



Production et sauvegarde des données de la science

Alexandra Niedzwiedz

► To cite this version:

Alexandra Niedzwiedz. Production et sauvegarde des données de la science : Atelier des doctorants du BETA. Doctorat. Saint-dié-des-Vosges, France. 2023, pp.34. hal-04342272

HAL Id: hal-04342272

<https://hal.science/hal-04342272>

Submitted on 13 Dec 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Open licence - etalab

Production et sauvegarde des données de la science

Alexandra NIEDZWIEDZ

BETA/OLEF 

Observatoire pour L'Economie de la Forêt



Etalab 2.0



BETA
Bureau
d'économie
théorique
et appliquée

12 décembre 2023
Atelier doctorants 2^{ème} année



STRUCTURE D'APPUI
À LA RECHERCHE LUE



Petit QUIZ avant de commencer !

- Vous ❤️ faire avancer la recherche ? ✓
- Vous ❤️ disposer de données utilisables ? ✓
- Vous ❤️ partager vos idées/résultats ? ✓
- Vous ❤️ travailler en équipe ? ✓



Autant de bonnes raisons pour **bien gérer**
les données que vous produisez !



Objectif : Bien gérer ses données

On ne vous demande pas la Lune !

Pour y arriver, suivre le plan de vol :

LA THÉORIE

- ▶ La gestion de données de la recherche : concepts
- ▶ Les enjeux

EN PRATIQUE

- ▶ Les outils à utiliser
- ▶ Les ressources pour vous accompagner

Un peu de théorie...

La gestion de données de la recherche : Définitions et principes de base



- **Donnée** = enregistrement de tout type

Données d'observation, expérimentales, computationnelles ou de simulation, compilées ou dérivées

Formats divers : image, son, texte, jeu de données (données encodées dans une structure définie), modèle, programme, logiciel, etc.

- **Gestion des données** = partie intégrante d'un projet de recherche

Processus complexe comprenant différentes étapes (création, organisation, documentation, partage, conservation) tout au long du projet.

Assure la qualité des données, nécessaire dans le contexte d'Open Science / Open Data.

- **Principes et outils liés :**

Cycle de vie des données ; Principes FAIR ; Données à Caractère Personnel ; Science Ouverte

Plan de Gestion de Données (PGD) = Data Management Plan (DMP) ; Licences

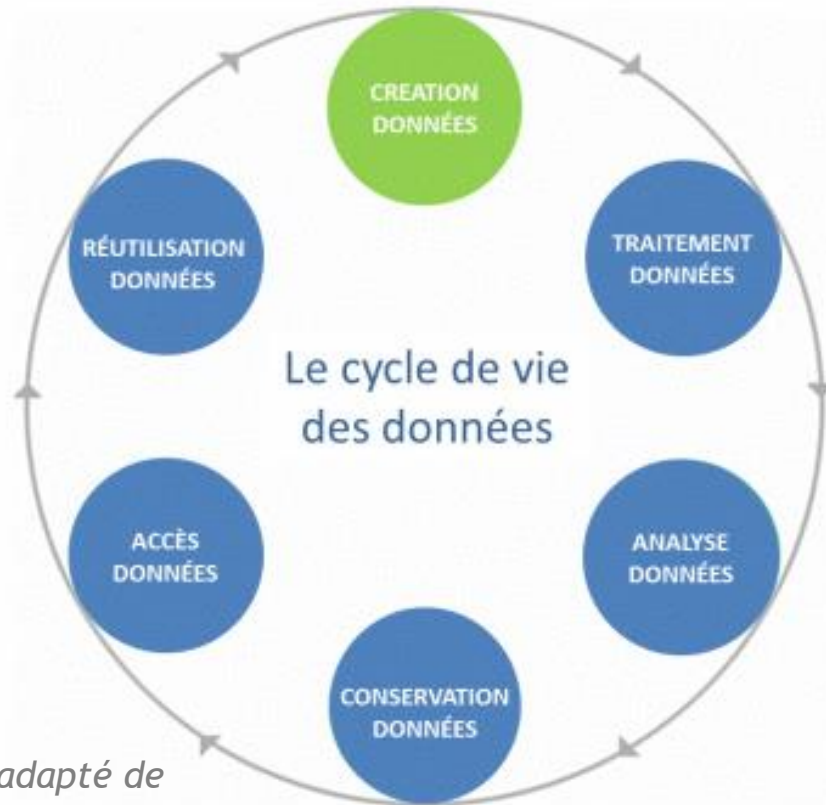
Les enjeux de la gestion des données de la recherche



