : un jeu de données correspond à l'ensemble des séries mesurées sur un bassin hydrologique / hydrogéologique donné, pour une catégorie de paramètres donnée. Exemple : *données chimiques mesurées sur le bassin du Lez (observatoire MEDYCYSS)* .

Les fiches de métadonnées du catalogue sont générées automatiquement à partir des métadonnées stockées dans la base de données : un script R disponible en ligne extrait les métadonnées, les formate et les envoie au catalogue (librairies R geometa et geonapi). Le catalogue est mis à jour régulièrement (au moins 1 fois par an).

Les métadonnées du catalogue (outil GeoNetwork) respectent la norme ISO 19115 et la directive européenne INSPIRE, et sont diffusables et interrogeables selon le standard CSW (Catalogue Service for the Web) de l'Open Geospatial Consortium.

Un moissonnage du catalogue de l'OSU OREME par le portail data.gouv.fr est en cours de réalisation.

Indexation des métadonnées sur Datacite

Enfin, les métadonnées de chaque jeu de données produit par les observatoires et les partenaires du SNO KARST sont associées aux DOI de ces jeux, sous le format Datacite (voir **5. Partage des données et conservation à long terme**).

https://dmp.opidor.fr/plans/12770/export.pdf

6.6. Pour résumer

Les métadonnées sont utiles pour :

- Comprendre l'origine des données et leur contexte de création ou de collecte
- Améliorer le moissonnage (extraction automatique) par les machines (moteur de recherche)
- Garantir l'interopérabilité
- Connaitre les conditions de réutilisation et de partage des données
- Accéder à des informations très utiles lorsqu'on ne peut pas partager ses données (embargo, accès restreint) ou lors du retrait des données (données obsolètes, etc.).

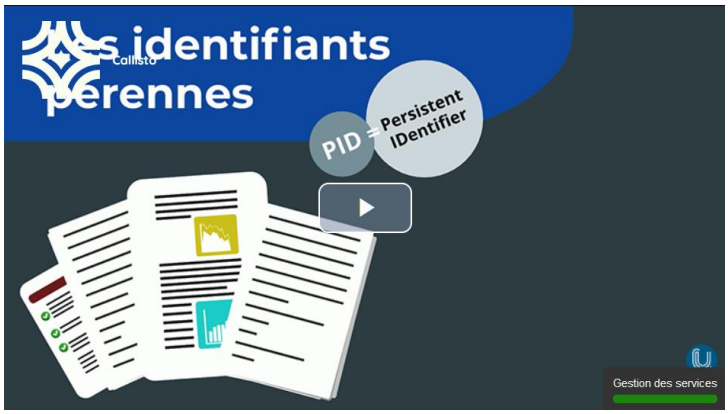
7. Identifiants pérennes

7.1. Les identifiants pérennes (PID) dans le cycle de vie des données

C'est au moment du partage des données qu'un identifiant pérenne est attribué aux données.

7.2. La minute PID

Cette vidéo (5 min) vous permet de mieux comprendre ce que sont les identifiants pérennes ou PID (pour Persistent IDentifiers) :



https://www.canal-u.tv/chaines/callisto/la-minute-identifiants-perennes-0

7.3. Des identifiants pérennes pour les données

Il est recommandé d'attribuer un identifiant pérenne à chacun des jeux de données.

Le plus utilisé est le DOI.

Un identifiant pérenne facilite le suivi, la localisation, l'accès et la citation des données lors de leur publication ou à des fins de réutilisation.

Le plus souvent, **un identifiant pérenne est attribué aux données lors du dépôt dans un entrepôt de manière automatique** !

Un chercheur ne peut pas faire une demande de DOI à titre individuel.
Seuls les institutions et centres de données peuvent obtenir des DOI directement en contactant une agence membre de DataCite, l'Inist-CNRS étant l'agence d'attribution de DOI pour la France.
https://opidor.fr/identifier/

Cette présentation vous propose un focus sur l'identifiant DOI (pour les productions scientifiques) :

https://doranum.fr/identifiants-perennes-pid/zoom-doi_10_13143_j5xt-6j41/

À noter : Depuis 2021, un DOI peut être attribué à un plan de gestion de données.

7.4. Un identifiant pérenne pour les logiciels

Il existe un identifiant pérenne unique dédié aux logiciels : SWHID.

Cette présentation vous propose un focus sur cet identifiant :

https://doranum.fr/identifiants-perennes-pid/zoom-swhid_10_13143_3qqg-yx41/

7.5. Des identifiants pérennes pour les auteurs

Avoir un identifiant auteur permet :

- De faire le lien avec ses productions scientifiques
- D'être bien identifié et cité.

Le plus utilisé est ORCID, un identifiant international, neutre et indépendant.

Cette présentation vous propose un focus sur l'identifiant ORCID :

https://doranum.fr/identifiants-perennes-pid/zoom-orcid_10_13143_c6rx-9w77/

À noter : Le “Plan de gestion des données” a été intégré par ORCID dans la liste des “Types de travaux” (catégorie de travaux “Autre”) qui peuvent être associés à une personne.
ORCID peut ainsi récupérer automatiquement les informations du PGD si un DOI lui a été attribué.

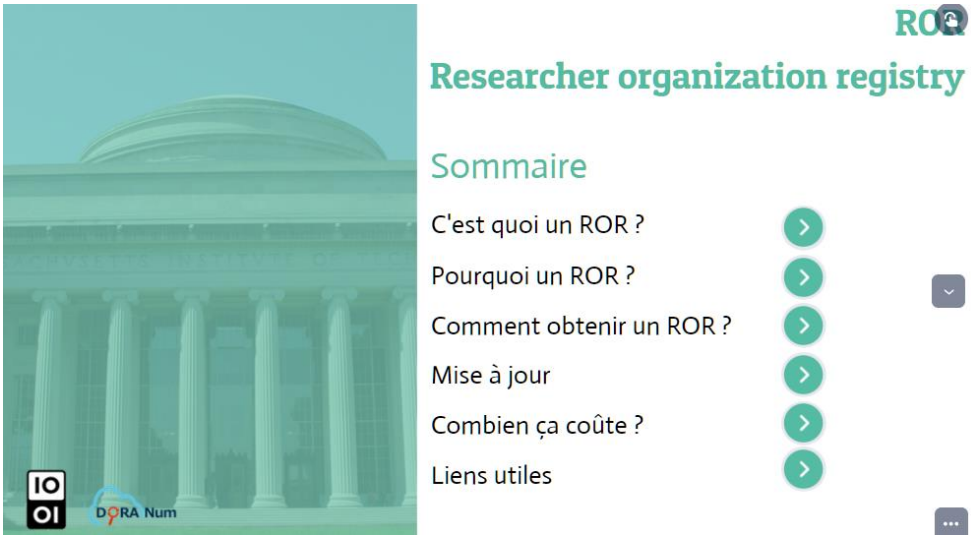
7.6. Des identifiants pérennes pour les institutions

Avoir un identifiant pour les institutions permet :

- D'identifier de manière unique les affiliations de chercheurs et les résultats de recherche
- Une découverte et un suivi plus efficaces des résultats de la recherche par les institutions.

Le plus connu est ROR (Research Organization Registry), un identifiant universel, international et indépendant.

Cette présentation vous propose un focus sur l’identifiant ROR :



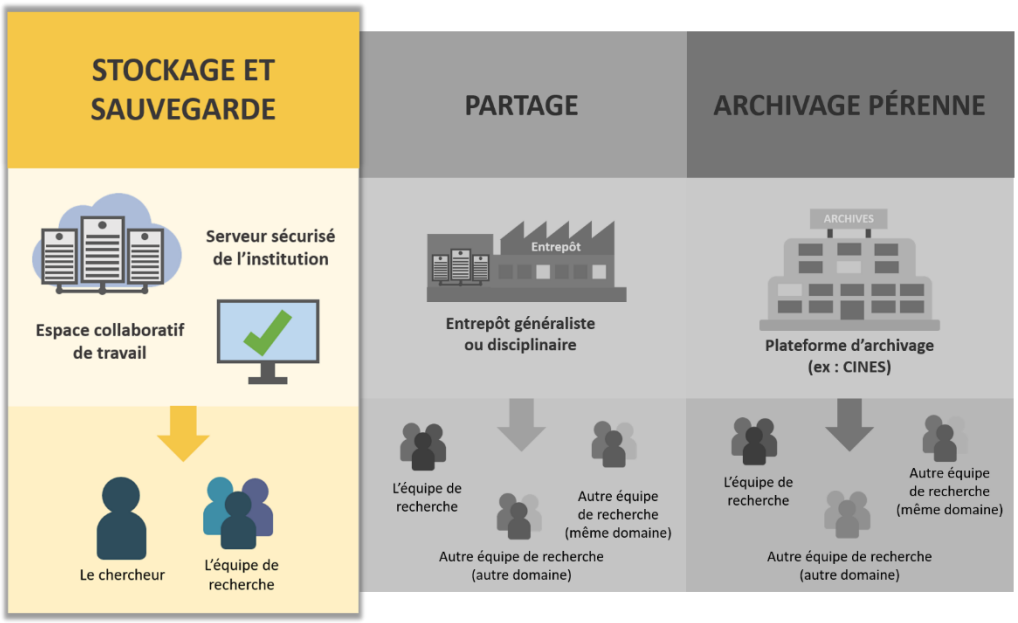
https://doranum.fr/identifiants-perennes-pid/zoom-sur-ror_10_13143_h5zy-bn73/

8. Stockage et sauvegarde des données durant le projet

8.1. La 1ère étape : le stockage et la sauvegarde des données durant le projet

La gestion des données de recherche du projet doit être réfléchie et organisée différemment en fonction de l'étape à laquelle on se situe, pendant et après le projet.

La première étape concerne le stockage sécurisé et la sauvegarde des données durant toute la durée du projet.



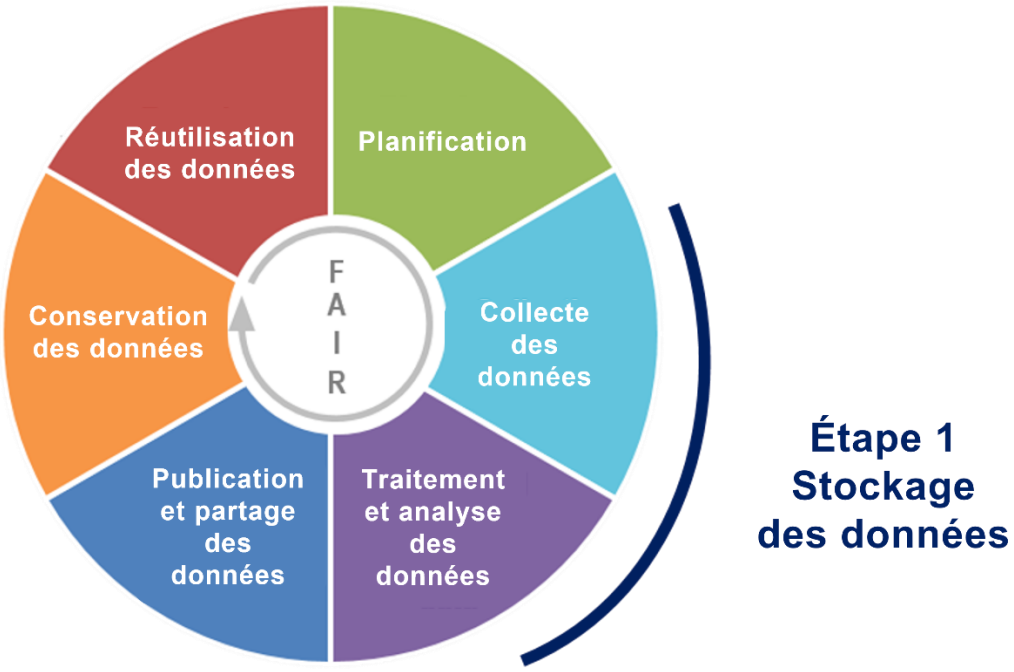
https://doranum.fr/stockage-archivage/stockage-partage-archivage-quelles-differences_10_13143_5dax-qp58/

Les objectifs sont de :

- Garantir la sécurité des données
- Faciliter l'accès pour l'ensemble des collaborateurs du projet.

8.2. Le stockage sécurisé des données dans le cycle de vie des données

Cela concerne la première moitié du cycle de vie des données.



Il est très important de prévenir la perte et la dégradation des données pendant le projet du fait :

- Du matériel
- Du logiciel utilisé
- Du format de fichier
- De la perte de la signification du contenu
- D'un mauvais étiquetage
- D'un manque de rigueur dans le nom des fichiers

Quidoz Marie-Claude. Atelier « Carnets de terrain électroniques ». Sécuriser les données produites par les carnets de terrain électroniques. 28-29 mars 2018.
https://oreme.org/app/uploads/Quidoz_Atelier2018.pdf

8.3. Mesures de sauvegarde à mettre en place

Une sauvegarde efficace signifie qu'il faut dupliquer et stocker les données à différents endroits sur différents supports selon une temporalité pertinente pour le projet.

L'idéal est d'appliquer la règle du 3-2-1, ce qui veut dire :

- Garder 3 exemplaires des données,
- Sur 2 supports ou technologies différents (les tester régulièrement et migrer les fichiers sur un autre support si nécessaire),
- Dont 1 se trouve hors site.

Cette vidéo (2 min) détaille la règle du 3-2-1 :

https://www.canal-u.tv/chaines/callisto/la-sauvegarde-3-2-1

Dans tous les cas, il faut organiser et planifier ces sauvegardes en veillant à bien gérer les versions. À chaque point d'étape du projet, sélectionner les données à sauvegarder, à supprimer. Les différents états des données sont conservés en corrélation avec les différentes étapes de traitement, ce qui permet de revenir à une version antérieure si besoin.

Cela nécessite aussi de définir un hébergement et une politique de sauvegarde adaptés aux besoins du projet concernant les spécificités de stockage des données (par exemple en cas de données sensibles, de grosse volumétrie...). Cela peut être sur des serveurs locaux (machines virtuelles), un cloud institutionnel avec accès sécurisé...

Il est recommandé d'éviter au maximum les outils du type One Drive, Google Drive, Dropbox, etc.

Ne pas hésiter à **se rapprocher de son établissement afin de connaitre les espaces de stockage sécurisés mis à disposition**.

8.4. Nommage des fichiers

La fiabilité d'accès passe par un nommage unique et précis des fichiers de données. Il est préférable d'utiliser une nomenclature explicite et commune à tous les partenaires du projet.

Bonnes pratiques :

- 30 caractères maximum
- Lettres majuscules et minuscules, chiffres, tirets (- ou _)
- Acronyme ou numéro du projet
- Description brève du contenu
- Dates au format : AAAA-MM-JJ ou AAAAMMJJ
- Numéro de version le cas échéant

À éviter :

- Pas de caractères spéciaux ou accentués du type ùéàç+'@°[ ] :</* »& !$ …
- Séparateurs : pas d'espace, pas de point, pas de mots vides
- Pas de dénomination vague : divers, autres, à classer…

8.4.1. Exemples

La nomenclature des échantillons doit contenir a minima les éléments suivants :

- La localisation (ex : le nom du lieu, de la parcelle ou de la station)
- La date de prélèvement (yyyymmdd ou yyyy-mm-dd)
- Le type d'échantillon (ex : sol, terre, eau…)

Échantillon d'eau (W) prélevé à Joeuf Abattoir (JOAB) le 7 mai 2015 dans le cadre du projet Mobised : JOAB_20150507_W

Il peut aussi être utile de signifier d'autres éléments comme par exemple une modalité de prélèvement ou d'analyse.

Échantillon de particules en suspension (SPM) après centrifugation de terrain (FC) à Joeuf Abattoir (JOAB) le 7 mai 2015 : JOAB_20150507_SPM_FC_1

Relevé de biomasse de la luzerne effectué par Pierre Martin (PM) dans le cadre du projet Multipolsite (MPS) sous format csv : MPS_2011-05-30_biomasse-v1.csv

Arnould Pierre-Yves, Jacquemot-Perbal Marie-Christine. Guide de bonnes pratiques. Gestion et valorisation des données de la recherche. 23 février 2016. https://hal.science/hal-01275841/

8.5. Formats de fichiers

8.5.1. Comment choisir le format d'un fichier

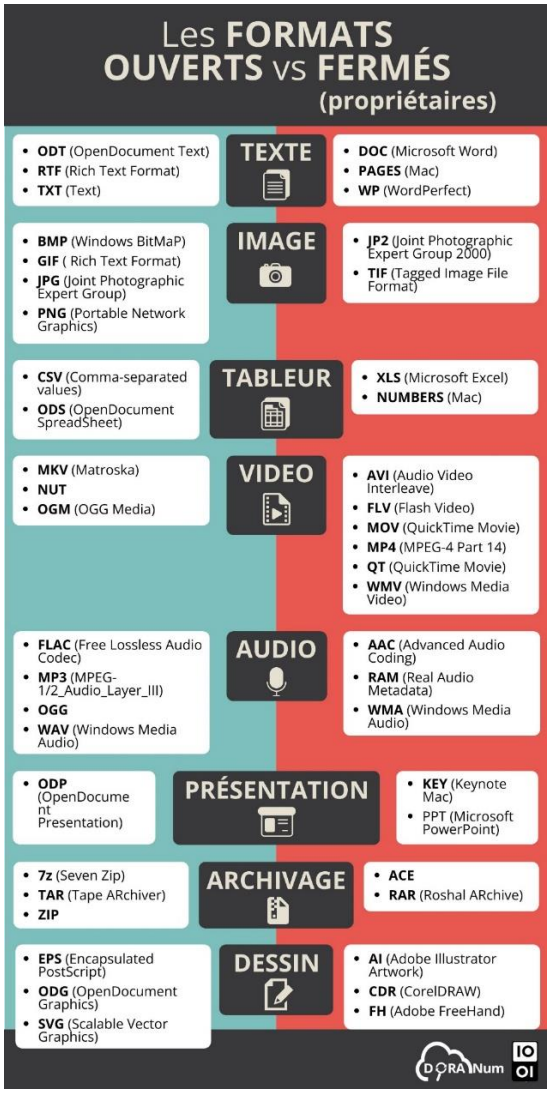
Le choix d'un format peut être guidé par :

- Les recommandations de son institution,
- Les usages de la communauté scientifique de la discipline,
- Les logiciels ou équipements utilisés.

L'idéal est d'opter pour des formats de fichiers les plus ouverts possible (non propriétaires), standardisés et pérennes, par exemple :

- Privilégier .csv à .xls
- Privilégier .odt à .doc
- Privilégier .jpg à .tif

Voici Une infographie non exhaustive sur les formats fermés et de leurs équivalents ouverts :



https://doranum.fr/stockage-archivage/quiz-format-ouvert-ou-ferme_10_13143_mcwq-qs64/

Dans tous les cas, **il faut mentionner dans le PGD quels formats seront utilisés**.

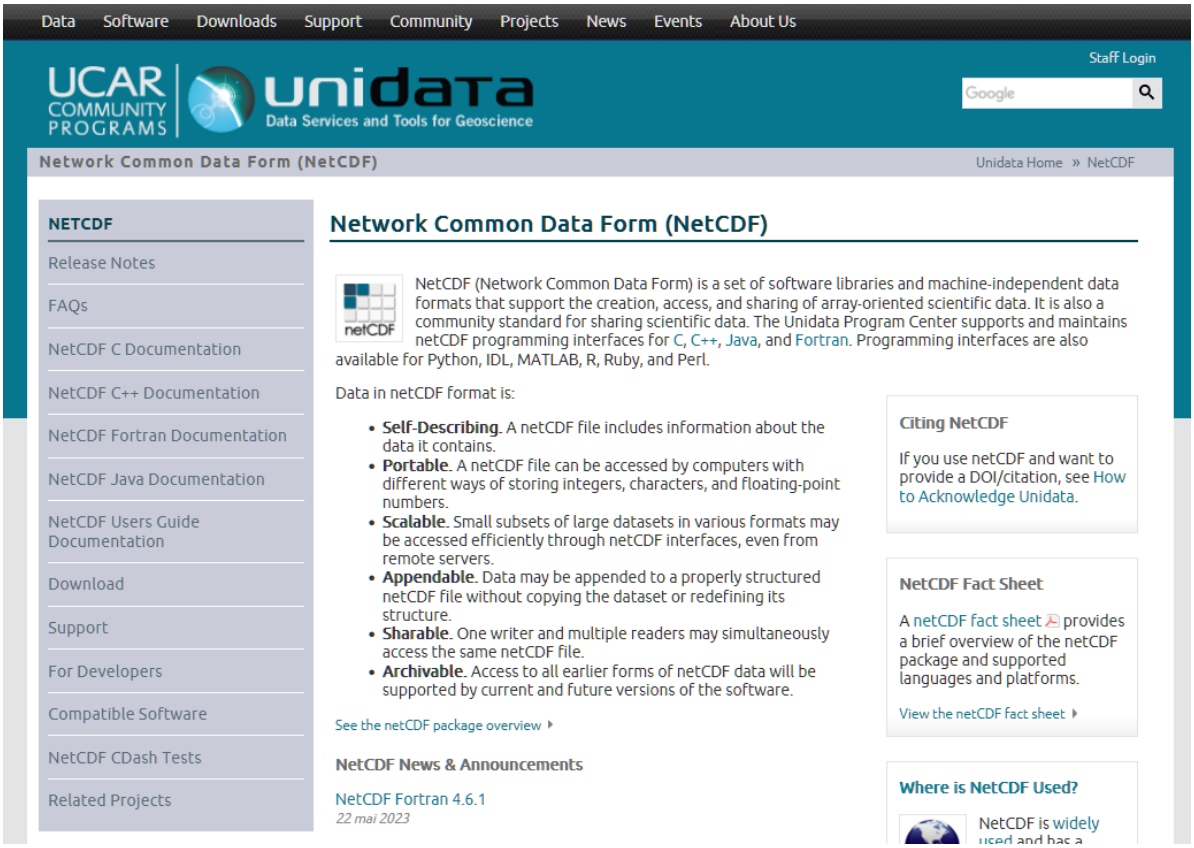
8.5.2. Exemples de formats utilisés dans le domaine de l'environnement

NetCDF

Le format NetCDF (Network Common Data Form) est un format ouvert, auto-documenté et très utilisé en sciences de l'environnement. C'est aussi un modèle de représentation des données qui s'applique bien pour structurer des données qui évoluent en fonction de certaines dimensions (temps, altitude, profondeur, latitude, longitude etc.). Il est très bien adapté et utilisé, par exemple pour représenter et formater des données de type profils verticaux, des séries temporelles, des trajectoires, ou encore des surfaces maillées en 2D. Ce format est dit "auto-descriptif" car les métadonnées sont insérées dans l'entête du fichier, avec les données elles-mêmes. En ce sens il permet de ne pas avoir besoin d'un fichier de description complémentaire. On peut ainsi décrire de manière assez précise les données du fichier, par exemple en insérant les unités de mesure des paramètres mesurés, la licence de diffusion, les propriétaires, etc., ainsi que l'organisation des données.
Il est recommandé par l'e-Infrastructure de recherche Data Terra car il procure un cadre de standardisation international qui permet l'interopérabilité, la pérennité et la réutilisation des données.

Hadrossek Christine, Janik Joanna, Libes Maurice, Louvet Violaine, Quidoz Marie-Claude, Rivet Alain, Romier Geneviève. Atelier Données. Guide de bonnes pratiques sur la gestion des données de la recherche. Version 2.0. 23 août 2023. https://mi-gt-donnees.pages.math.unistra.fr/guide/04-traiter.html

Libes Maurice, Jeannaud Viêt. netCDF, format de fichier interopérable pour la science ouverte (Version 1). Callisto Formation. Novembre 2022. https://doi.org/10.60538/NETCDF-INTRODUCTION-CALLISTO



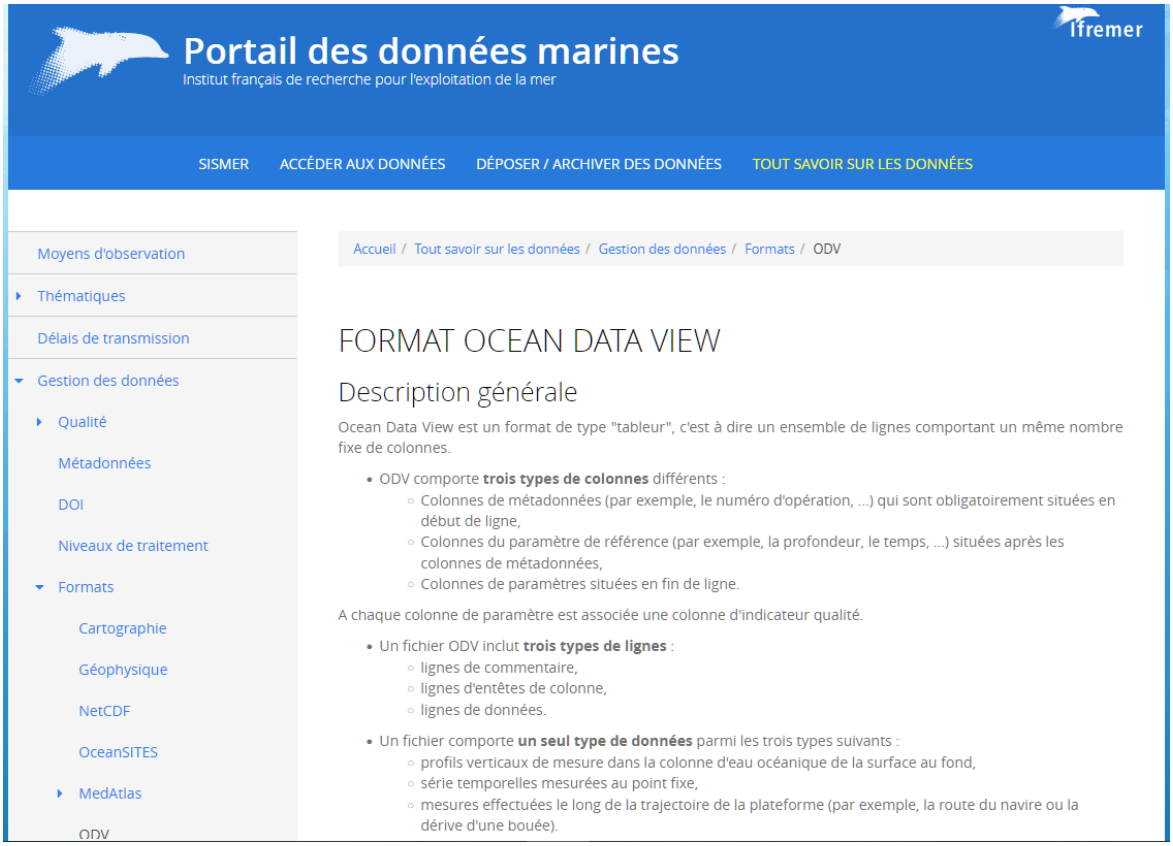
https://www.unidata.ucar.edu/software/netcdf/

ODV

" Le format ODV (Ocean Data View) est également un format standard ouvert intéressant. C'est un format de type "tableur", ensemble de lignes comportant un nombre fixe de colonnes qui se rapproche d'un format CSV, composé de colonnes de données séparées par des virgules (ou tout autre séparateur), à cette différence près que le format ODV permet l'insertion d'un entête assez riche permettant de placer des métadonnées en début de fichier. On trouvera un exemple sur le Portail des données marines.

Le format de données ODV permet un stockage dense et un accès très rapide aux données. De grandes collections de données comprenant des millions de stations peuvent être facilement entretenues et explorées sur des ordinateurs de bureau. "

Hadrossek Christine, Janik Joanna, Libes Maurice, Louvet Violaine, Quidoz Marie-Claude, Rivet Alain, Romier Geneviève. Atelier Données. Guide de bonnes pratiques sur la gestion des données de la recherche. Version 2.0. 23 août 2023. https://mi-gt-donnees.pages.math.unistra.fr/guide/04-traiter.html



https://data.ifremer.fr/Tout-savoir-sur-les-donnees/Gestion-des-donnees/Formats/ODV

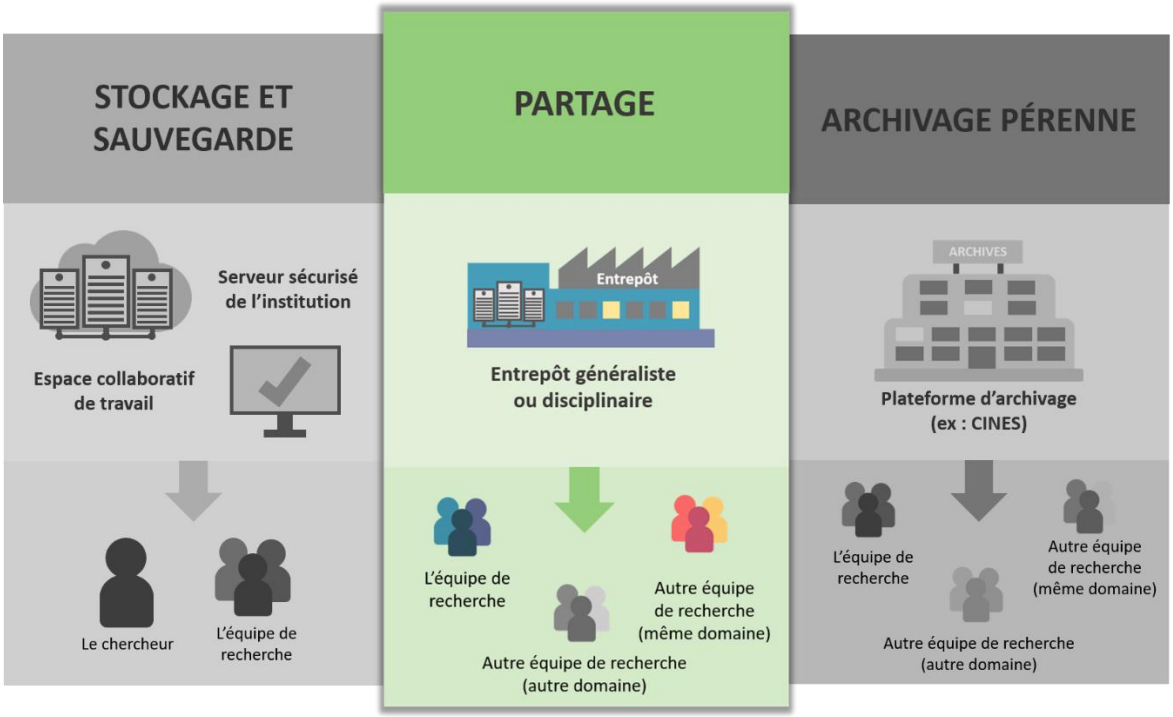
9. Dépôt des données dans un entrepôt pour le partage

9.1. L'étape du partage des données

La gestion des données de recherche du projet doit être réfléchie et organisée différemment en fonction de l'étape à laquelle on se situe, pendant et après le projet.