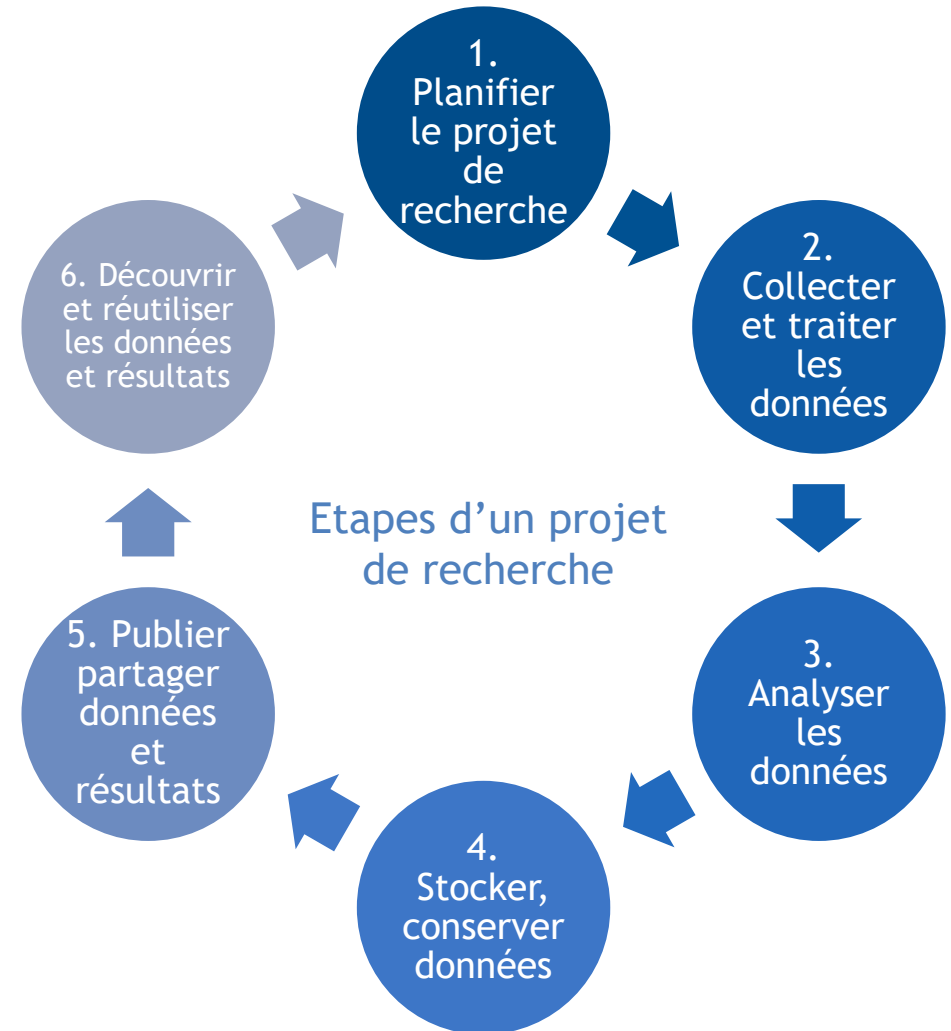
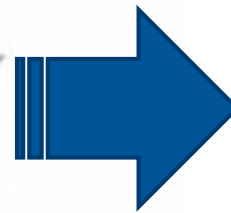
- ▶ Accroître l'**efficacité** de la recherche à court et long terme : accès et analyse facilités ; formatage et documentation garantissant la valeur des données
- ▶ Assurer la continuité de la recherche : **réutilisation** des données
- ▶ Assurer l'**intégrité** des données et la validation des résultats, limite la fraude : suivi du processus de production de la donnée, versionnage
- ▶ Renforcer la **sécurité** des données : organisation et stockage adapté réduisant le risque de perte d'informations
- ▶ Faciliter la **publication** de résultats et réduction du risque de plagiat : documentation utile pour les articles scientifiques et **data papers** ; métadonnées précisant auteurs
- ▶ Répondre aux exigences des financeurs : vers davantage d'**ouverture** des données (**transparence**, mise à disposition)
- ▶ Attester de la **responsabilité** des chercheurs : preuve d'une utilisation responsable du financement public de la recherche