Il est important de partager ses données en les déposant dans un entrepôt :

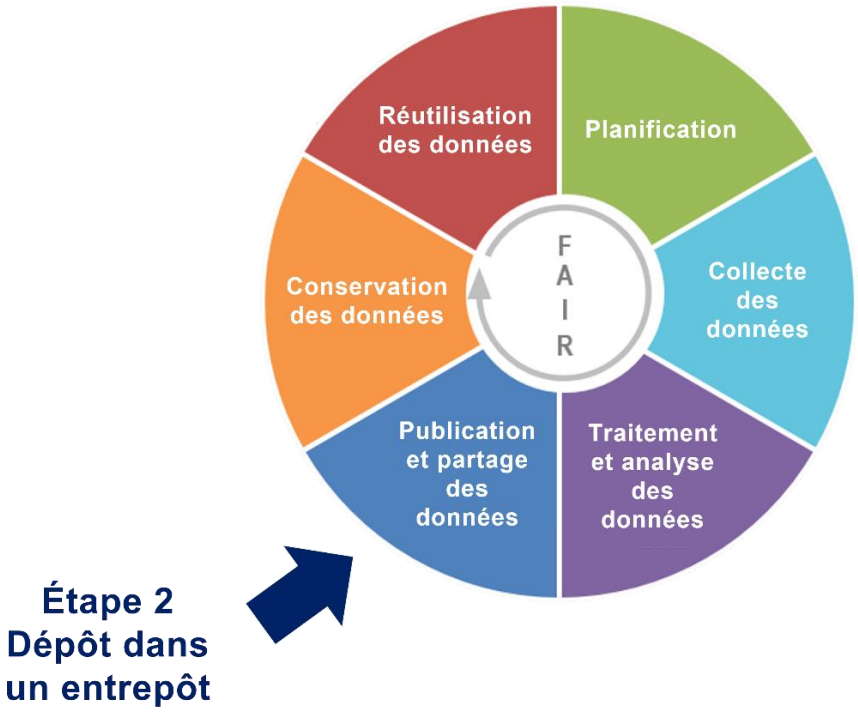
- Afin de les rendre accessibles facilement (principes FAIR)
- Afin de permettre leur réutilisation (principes FAIR)
- Sur le court et le moyen terme (5 à 10 ans)
- Pour des chercheurs du même domaine, ou d'un autre domaine.



https://doranum.fr/stockage-archivage/stockage-partage-archivage-quelles-differences_10_13143_5dax-gp58/

9.2. Le partage des données dans le cycle de vie des données

Il intervient le plus souvent à la fin du projet de recherche : le partage des données est complémentaire de la publication scientifique.



9.3. Dépôt des données dans un entrepôt

Le but est de partager les données de recherche du projet dans des conditions optimales.

Pour cela, il faut **déposer les données et les métadonnées associées** dans l'entrepôt choisi, sans oublier les **codes sources** (scripts) nécessaires à la lecture et à la compréhension de celles-ci.

Les entrepôts permettent de stocker des données de recherche, d'y accéder et de les réutiliser. Il existe différentes catégories d'entrepôts : les entrepôts propres à un éditeur, à une discipline, à une institution ou multidisciplinaires.

Il est recommandé de déposer vos données de préférence dans un **entrepôt disciplinaire** ou **institutionnel**.

9.4. Préparer ses données selon les principes FAIR

Pour que le partage des données soit efficace, il faut les préparer selon les principes FAIR, et ce, quelle que soit l'ouverture des données envisagée, totale ou partielle, ouverte ou sécurisée...

Voici une check-list pour bien préparer ses données :

1) **Sélectionner les données à partager**
   L'ensemble des données n'est pas nécessairement destiné à être partagé. L'équipe de recherche doit sélectionner les jeux de données qu'elle souhaite partager et, pour chacun d'eux, définir les modalités d'accès.

2) **Formats de données**
   - Vérifier la compatibilité et l'interopérabilité des formats de données,
   - Migrer si nécessaire vers un format adapté, le plus ouvert possible.

3) **Codes sources**

Préparer si nécessaire les codes sources (ex : scripts) qui permettront de lire et traiter les données.

4) **Métadonnées**

Compléter et enrichir les métadonnées en fonction de l'entrepôt choisi :

- Si ce n'est pas déjà fait, choisir un standard de métadonnées,
- S'il n'en existe pas d'adapté, créer un schéma de métadonnées,
- Compléter les champs pour chaque jeu de données, suivant le standard adopté.

9.5. Choix de l'entrepôt

Pour choisir le bon entrepôt, appuyez-vous sur les **pratiques de votre communauté scientifique** !

1) **Entrepôt disciplinaire**

Certaines disciplines sont bien organisées pour la gestion des données, et proposent des entrepôts disciplinaires spécifiques. Vous pouvez ainsi vous appuyer sur un ensemble de bonnes pratiques et de standards bien définis, ce qui facilitera grandement la préparation, la documentation et le dépôt des données.
Si aucun entrepôt n'est recommandé par votre communauté, vous pouvez identifier celui qui pourrait convenir à vos besoins grâce aux annuaires dédiés comme re3data, OAD, OpenDOAR, FAIRsharing...

2) **Entrepôt institutionnel**

Si aucun entrepôt disciplinaire ne convient, il est conseillé de déposer vos données dans l'entrepôt de votre institution, s'il existe.

3) **Recherche Data Gouv**

Si aucun entrepôt disciplinaire ou institutionnel ne correspond à vos besoins, il est recommandé de **déposer dans l'entrepôt national pluridisciplinaire** *Recherche Data Gouv*. Il permet à la communauté scientifique française de déposer et d'ouvrir ses données de recherche.

Développé à partir de l'application web open source Dataverse, l'entrepôt *Recherche Data Gouv* est organisé en espaces institutionnels de publication et de signalement des données des établissements qui souhaitent participer.

Si vous avez des difficultés à choisir un entrepôt, vous pouvez vous faire aider
- par les services de documentation / professionnels de l'IST de votre institution
- par les équipes des ateliers de la donnée
- par les services de soutien locaux. Pour les identifier, consulter la page SOS-PGD du site Couperin.

9.5.1. Exemples d'entrepôts disciplinaires

Il est conseillé de privilégier le dépôt de vos données dans un **entrepôt disciplinaire**.

DRYAD : Entrepôt en Sciences de la Vie, Agronomie, Géosciences, Anthropologie et Sciences comportementales.

GBIF (Global Biodiversity Information Facility) : Entrepôt international certifié sur la Biodiversité.

PANGAEA : Entrepôt certifié de données géoréférencées issues de la recherche sur le système terrestre.

Data.InDoRES (Inventaire des Données de la Recherche en Environnement et Sociétés).

Portail des données marines de l'Ifremer

ORDaR (OTELo Research Data Repository) : Entrepôt de l'Observatoire Terre et Environnement de Lorraine.

9.5.2. Critères pour choisir un entrepôt

Si le choix de l'entrepôt vous revient, voici une liste de critères pour vous aider :

- Privilégier un entrepôt disciplinaire

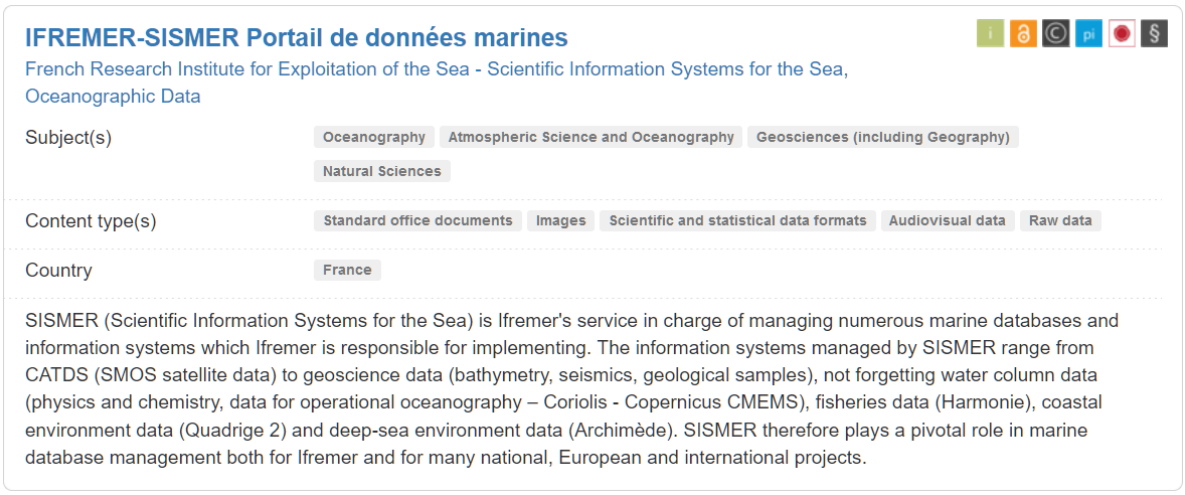
- Choisir un entrepôt en fonction des types de données acceptés
- Vérifier la qualité des métadonnées requise
- Vérifier si l'entrepôt envisagé est certifié
- Vérifier si l'entrepôt envisagé est labellisé "entrepôt de confiance"
- Vérifier la pérennité proposée par l'entrepôt
- Voir si le dépôt de données dans l'entrepôt envisagé génère un identifiant pérenne
- Vérifier si l'entrepôt envisagé permet la gestion des versions si tel est votre besoin.

Pour en savoir plus, consultez cette ressource :

https://doranum.fr/depot-entrepots/criteres-pour-choisir-entrepot_de_donnees_10_13143_zgpb-9449/

9.5.3. Exemple de recherche dans l'annuaire re3data

La recherche s'effectue grâce à des filtres. La liste des résultats apparait sous forme de brèves fiches descriptives présentant, pour chaque entrepôt, le sujet, le type de contenu, le pays, un petit résumé et des icônes symbolisant les critères auxquels répond l'entrepôt.



https://www.re3data.org/repository/r3d100012965

Dans notre exemple, l'entrepôt IFREMER-SISMER Portail des données marines répond aux critères suivants :

- Informations complémentaires fournies
- Libre accès aux données
- Conditions d'utilisation et licence fournis
- Génération d'un DOI
- Certification
- Politique de l'entrepôt fournie

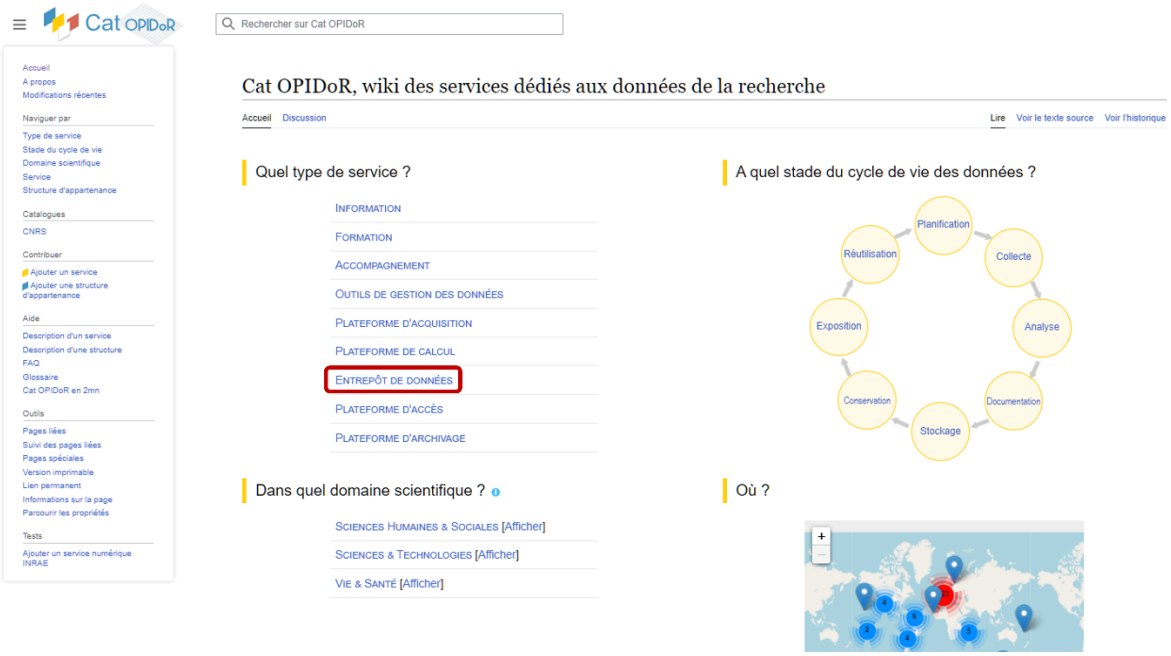
9.5.4. Recherche d'un entrepôt dans Cat OPIDoR

Pour trouver un entrepôt de données français, vous pouvez aussi utiliser l'outil **Cat OPIDoR** (Catalogue pour une Optimisation du Partage et de l'Interopérabilité des Données de la Recherche).

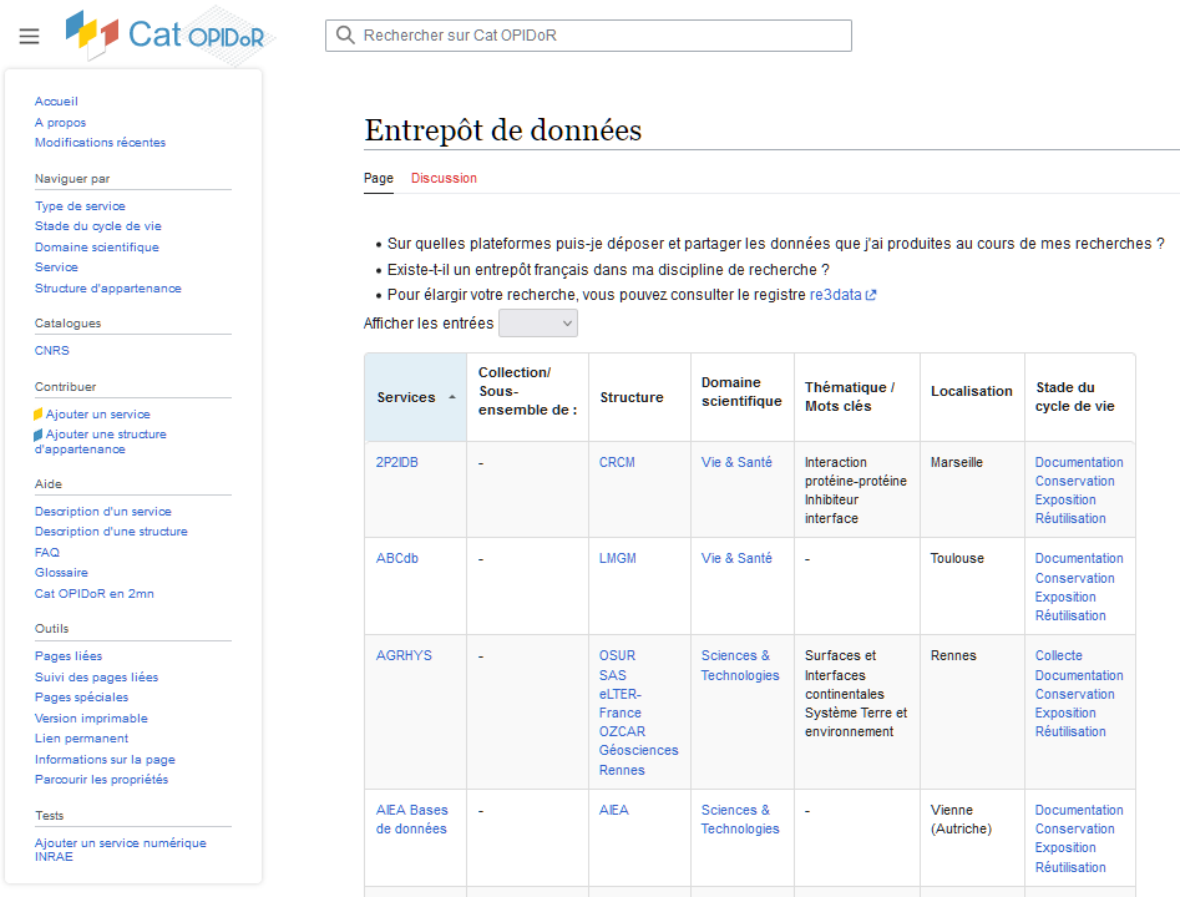
Proposé sous forme d'un wiki, cet outil collaboratif, gratuit et ouvert à tous permet de repérer et ajouter des services utiles dans le cadre d'un projet de recherche.

Cat OPIDoR recense et décrit les **services français** dédiés aux données scientifiques et présente, par domaine scientifique : des sites d'information, de formation, des outils de gestion, des plateformes... pour accompagner les chercheurs sur l'ensemble des étapes clés de la gestion, collecte, stockage, conservation et ouverture des données.

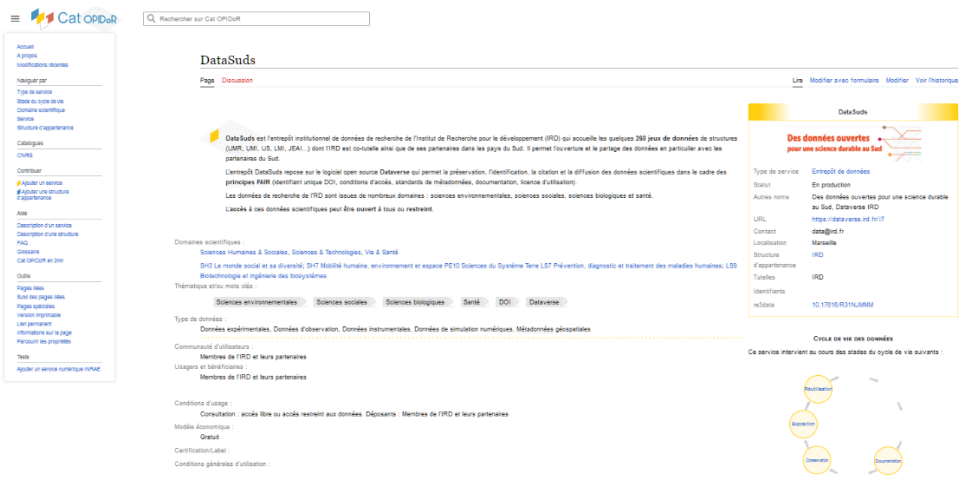
Les entrepôts de données figurent parmi les types de services recensés.



Écran d'accueil de Cat OPIDOR avec les 4 modes de recherche : par type de service, par domaine scientifique, par stade du cycle de vie des données, par localisation



Page "Entrepôt de données" avec présentation des résultats sous forme de tableau



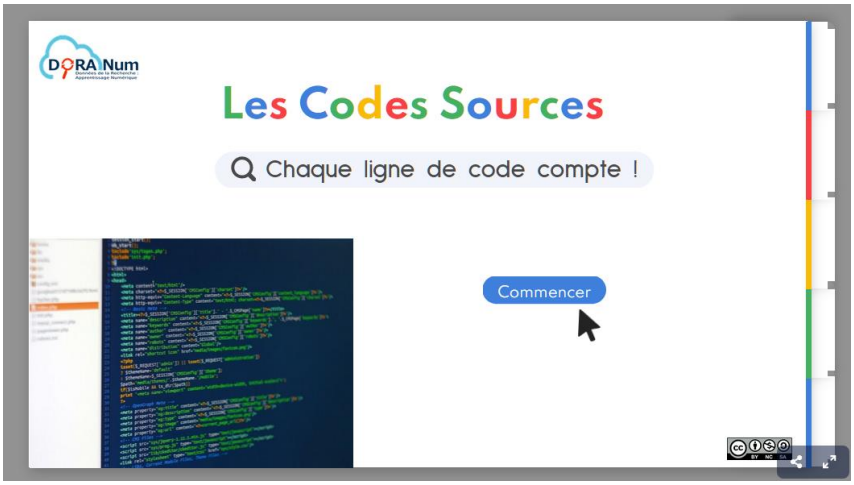
Fiche descriptive de l'entrepôt de l'IRD DataSuds

Astuce : vous pouvez éventuellement trouver un entrepôt de données à votre convenance en effectuant une recherche sur le sujet que vous traitez via Google Dataset Search ou OpenAIRE Explore. Ces moteurs de recherche sont conçus pour chercher directement des jeux de données.

9.6. Logiciels et codes sources

Software Heritage et l'archive ouverte HAL assurent l'accessibilité et la conservation des codes sources en libre accès, tout en gérant les versions.

Pour en savoir plus, consultez la ressource ci-dessous :



https://doranum.fr/stockage-archivage/les-codes-sources-definition-enjeux-et-preservation_10_13143_7tj2-gw58/

9.7. Pour résumer

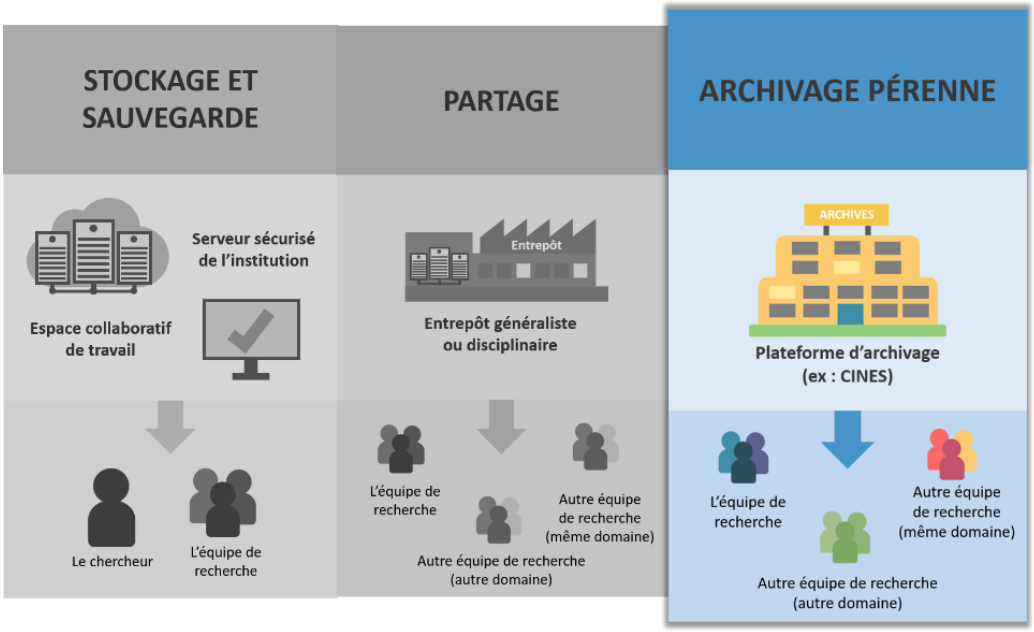
Voici une vidéo (4 min) qui synthétise les points essentiels à connaître pour un partage optimal des données de recherche grâce à leur dépôt dans un entrepôt :

https://www.canal-u.tv/chaines/callisto/deposer-ses-donnees-de-recherche-pourquoi-quoi-quand-ou-et-comment

10. Archivage pérenne

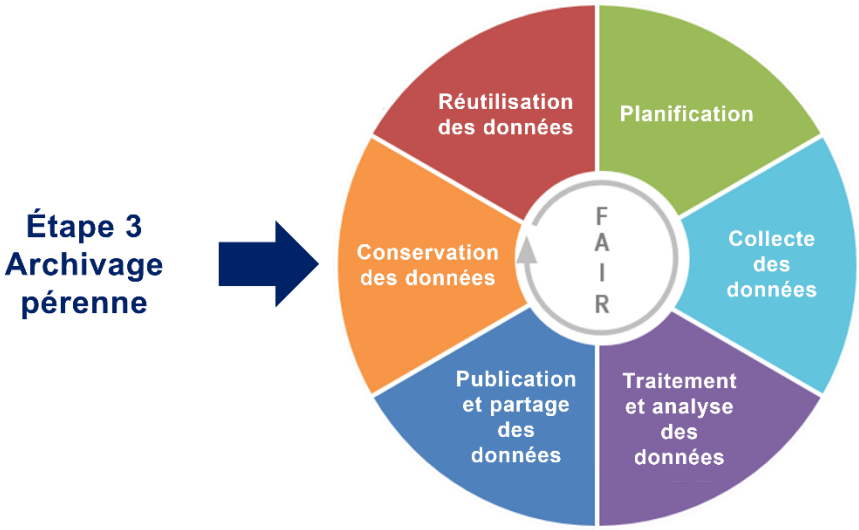
10.1. L'ultime étape : l'archivage pérenne

La gestion des données de recherche du projet doit être réfléchie et organisée différemment en fonction de l'étape à laquelle on se situe, pendant et après le projet. Contrairement au stockage, à la sauvegarde et au dépôt des données dans un entrepôt pour le partage, l'étape de l'archivage pérenne ne concerne pas tous les projets : il est à réaliser uniquement en cas de nécessité avérée.



https://doranum.fr/stockage-archivage/stockage-partage-archivage-quelles-differences_10_13143_5dax-qp58/

10.2. L'archivage pérenne dans le cycle de vie des données



10.3. Définition et périmètre

L'archivage pérenne ne concerne en général qu'une partie des données produites par un projet. Pour certains projets, il n'est d'ailleurs pas nécessaire de prévoir d'archivage pérenne.

En effet, la question de l'archivage pérenne se pose uniquement pour les données présentant une **valeur scientifique reconnue** par la communauté d'où elles proviennent et qui nécessitent une conservation pour **au moins 30 ans**.

C'est une **opération coûteuse qui nécessite un budget alloué**. Elle **se décide à l'échelle du laboratoire ou de l'institution et non pas à l'échelle du chercheur**.

Concrètement, l'archivage numérique pérenne consiste à conserver le document et l'information qu'il contient :

- Dans son aspect physique comme dans son aspect intellectuel
- Sur le très long terme
- De manière à ce qu'il soit en permanence accessible et compréhensible.

10.4. Le CINES

Le CINES (Centre Informatique National de l'Enseignement Supérieur) est l'opérateur mandaté par le Ministère pour opérer la mission d'archivage pérenne pour l'Enseignement Supérieur et la Recherche.
Il propose VITAM, le Programme interministériel d'archivage numérique.

Selon son institution, sa discipline ou l'entrepôt choisi, il existe déjà des partenariats avec le CINES, proposant un accompagnement pour l'archivage.

Exemple : Huma-Num en SHS

10.5. Préparation de l'archivage

L'archivage pérenne nécessite une préparation en amont et requière de multiples compétences.

Cette vidéo (2 min) vous aidera à mieux comprendre ce qu'est l'archivage pérenne des données et quels sont les acteurs susceptibles d'intervenir à cette étape :

https://www.canal-u.tv/chaines/callisto/les-minutes-doranum/la-minute-archivage-perenne-des-donnees

10.6. Sélection des données à archiver

La valeur des données est à considérer afin de procéder à la sélection des données qu'il sera pertinent d'archiver sur le long terme.

Valeur scientifique des données

- Les données sont-elles uniques, non reproductibles (ou à des coûts trop élevés) ?
- Les données ont-elles une valeur historique, c'est-à-dire représentent-elles un point de repère dans les découvertes scientifiques ?
- Les données comprennent-elles des changements dans les méthodes de traitement, de nouvelles normes ou créent-elles des précédents ?
- Les données appuient-elles les projets en cours ou les tendances scientifiques ?
- Les données sont-elles susceptibles de répondre aux besoins/orientations futurs de la communauté scientifique (potentiel de réutilisation) ?
- Les données sont-elles susceptibles d'être citées ou référencées dans une publication ?
- …

Mesures de contrôle de la qualité des données

- La qualité et la conformité de la collecte des données doivent être contrôlées et documentées. Il peut s'agir des processus comme la calibration, la répétition des échantillons ou des mesures, la capture standardisée des données, la validation de saisie des données, la revue par les pairs …
- La qualité, l'intégrité physique des données (non endommagées, lisibles…).

Considérations politiques / institutionnelles

- Quelle est la politique du financeur, de l'institution ?
- Les données sont-elles conformes à la stratégie de l'institution ?

Considérations juridiques / statutaires

- Y a-t-il une raison légale ou législative pour conserver les données ?
- Existe-t-il une raison évidente pour laquelle les données peuvent être utilisées dans le cadre de litiges, d'enquêtes publiques, d'enquêtes policières ou de tout rapport ou document qui pourrait être contesté en justice ?
- Existe-t-il des obligations financières ou contractuelles qui obligent à conserver les données ?

Considérations financières

Lorsqu'on envisage la préservation des données, le coût de conservation (identifié non seulement comme étant le stockage, mais aussi la gestion, le partage, l'accès, la sauvegarde et la maintenance à long terme des données) doit être mis en balance avec les preuves d'une réutilisation potentielle des données.

Règles de tri et de conservation des archives

Consulter le référentiel de gestion des archives de la recherche, Association des archivistes français, Section Aurore.

NERC Data Value Checklist. *https://www.ukri.org/publications/nerc-data-value-checklist/*
DoRANum. Le Référentiel de gestion des archives de la recherche. 11 septembre 2019. *https://doranum.fr/stockage-archivage/referentiel-de-gestion-des-archives-de-la-recherche_10_13143_pcgd-hy47/*

10.7. Préparation des données à archiver

Voici une check-list pour bien préparer ses données dans une optique d'archivage pérenne :

1) **Sélection des jeux de données**
   Sélectionner les jeux de données (et métadonnées associées) à conserver à long terme, sachant qu'ils peuvent être différents des jeux de données partagés.

2) **Volumétrie**
   Prévoir la volumétrie des données et le budget nécessaire.

3) **Traitement des données**
   Traiter les données si cela est nécessaire.
   Exemple : Données personnelles (nécessitent une anonymisation)

4) **Format de fichiers**
   Vérifier la validité des formats de fichiers de données avec l'outil FACILE mis en place par le CINES.

5) **Logiciels**
   Documenter également les logiciels permettant l'accès aux données.

6) **Métadonnées**
   Compléter et enrichir si besoin les métadonnées.
   Les données doivent posséder une description minimale imposée par le CINES.

Pensez à vous rapprocher des archivistes de votre établissement pour vous aider dans les bonnes pratiques d'archivage pérenne et pour savoir où archiver vos données.

11. Réutilisation et valorisation des données

« Les données de la recherche sont la matière première de la connaissance. Les partager, c'est ouvrir de nouvelles perspectives scientifiques. »
MESR, Ministère de l'Enseignement Supérieur et de la Recherche. Plan national pour la Science Ouverte. 4 juillet 2018. https://www.ouvrirlascience.fr/plan-national-pour-la-science-ouverte/

11.1. Réutilisation et valorisation des données dans le cycle de vie des données

C'est l'ultime étape du cycle de vie des données mais également le point de départ d'un nouveau cycle si celles-ci sont réutilisées pour un nouveau projet de recherche.

Les données déposées doivent trouver leur voie jusqu'aux chercheurs intéressés et réciproquement, les chercheurs doivent trouver parmi les données déposées celles qui sont pertinentes pour leur recherche.

11.2. Réutilisation et citation des données

11.2.1. Du côté du chercheur

Pour que les données de recherche qu'il a produites soient réutilisées dans de bonnes conditions, le chercheur doit adopter plusieurs bonnes pratiques :

- Rendre ses données FAIR
- Documenter ses jeux de données avec un fichier Readme.txt ou Lisez-moi.txt
- Les déposer dans un entrepôt
- Appliquer une licence de diffusion
- Bien renseigner les métadonnées
- Associer le(s) logiciel(s) nécessaire(s) à leur lecture / compréhension
- Appliquer un identifiant pérenne.

11.2.2. Du côté des ré-utilisateurs

Il existe plusieurs manières pour un chercheur de trouver des jeux de données réutilisables.
Parmi celles-ci, une bonne option consiste à consulter les entrepôts dédiés, en particulier ceux du domaine scientifique qui l'intéresse.
Il est possible aussi de faire une recherche depuis les annuaires d'entrepôts (comme re3data, OAD, OpenDOAR, FAIRsharing...).
Des moteurs de recherche comme OpenAIRE Explore et Google Dataset Search sont également des outils très utiles pour la communauté scientifique.
Enfin, une autre manière moins connue de trouver des jeux de données intéressants et bien documentés, c'est de prospecter du côté des **data papers**.

Dans tous les cas, il est à noter que les ré-utilisateurs doivent s'engager à respecter certaines règles :

- Respecter la propriété intellectuelle des auteurs telle que mentionnée dans la licence
- Citer les données si la licence l'exige (il est recommandé de toujours citer ses sources)
- Lier les données aux publications.

11.3. Data papers

11.3.1. Définition de data paper

« À la différence d'un article scientifique classique qui exploite, analyse et interprète les données scientifiques, un article de données (data paper) décrit finement un/des jeu(x) de données de façon à en faciliter la compréhension et l'éventuelle réutilisation. »
MESR, Ministère de l'Enseignement Supérieur et de la Recherche. Deuxième Plan national pour la science ouverte. Juillet 2021. *https://www.ouvrirlascience.fr/deuxieme-plan-national-pour-la-science-ouverte/*

La rédaction et la publication d'un data paper sont une bonne façon pour un chercheur de valoriser ses données de recherche.

Un **data paper** est une publication qui décrit des jeux de données de recherche et les métadonnées associées. C'est un article à part entière, suivant le même processus éditorial que les articles scientifiques classiques :

- Éléments communs aux articles classiques : titre, résumé, mots-clés, révision par les pairs, citation dans des revues académiques ou savantes…
- Éléments spécifiques liés aux données : types de données, formats, acquisition, processus et méthodes de production, métadonnées, réutilisation…

Un data paper peut être publié :

- Soit dans un **data journal** (revue dédiée à ce type de publication),
- Soit dans une **revue classique** qui accepte les data papers.

Les **données** sont **déposées** de préférence **dans un entrepôt de données** et **c'est l'identifiant pérenne** (exemple : le DOI) **qui permet d'établir le lien entre le data paper et les données**.

Point de vigilance : Il vaut mieux **éviter de publier les données sous forme de matériel supplémentaire** (supplementary data), car il y a un risque que les éditeurs prennent la main sur le copyright des données.

Cette vidéo (2 min) présente brièvement comment se déroule la publication d'un data paper :

https://www.canal-u.tv/chaines/callisto/les-minutes-doranum/la-minute-publier-un-data-paper

11.3.2. Avantages d'un data paper

- DOI permettant l'indexation, l'accessibilité et la citation ;
- Promotion, valorisation et visibilité des données ;
- Permet de décrire des types de données hétérogènes, déposées dans différents entrepôts disciplinaires ;
- Facilite la reproductibilité et la réplicabilité sur d'autres terrains d'étude en fournissant les protocoles ;
- Reconnaissance des éditeurs de données (y compris les personnels de laboratoire, des citoyens...) via une publication scientifique avec comité de lecture ;
- Décrit les données sous forme structurée et lisible par un humain, en y ajoutant des aspects techniques facilitant la réutilisation ;
- Possibilité de rajouter des données annexes : analyses statistiques, représentations graphiques... ;
- Lien avec d'autres ressources scientifiques : publications, données d'autres disciplines...

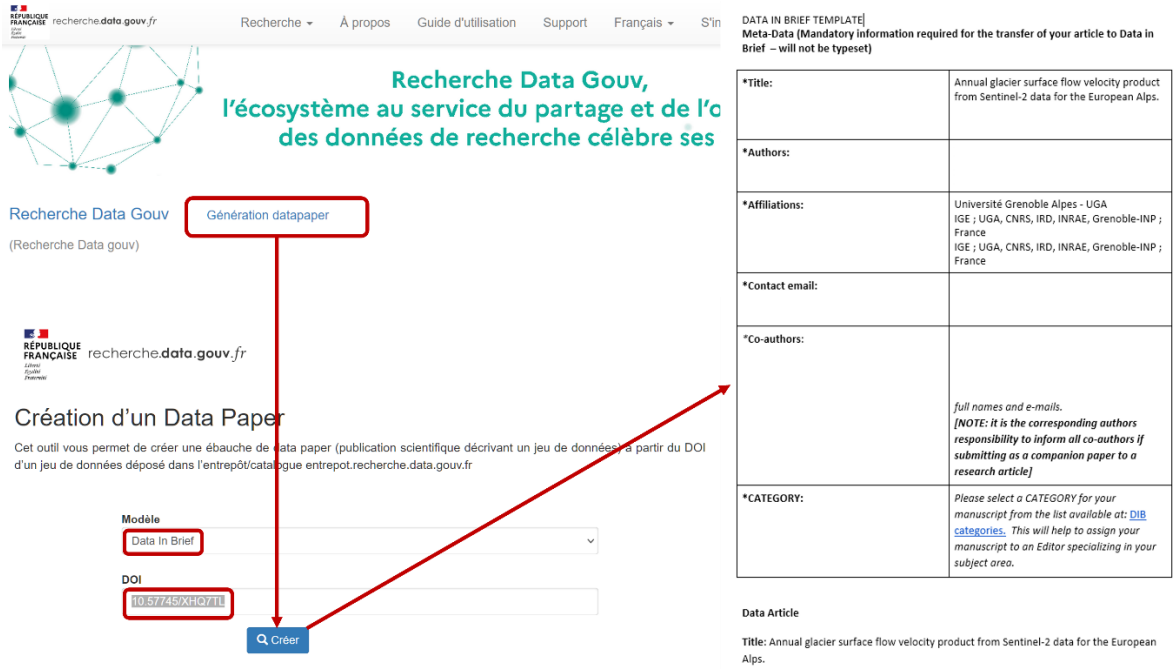
11.3.3. Outils de rédaction d'un data paper

Recherche Data Gouv :

L'outil de génération d'un datapaper est une fonctionnalité spécifique à l'entrepôt Recherche Data Gouv.

Il y a actuellement 2 modèles disponibles :

- Celui de Recherche Data Gouv qui inclut toutes les métadonnées permettant de générer un data paper.
- Celui de Data in Brief qui permet la ventilation des métadonnées suivant un formulaire plus spécifique à cette revue.



https://entrepot.recherche.data.gouv.fr/datapartage-datapapers-web/

Réseau GBIF (Global Biodiversity Information Facility) :

Programme intergouvernemental et infrastructure de données regroupant 64 pays et 42 organisations associées, dont l'objectif est de promouvoir l'accès libre et ouvert aux données sur la biodiversité.

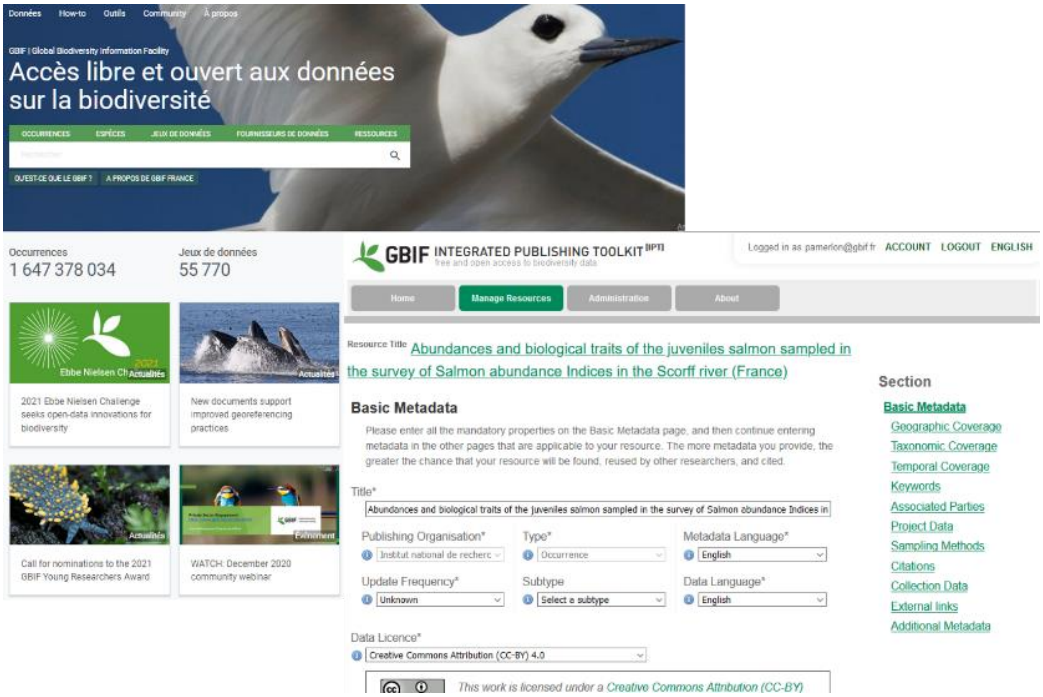
Il propose :

- Un portail d'accès à plus de 2,3 milliards de données primaires sur la biodiversité avec des requêtes interopérables, des cartes interactives et des web services facilitant la réutilisation des données

- L'outil **Integrated Publishing Toolkit** (IPT) : permet de formater les données, facilite le remplissage des métadonnées associées et la production automatisée d'un manuscript de data paper.

La rubrique Data papers propose une liste de revues qui acceptent les data papers.

Archambeau Anne-Sophie. GBIF (Système Mondial d'Information sur la Biodiversité) : bonnes pratiques pour l'accès libre et ouvert aux données de biodiversité. 7 décembre 2022. https://doi.org/10.5281/zenodo.7410132



Pamerlon Sophie. Exemple d'intégration du data paper à un workflow de publication de jeux de données. 5 novembre 2020. https://gt-atelier-donnees.miti.cnrs.fr/download/GBIF_IPT_Sophie_Pamerlon.pdf

Les outils de rédaction d'un data paper sont souvent liés à l'éditeur choisi.

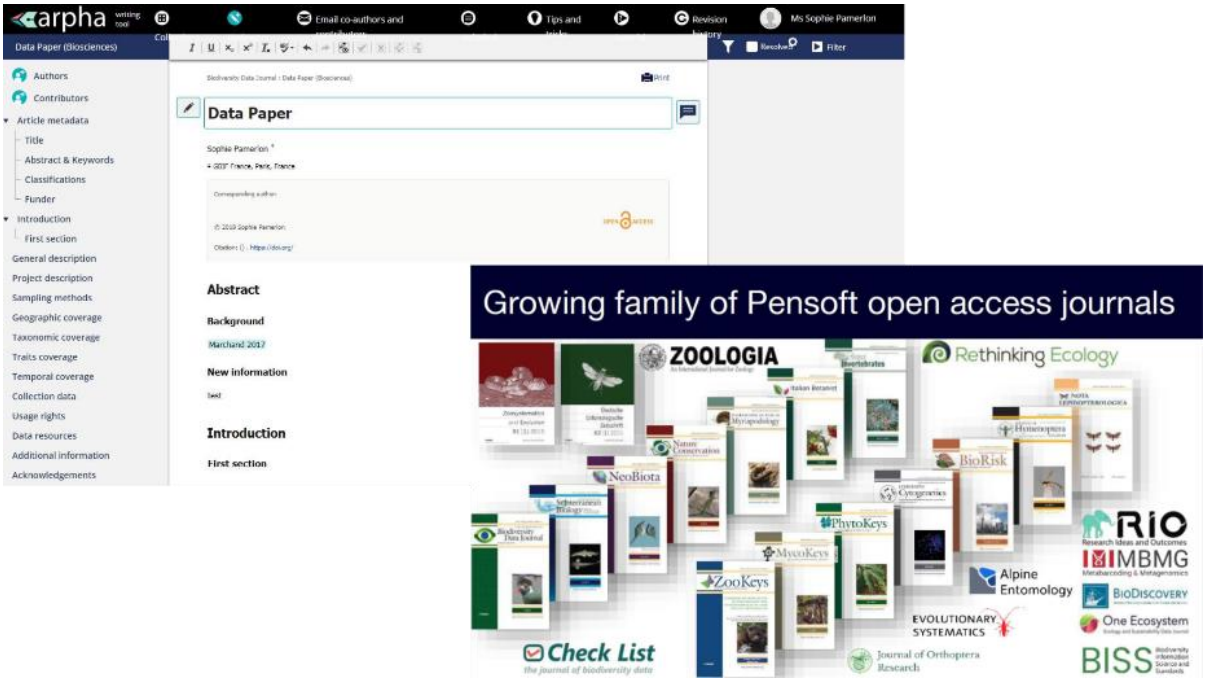
Des modèles sont fournis qui permettent de suivre la structure d'un data paper.

Pensoft (maison d'édition) a mis en place un workflow pour intégrer les données ainsi que les data papers. Pensoft a développé les data papers sur les données de biodiversité en partenariat avec le GBIF.

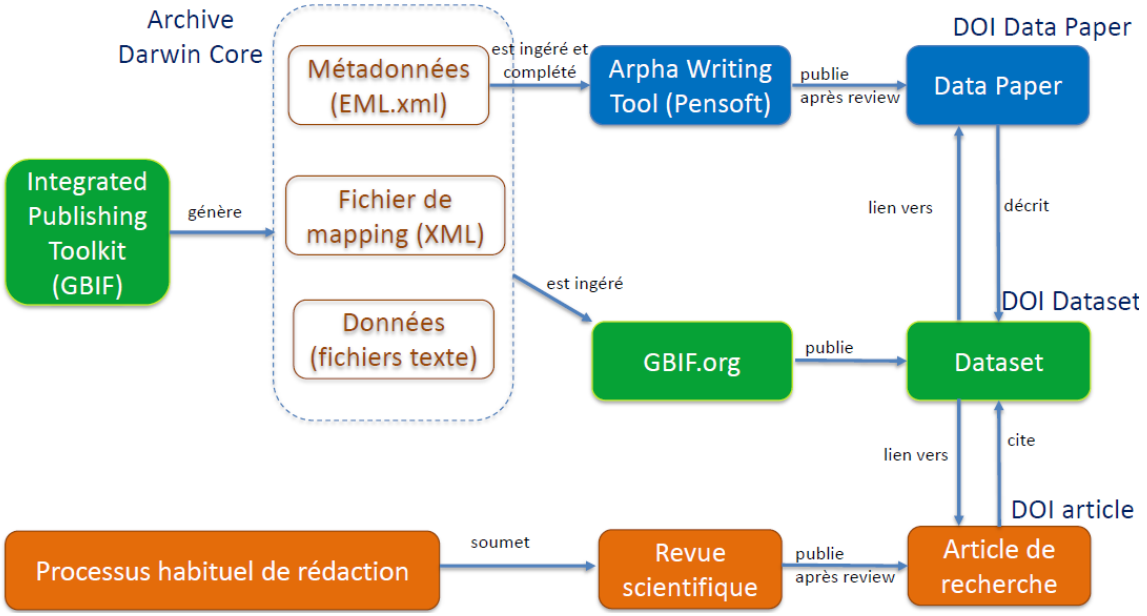
- **Arpha Writing Tool** : outil d'aide à la rédaction de data papers. Il facilite la mise en page, la soumission, le processus de relecture, la publication, la citation, l'hébergement et l'archivage d'articles scientifiques.

  Possibilité de choisir deux voies de publication :

- Document textuel
- Envoi d'un fichier EML.xml des métadonnées à intégrer directement dans l'outil qui va l'interpréter et proposer un brouillon de data paper en fonction de ce qui a été renseigné dans les métadonnées. Les illustrations et graphiques peuvent être rajoutés ensuite.



En résumé

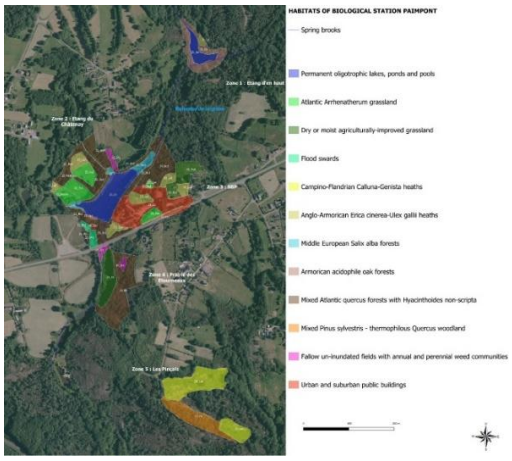


Pamerlon Sophie. Exemple d'intégration du data paper à un workflow de publication de jeux de données. 5 novembre 2020 https://gt-atelier-donnees.miti.cnrs.fr/download/GBIF_IPT_Sophie_Pamerlon.pdf

11.3.4. Exemples de data papers

- **Data paper issu de la science citoyenne utilisant le GBIF**

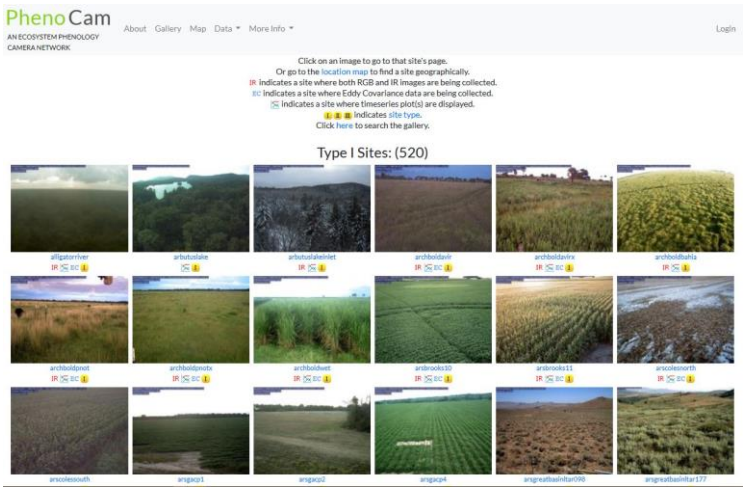
Un inventaire " éclair " de la biodiversité a été réalisé dans un espace donné (Station Biologique de Paimpont), faisant appel à des scientifiques, naturalistes, étudiants et citoyens curieux qui participent aux 3 étapes d'échantillonnage (diversité taxonomique), de tri et identification et de saisie et cartographie des données.
Cet inventaire a permis d'augmenter la part française des données issues de sciences citoyennes dans le GBIF.
43 zones ont été échantillonnées dans 13 habitats sur 17,3 ha. Ces données ont été complétées par les archives de la station sur 60 ans.
Des métadonnées supplémentaires ont été ajoutées pour faciliter la réutilisation : espèces sur liste rouge, espèces natives ou introduites, classification EUNIS des habitats, associations d'espèces...
Des données annexes ont également été décrites dans le data paper : paramètres physiques, chimiques, données sociologiques, cartographie du site d'étude, détails des protocoles, présence ou disparition d'espèces, espèces remarquables, rares, nouvelles, données météo, paramètres du sol, profils hydrologiques, observations du paysage (photos)...
Nicolai Annegret, Guernion Muriel, Supper Régis. BioBlitz 2017 à la Station Biologique de Paimpont - un data paper de science citoyenne. 5 novembre 2020. [*https://gt-atelier-donnees.miti.cnrs.fr/download/BioBlitzSBP_Annegret_Nicolai.pdf*](https://gt-atelier-donnees.miti.cnrs.fr/download/BioBlitzSBP_Annegret_Nicolai.pdf)



Transdisciplinary Bioblitz: Rapid biotic and abiotic inventory allows studying environmental changes over 60 years at the Biological Field Station of Paimpont (Brittany, France) and opens new interdisciplinary research opportunities. [*https://bdj.pensoft.net/article/50451/*](https://bdj.pensoft.net/article/50451/)

- **Data papers pour un même projet sur la phénologie de la végétation en Amérique du Nord**

Deux data papers ont été rédigés sur des données photographiques permettant d'étudier l'évolution de la phénologie de la végétation dans différents écosystèmes à travers l'Amérique du Nord.
Les données sont dérivées d'images numériques automatisées (prises toutes les 30 mn), collectées via le réseau PhenoCam. Les données sont des séries temporelles caractérisant la couleur de la végétation, notamment le degré de verdissement.
L'interface PhenoCam Explorer a été développée afin de faciliter l'exploration et la visualisation des données, à partir desquelles l'utilisateur peut également télécharger des données site par site.
Les images sont également visualisables en temps réel sur la page du projet PhenoCam.



https://phenocam.nau.edu/webcam/gallery/

1er data paper :
Tracking vegetation phenology across diverse North American biomes using PhenoCam imagery
Ce premier data paper paru en 2018 présente la version 1.0 d'une série de jeux de données constituant, tous assemblés, environ 750 ans d'observations (plus de 15 millions d'images produites par environ 400 caméras numériques automatisées).
Les données ont été déposées dans l'entrepôt ORNL DAAC de la NASA.
2ème data paper :
Tracking vegetation phenology across diverse biomes using Version 2.0 of the PhenoCam Dataset

Ce deuxième data paper paru en 2019 présente la version 2.0 d'une série de jeux de données de 1783 sites et 393 caméras numériques situées dans différents écosystèmes sous des climats très variés.
La qualité des jeux de données a été améliorée.
Les données ont été déposées dans l'entrepôt ORNL DAAC de la NASA.

11.4. Data journals

11.4.1. Définition de data journal

« Un data journal est une revue (toujours en libre accès) qui publie des articles de données (data papers). Il fournit habituellement des modèles de description des données et guide les chercheurs sur les lieux de dépôt et sur la façon de décrire et de présenter leurs données. »
L'Hostis Dominique, Hamelin Marjolaine, Lelievre Virginie, Aventurier Pascal. Publier un Data Paper pour valoriser ses données (Cours). Octobre 2016. https://hal.science/hal-02801638/

11.4.2. Exemples de data journals (ou de revues publiant des data papers)

- **Biodiversity Data Journal**

Revue en libre accès dans les domaines de la biodiversité (données taxonomiques, floristiques/faunistiques, morphologiques, génomiques, phylogénétiques, écologiques ou environnementales sur tout taxon de toute époque géologique et de toute partie du monde), sans limite de taille de manuscrit.
Coût de publication (APC) en janvier 2022 : 650 € + 10 € par tranche de 1000 caractères au-delà de 40 000.

- **BioInvasions Records**

Revue internationale en libre accès, évaluée par des pairs, qui présente des recherches de terrain sur les invasions biologiques dans les écosystèmes aquatiques et terrestres du monde entier. Elle se consacre principalement à la publication de travaux de recherche et de data papers sur les données d'espèces non indigènes.
Coût de publication (2023) : 70 €/page, avec un minimum de 840 € pour un article inférieur à 12 pages.

Nature Conservation

Revue en libre accès, évaluée par des pairs, qui traite de tous les aspects de la conservation de la nature. Elle publie des articles dans toutes les disciplines qui s'intéressent à l'écologie fondamentale et appliquée de la conservation et à la conservation de la nature en général, à diverses échelles spatiales, temporelles et évolutives, des populations aux écosystèmes et des micro-organismes et champignons aux plantes et animaux supérieurs.
Coût de publication (février 2022) : 950 € pour 1 à 40 pages. 25 € / page pour chaque page supplémentaire (au-dessus de 40).

One Ecosystem

Revue scientifique innovante, en libre accès, qui offre un forum pour les études dans le domaine de l'écologie et du développement durable. En plus des articles de recherche conventionnels, la revue accueille les contributions documentant l'ensemble du cycle de la recherche, y compris les données, les modèles, les méthodes, les flux de travail, les résultats, les logiciels, les perspectives et les recommandations politiques.
Coût de publication (février 2022) : 550 € + 10 € par tranche de 1000 caractères au-delà de 40 000.

RIO Journal (Research Ideas and Outcomes)

Il a pour but de catalyser le changement dans la communication de la recherche en publiant des idées, des propositions et des résultats de manière exhaustive. Ce faisant, ils espèrent accroître la transparence, la confiance et l'efficacité de l'ensemble de l'écosystème de la recherche.
Coût de publication : 299 € + 10 € par tranche de 1000 caractères supplémentaires au-delà de 20 000 (+ frais de conversion de 30%).

11.4.3. Ressources complémentaires

Site CoopIST du CIRAD : Rédiger et publier un data paper

Site GBIF : Publier un data paper

11.5. Exposition et visualisation des données

En complément du dépôt dans un entrepôt, et peut-être de la publication d'un data paper, l'exposition des données est une autre bonne façon de les valoriser.

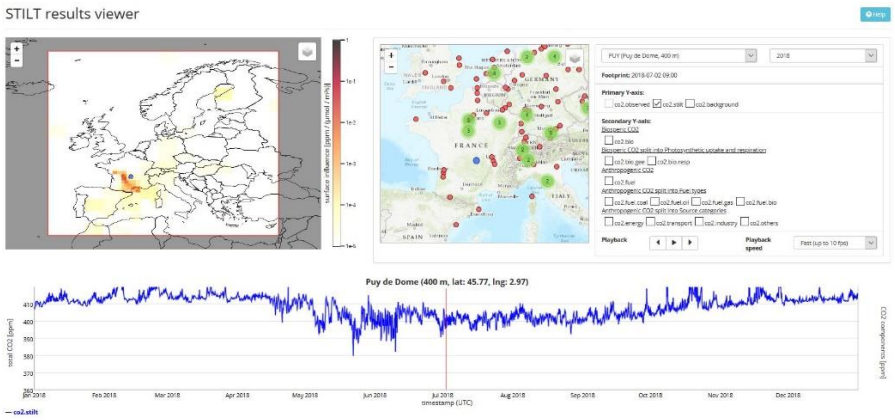
En effet, il peut s'avérer utile, surtout dans le cas de données nombreuses et complexes, d'exposer ses données sous forme visuelle (cartographies, graphiques, etc.) via une plateforme.

11.5.1. Exemple 1

Les données disponibles sur le portail ICOS Carbon Portal proviennent de séries chronologiques de valeurs sur des centaines de paramètres.

On peut voir par exemple à l'aide d'outils de visualisation l'évolution des concentrations de CO2 sur une année, couplée à l'origine de la masse d'air. Ce serait très difficile à appréhender sans passer par la visualisation de données.

Il faut un login – mot de passe pour utiliser ce service.



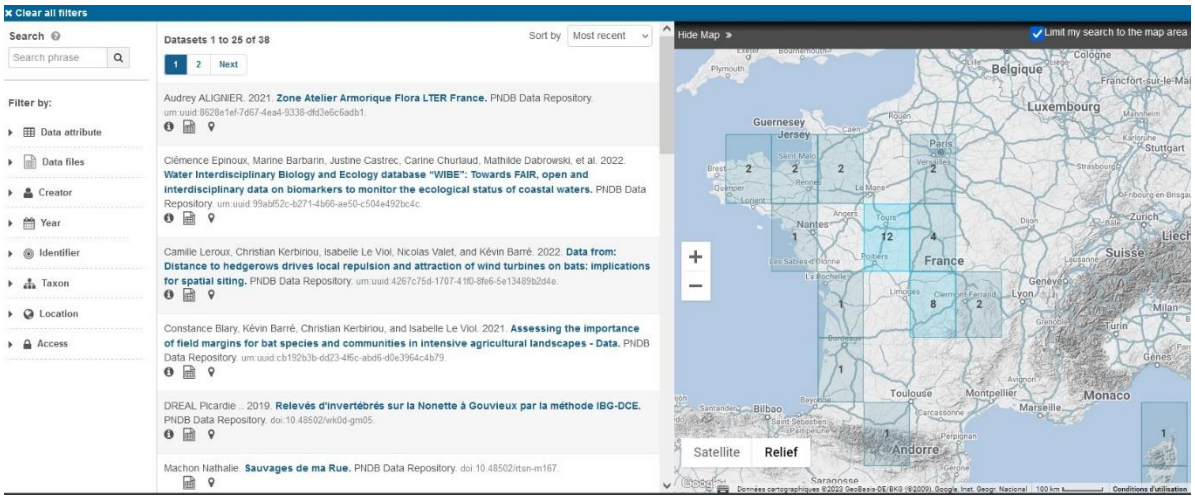
https://stilt.icos-cp.eu/viewer/

11.5.2. Exemple 2

L'infrastructure Pôle National de Données de Biodiversité (PNDB), qui s'inscrit dans une approche FAIR, est au service des scientifiques produisant, gérant et analysant des données sur la biodiversité.

Le portail PNDB propose un accès aux jeux de données et aux métadonnées associées, ainsi qu'à des services associés et à des produits dérivés des analyses.

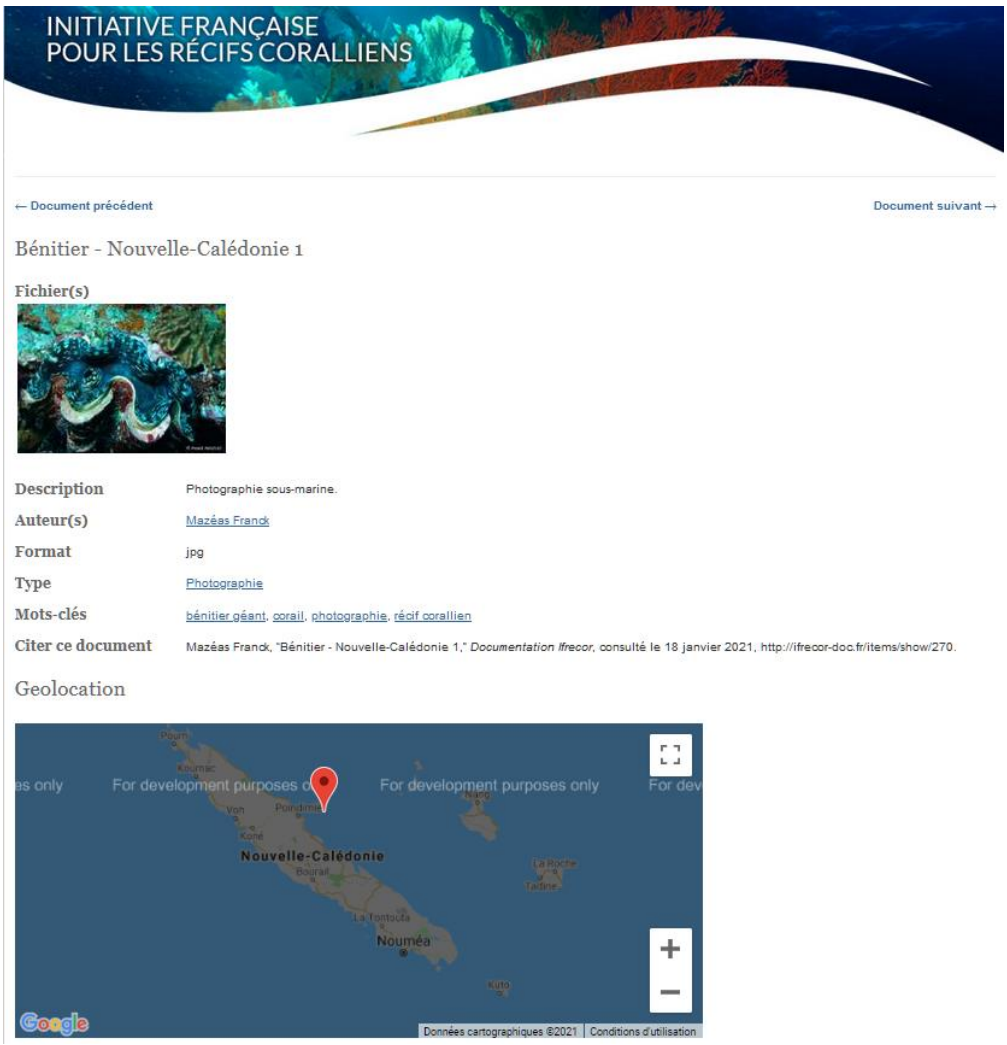
Pour que les données soient présentées dans le portail PNDB, il faut utiliser le standard de métadonnées international EML (Ecological Metadata Language) et associer les données et métadonnées au sein du package de données. Chaque jeu de données est géolocalisé. La connexion au portail se fait via un compte ORCID.



https://data.pndb.fr/data

11.5.3. Exemple 3

Le site IFRECOR Documentation, réalisé avec le logiciel Omeka, valorise la documentation produite ou financée par l'IFRECOR (Initiative Française pour les REcifs CORaliens), ainsi que celle de ses membres partenaires. Il présente également des documents sur les récifs coralliens produits par d'autres institutions que l'IFRECOR et ses partenaires. Cette documentation aussi diverse que variée concerne les principales actions nationales et locales de l'initiative. Les photographies ont la particularité d'être géolocalisées, vous pouvez accéder à ces dernières directement depuis une carte située à l'onglet " Géolocaliser ".



http://ifrecor-doc.fr/items/show/270

11.5.4. Outil d'exposition et de visualisation des données

Omeka est un outil qui s'inscrit pleinement dans la science ouverte et qui permet la création de bases de données au plus près des principes FAIR.
À partir de données de recherche brutes, l'outil permet de créer des bases de données éditorialisées, autrement dit structurées, accessibles, et visibles sur le web.
L'outil offre une grande modularité des fonctionnalités grâce à de nombreux plugins, et traite les divers objets multimédia (textes, images, sons, vidéos).
L'outil offre plusieurs avantages techniques :

- L'interface est simple et intuitive ;
- Les métadonnées sont moissonnables, permettant notamment le référencement dans d'autres bases ;
- Une base Omeka peut être connectée à d'autres services grâce à une API REST.

VALIDATION

12. Testez vos connaissances (7 exercices)

1/7. Dans le cas d'un projet financé par l'ANR (Agence nationale de la recherche)...

Consigne : cochez la bonne réponse

- ☐ la rédaction d'un DMP est obligatoire
- ☐ la rédaction d'un DMP est fortement recommandée

Solution :

La rédaction d'un DMP est obligatoire. En effet, l'ANR a rendu la rédaction d'un DMP obligatoire pour tous les projets qu'elle finance depuis 2019.

2/7. À quoi correspondent les licences suivantes ?

Consigne : glissez et accolez chaque élément comportant le nom d'une licence (à gauche) avec sa définition (à droite).

Licence ouverte (Etalab)	Licence de logiciel libre
CC-BY-SA	Licence libre pour les bases de données
CC-By-NC	Licence ouverte française pour les données publiques
CeCILL-B	Attribution - Partage dans les mêmes conditions
ODBL	Attribution - Pas d'utilisation commerciale

Solution :

Licence ouverte (Etalab)	Licence ouverte française pour les données publiques
CC-BY-SA	Attribution - Partage dans les mêmes conditions
CC-By-NC	Attribution - Pas d'utilisation commerciale
CeCILL-B	Licence de logiciel libre
ODbL	Licence libre pour les bases de données

3/7. Bonnes pratiques de gestion et partage des données

Consigne : testez votre connaissance des bonnes pratiques de gestion et partage des données en répondant aux questions de Nora. Cliquez pour cela sur la proposition qui vous semble correcte.

Cas n°1. Comment partager ses données ?

Nora est une jeune chercheuse. Vous devez l'aider à adopter de bonnes pratiques pour bien partager ses données de recherche.

- **Nora : Où dois-je déposer mes données pour bien les partager ?**

1) Il faut déposer ses données dans un entrepôt de données.
2) Il faut déposer ses données dans une archive pérenne.

Solution :

Si vous avez choisi la proposition 1) Oui ! C'est bien ça, déposer ses données dans un entrepôt permet de les partager efficacement.

Si vous avez choisi la proposition 2) Non, l'archivage pérenne a pour objectif premier de conserver les données sur le long terme et d'en préserver l'accès et l'intelligibilité. Pour partager ses données, il faut déposer dans un entrepôt.

- **Nora : Et quel entrepôt me conseillez-vous de choisir ?**

1) Il est préférable de choisir un entrepôt disciplinaire.
2) Il est préférable de choisir un entrepôt généraliste.

Solution :

Si vous avez choisi la proposition 1) Oui, il est effectivement conseillé de choisir un entrepôt disciplinaire et s'il n'en existe pas dans le domaine souhaité, on peut opter pour un entrepôt généraliste.

Si vous avez choisi la proposition 2) L'entrepôt généraliste est une bonne option uniquement quand il n'existe pas d'entrepôt disciplinaire.

Cas n°2. Nora a trouvé un entrepôt dans sa discipline. Comment doit-elle préparer ses données en vue du partage ?

- **Nora : Dois-je déposer toutes mes données dans l'entrepôt ?**

1) Oui, tout à fait, toutes les données de recherche doivent être déposées dans l'entrepôt !
2) Non, il faut faire un tri parmi ses données.

Solution :

Si vous avez choisi la proposition 1) Non. Un choix doit être opéré pour déterminer quels jeux de données il est pertinent de déposer dans l'entrepôt.

Si vous avez choisi la proposition 2) En effet, toutes les données produites au cours d'un projet n'ont peut-être pas vocation à être déposées dans un entrepôt pour être partagées. Il faut opérer le plus souvent une sélection.

- **Nora : Mes données sont en format .xls. Cela convient-il ?**

1) Oui, c'est très bien. Il suffit d'indiquer que les données sont en .xls dans les métadonnées.
2) Non, il faut convertir les données dans un format plus adapté.

Solution :

Si vous avez choisi la proposition 1) C'est bien de renseigner le format dans les métadonnées mais ce n'est pas suffisant. Pour que les données soient FAIR, il faut les migrer dans un format ouvert et non propriétaire. En .csv ici.

Si vous avez choisi la proposition 2) Oui, tout à fait ! Pour que les données soient FAIR, il faut les migrer dans un format ouvert et non propriétaire, c'est-à-dire en .csv dans notre cas.

- **Nora : Comment dois-je faire pour les métadonnées ?**

1) Il suffit de renseigner les 15 éléments Dublin Core.
2) Le mieux est de choisir un standard dans sa discipline !
3) L'idéal est de créer son propre schéma de métadonnées !

Solution :

Si vous avez choisi la proposition 1) Les 15 éléments du Dublin Core (standard généraliste) sont le minimum requis. Il est préférable de choisir un standard dans sa discipline afin d'avoir des métadonnées les plus riches possibles.

Si vous avez choisi la proposition 2) Bravo ! L'idéal est bien de choisir un standard dans sa discipline ou qui correspond à vos besoins et de bien le renseigner. Plus les métadonnées sont riches, mieux c'est.

Si vous avez choisi la proposition 3) Ce n'est pas tout à fait exact. Il est conseillé de créer votre propre schéma de métadonnées uniquement si aucun standard de métadonnées ne convient à vos besoins.

Nora : Je connais désormais les bonnes pratiques pour déposer mes données dans un entrepôt et bien les partager !

4/7. En règle générale, l'attribution de la propriété intellectuelle des données revient...

Consigne : cochez la bonne réponse

☐ au producteur de ces données

☑ à l'établissement de tutelle des producteurs de données

Solution :

Il ne s'agit pas du même droit que pour les publications.

Les données relèvent d'un **régime lié au droit des bases de données**. Dans ce cas, le droit de propriété appartient légalement au « producteur » de la base de données, compris au sens de la personne qui réalise l'investissement financier et matériel nécessaire à la constitution de la base. Il s'agira donc en général de **l'établissement de tutelle des chercheurs** qui **sera considéré comme le titulaire effectif du droit de propriété**.

Mais si ce droit existe formellement, il ne peut plus être opposé aux droits des ré-utilisateurs des données (principe d'ouverture des données). En effet, la **loi pour une République numérique** a explicitement « neutralisé » le droit des bases de données des administrations pour faire primer le **principe de libre réutilisation**. Il en résulte que les données produites par les chercheurs sont bien comprises dans le principe d'ouverture par défaut.

5/7. Que signifie l'acronyme FAIR ?

Consigne : écrivez votre réponse dans la case prévue à cet effet

F pour …

A pour…

I pour…

R pour …

Solution :

F pour Facile à trouver (FR) ou findable (EN)

A pour Accessible (FR) ou accessible (EN)

I pour Interopérable (FR) ou Interoperable (EN)

R pour Réutilisable (FR) ou reusable (EN)

6/7. S'agit-il d'un identifiant pérenne ou d'un standard de métadonnées ?

Consigne : glissez et déposez chaque carte sur une des deux zones grises. Dans ces tableaux, les réponses sont dans le désordre.

SWHID ISO 19115 ORCID	Identifiant pérenne
EML DOI Dublin Core	Standard de métadonnées

Solution :

Identifiant pérenne SWHID ORCID DOI
Standard de métadonnées ISO 19115 EML Dublin Core

7/7. À quoi servent ces outils ?

Consigne : glissez et accolez chaque élément comportant le nom d'un outil (à gauche) avec sa définition (à droite).

License Selector	Vérifier la validité des formats de fichiers de données pour l'archivage pérenne
re3data	Recense et décrit les services français dédiés aux données scientifiques
Omeka	Annuaire d'entrepôts
DMP OPIDoR	Choisir une licence
Cat OPIDoR	Exposer et visualiser les données
FACILE	Formater les données, remplir les métadonnées associées et produire un data paper
Integrated Publishing Toolkit	Outil de rédaction des plans de gestion de données

Solution :

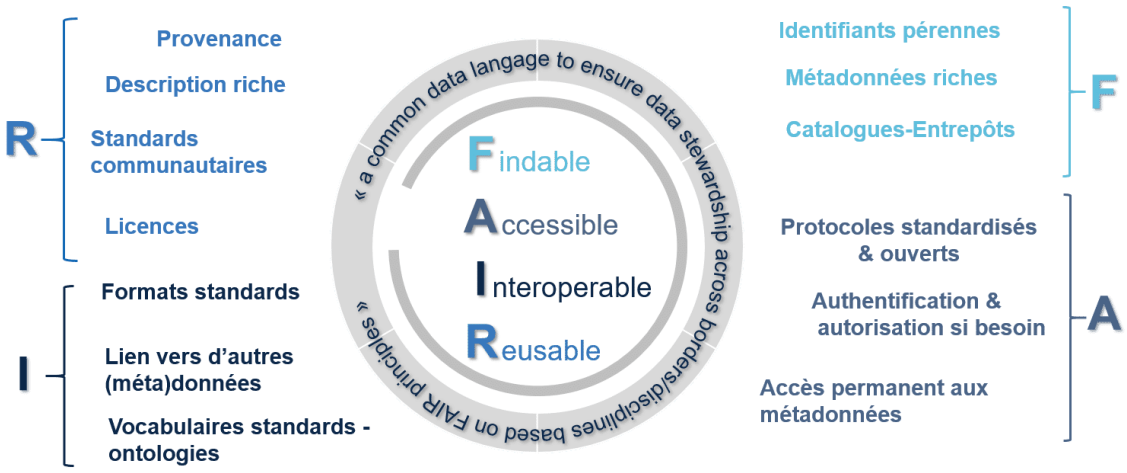
License Selector	Choisir une licence
re3data	Annuaire d'entrepôts
Omeka	Exposer et visualiser les données
DMP OPIDoR	Outil de rédaction des plans de gestion de données
Cat OPIDoR	Recense et décrit les services français dédiés aux données scientifiques
FACILE	Vérifier la validité des formats de fichiers de données pour l'archivage pérenne
Integrated Publishing Toolkit	Formater les données, remplir les métadonnées associées et produire un data paper

POUR TERMINER

13. À retenir

Voici deux schémas récapitulatifs pour résumer l'essentiel :

13.1. Les principes FAIR



Implementation Roadmap for the European Science Cloud (Staff Working Document SWD(2018) 83), 14 March 2018

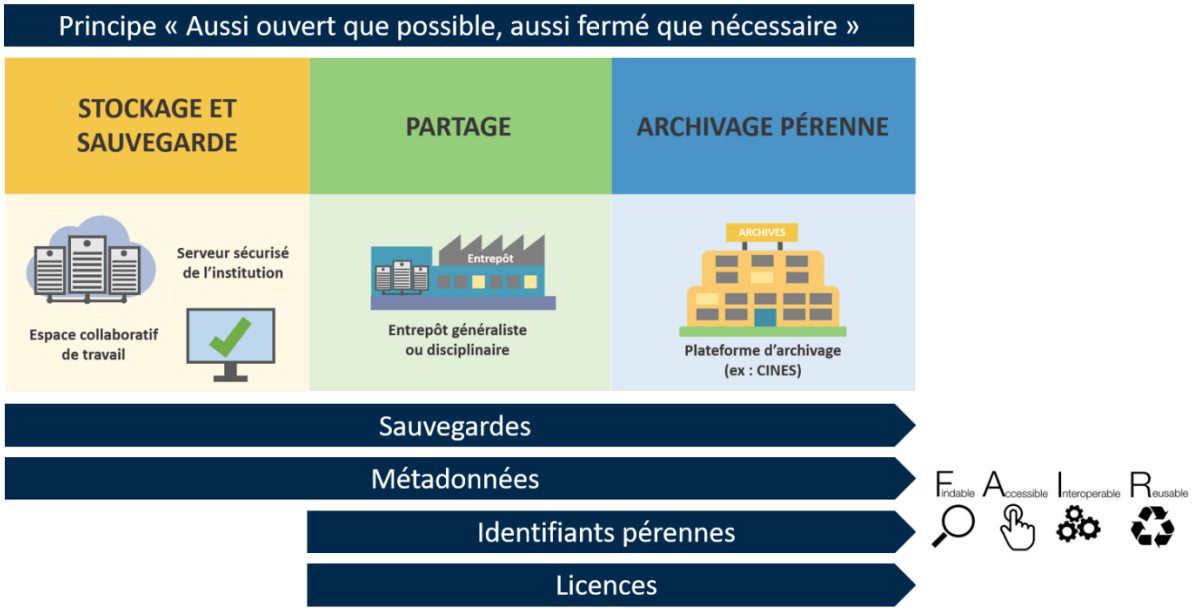
Traduction INRAE : https://science-ouverte.inrae.fr/les-donnees-et-le-numerique-scientifiques/produire-des-donnees-fair

4 grands principes sont à respecter pour garantir une utilisation optimale des données de recherche et des métadonnées associées, à la fois par les hommes et par les machines :

- F (Findable) = Facile à trouver
- A (Accessible) = Accessible
- I (Interoperable) = Interopérable
- R (Reusable) = Réutilisable

Ces principes, admis par les différentes communautés scientifiques au niveau international ainsi que par les financeurs, sont applicables tout au long du cycle de vie des données.
Pour rendre les données de la recherche FAIR, un certain nombre de bonnes pratiques sont à mettre en place.

13.2. Bonnes pratiques pendant et après le projet de recherche



Logo FAIR : SangyaPundir

Durant le projet de recherche, il faut veiller au **stockage** et à la **sauvegarde** des données dans un espace collaboratif sécurisé, accessible aux partenaires du projet.

Le **partage** consiste au dépôt des données dans un entrepôt dédié, de préférence disciplinaire. La sélection des données se fait suivant le principe "aussi ouvert que possible, aussi fermé que nécessaire" avec un objectif de partage à court et moyen terme.
Pour un partage efficace et optimal, les données doivent répondre aux principes FAIR, être accompagnées de métadonnées, d'un identifiant pérenne unique ainsi que d'une licence.

L'**archivage pérenne** ne concerne que les données présentant un intérêt majeur et nécessitant la garantie d'une conservation à long terme, soit plus de 30 ans.

Tout au long du cycle de vie des données, il est important de bien renseigner les métadonnées, l'idéal étant de procéder au fur et à mesure de l'avancée du projet.

14. Webographie

- *ALLEA, All European Academies. The European Code of Conduct for Research Integrity. Revised Edition 2023. Berlin (DE). Juin 2023. https://allea.org/wp-content/uploads/2023/06/European-Code-of-Conduct-Revised-Edition-2023.pdf*
- *ANR, Agence nationale de la recherche. Science ouverte : point d'étape sur la politique commune du réseau des agences de financement françaises. 11 mars 2022. https://anr.fr/fr/actualites-de-lanr/details/news/science-ouverte-point-detape-sur-la-politique-commune-du-reseau-des-agences-de-financement-franca/*
- *Archambeau Anne-Sophie. GBIF (Système Mondial d'Information sur la Biodiversité) : bonnes pratiques pour l'accès libre et ouvert aux données de biodiversité. 7 décembre 2022. https://doi.org/10.5281/zenodo.7410132*
- *Arnould Pierre-Yves, Jacquemot-Perbal Marie-Christine. Guide de bonnes pratiques : Gestion et valorisation des données de recherche. 1er février 2016. https://hal.science/hal-01275841/*
- *Becard Nicolas, Castets-Renard Céline, Chassang Gauthier, Dantant Martin, Freyt-Caffin Laurence, Gandon Nathalie, Martin Caroline, Martelletti Andrea, Mendoza-Caminade Alexandra, Morcrette Nathalie, Neirac Claire. Ouverture des données de la recherche. Guide d'analyse du cadre juridique en France. Décembre 2017. https://hal.science/hal-02791224*
- *CIRAD. CoopIST, Coopérer en Information Scientifique et Technique. Gérer les données de la recherche. https://coop-ist.cirad.fr/gerer-des-donnees*
- *CIRAD, INRAE. Avis 8 sur les enjeux éthiques et déontologiques du partage et de la gestion des données issues de la recherche. Février 2016. https://hal.inrae.fr/hal-02796585*
- *CNRS, Centre national de la recherche scientifique. Responsabilité de recherche. https://www.cnrs.fr/fr/le-cnrs/responsabilites/responsabilite-de-recherche*
- *Cocaud Sylvie, L'Hostis Dominique. Pourquoi et comment rédiger un plan de gestion de données ? 5 avril 2019. https://hal.science/hal-02791507*
- *COMETS, Comité d'éthique du CNRS. Charte nationale de déontologie des métiers de la recherche. 26 janvier 2015. https://comite-ethique.cnrs.fr/charte/*
- *COMETS, Comité d'éthique du CNRS. Guide pratique. Pratiquer une recherche intègre et responsable. Mars 2017. https://comite-ethique.cnrs.fr/guide-pratique/*
- *Couperin. Groupe de travail science ouverte. Groupe données. https://gtso.couperin.org/groupe-donnees/*

- *Couperin. Groupe de travail science ouverte. SOS-PGD. https://gtso.couperin.org/gtdonnees/sos-pgd/*
- *Deboin Marie-Claude. Découvrir de nouveaux métiers liés aux données de la recherche. CIRAD. 5 p. 5 octobre 2018. https://doi.org/10.18167/coopist/0061*
- *Delplanque Catherine, Lamrini Nawale, Leclère Fabrice, Maurel Lionel, et al. Guide : Règlement Général pour la Protection des Données. 2019. https://www.u-plum.fr/guide-reglement-general-pour-la-protection-des-donnees/*
- *École thématique DATA-SDUE. Guide de Survie dans la jungle des données en Sciences de l'Univers et de l'Environnement (SDUE) : Comment gérer les données pour les valoriser ? 10 au 14 octobre 2022. https://data-sdue.sciencesconf.org/*
- *École thématique E-Envir Strasbourg. Les données ouvertes en sciences environnementales : exploration, cas d'études et applications. Du 2 au 5 novembre 2021. https://e-envir-21.sciencesconf.org/*
- *European Commission. Ethics and data protection. 14 novembre 2018. https://cache.media.education.gouv.fr/file/2018/54/9/h2020_hi_ethics-data-protection_en_1046549.pdf*
- *Gaillard Rémi. De l'open data à l'open research data : quelle(s) politique(s) pour les données de la recherche. Janvier 2014. http://www.enssib.fr/bibliotheque-numerique/documents/64131-de-l-open-data-a-l-open-research-data-quelles-politiques-pour-les-donnees-de-recherche.pdf*
- *Ginouvès Véronique, Gras Isabelle, et al. La diffusion numérique des données en SHS – Guide des bonnes pratiques éthiques et juridiques. Octobre 2018. https://hal-amu.archives-ouvertes.fr/page/guide-de-bonnes-pratiques*
- *Hadrossek Christine, Janik Joanna, Libes Maurice, Louvet Violaine, Quidoz Marie-Claude, Rivet Alain, Romier Geneviève. Atelier Données. Guide de bonnes pratiques sur la gestion des données de la recherche. Version 2.0. 23 août 2023. https://mi-gt-donnees.pages.math.unistra.fr/guide/*
- *Inist-CNRS. Cat OPIDoR, wiki des services dédiés aux données de la recherche. https://cat.opidor.fr/*
- *Inist-CNRS. DMP OPIDoR. https://dmp.opidor.fr/*
- *Inist-CNRS, GIS Urfist. DoRANum (Données de la Recherche : Apprentissage Numérique). https://doranum.fr/*
- *INRAE, Institut national de recherche pour l'agriculture, l'alimentation et l'environnement. Le numérique pour la science et les données scientifiques.*

https://science-ouverte.inrae.fr/le-numerique-pour-la-science-et-les-donnees-scientifiques

- *INRAE, Institut national de recherche pour l'agriculture, l'alimentation et l'environnement. Les aspects éthiques et juridiques, et la propriété intellectuelle.* https://science-ouverte.inrae.fr/les-donnees-et-le-numerique-scientifiques/partager-publier-des-donnees-et-des-codes
- *L'Hostis Dominique, Hamelin Marjolaine, Lelievre Virginie, Aventurier Pascal. Publier un Data Paper pour valoriser ses données (Cours). Octobre 2016.* https://hal.science/hal-02801638/
- *Maurel Lionel. La réutilisation des données de la recherche après la loi pour une République numérique. La diffusion numérique des données en SHS - Guide de bonnes pratiques éthiques et juridiques. 2018.* https://hal.science/hal-01908766
- *MESR, Ministère de l'Enseignement Supérieur et de la Recherche. Plan national pour la Science Ouverte. 4 juillet 2018.* https://www.ouvrirlascience.fr/plan-national-pour-la-science-ouverte/
- *MESR, Ministère de l'Enseignement supérieur et de la Recherche. Deuxième Plan national pour la science ouverte. Juillet 2021.* https://www.ouvrirlascience.fr/deuxieme-plan-national-pour-la-science-ouverte/
- *MESR, Ministère de l'Enseignement supérieur et de la Recherche. Recherche Data Gouv.* https://recherche.data.gouv.fr/
- *Mission pour les Initiatives Transverses et Interdisciplinaires du CNRS. Atelier "Données".* https://mi-gt-donnees.pages.math.unistra.fr/site/
- *Nicolai Annegret, Guernion Muriel, Supper Régis. BioBlitz 2017 à la Station Biologique de Paimpont - un data paper de science citoyenne. 5 novembre 2020.* https://gt-atelier-donnees.miti.cnrs.fr/download/BioBlitzSBP_Annegret_Nicolai.pdf
- *Ofis, Office français de l'intégrité scientifique.* https://www.ofis-france.fr/
- *OCDE, Organisation de Coopération et de Développement Économiques. Principes et lignes directrices de l'OCDE pour l'accès aux données de la recherche financée sur fonds publics. 2007.* https://doi.org/10.1787/9789264034020-en-fr
- *Pamerlon Sophie. Exemple d'intégration du data paper à un workflow de publication de jeux de données. 5 novembre 2020.* https://gt-atelier-donnees.miti.cnrs.fr/download/GBIF_IPT_Sophie_Pamerlon.pdf
- *Quidoz Marie-Claude. Atelier « Carnets de terrain électroniques ». Sécuriser les données produites par les carnets de terrain électroniques. 28-29 mars 2018.* https://oreme.org/app/uploads/Quidoz_Atelier2018.pdf

- *Science Europe. Guide pratique pour une harmonisation internationale de la gestion des données de recherche. Juillet 2019. https://www.ouvrirlascience.fr/science-europe-guide-pratique-pour-une-harmonisation-internationale-de-la-gestion-des-donnees-de-recherche/*
- *Stérin Anne-Laure. Diffuser des données de la recherche dans le respect du droit et de l'éthique – Comment faire lorsqu'on n'est pas juriste ? La diffusion numérique des données en SHS - Guide de bonnes pratiques éthiques et juridiques. Octobre 2018. https://hal.science/hal-02050510*
- *Vachez Dominique. Étude comparative de thésaurus en Sciences de l'Environnement - Bonnes pratiques de conception et FAIRisation de thésaurus. Juin 2021. https://hal.science/hal-03264803/*