Le cycle de vie des données

- ▶ « Processus d'utilisation des données de leur création à la publication et à leur réutilisation ultérieure. »
- ▶ Modèle pour gérer les données de la recherche.



Inist (adapté de
UK Data Archive)

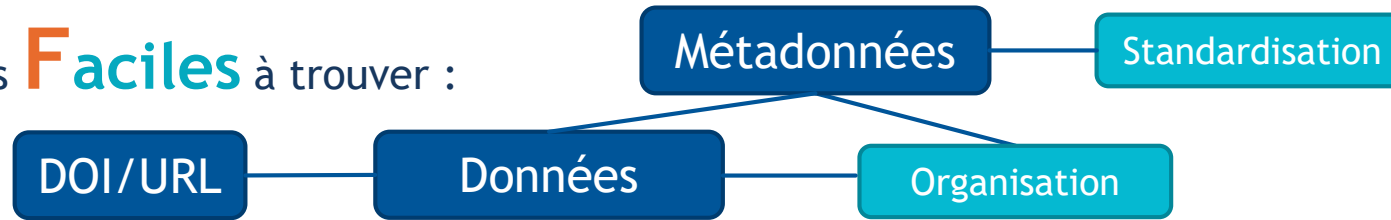


Gestion durable des données : principes FAIR

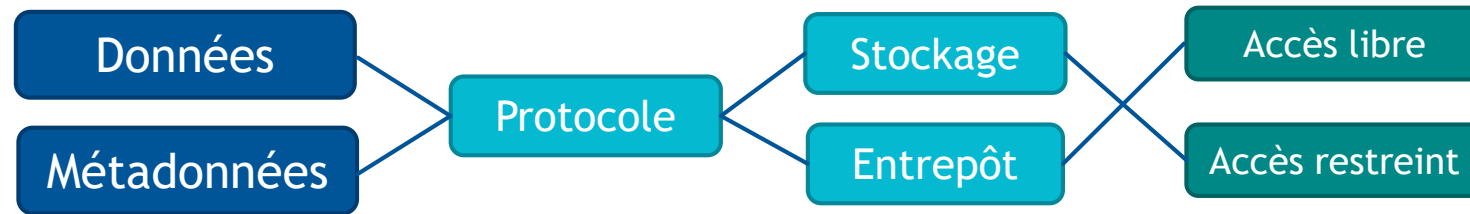


Lien interactif sur principes FAIR

- Données **F**aciles à trouver :



- Données **A**ccessibles : stockage durable et conditions d'accès



- Données **I**nteropérables : téléchargeables, utilisables, combinables



- Données **R**éutilisables

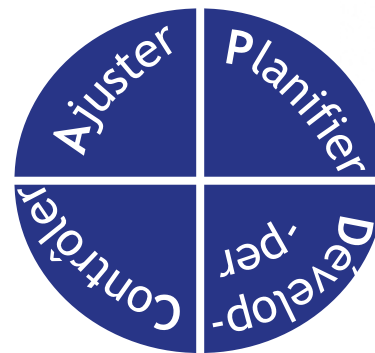


.xml, .csv, .ops (données tabulaires) ;
.pdf, .txt, .docx, .odt, .rtf, (données textuelles) ;
.gif, .png, .jpg (image) ;
.ora, .xcf (retouche d'images) ;
.mp3, .wav, .zip. (sons) ;
.mp4 (vidéos)

Plan de Gestion de Données (PGD) = Data Management Plan (DMP)

- Document synthétique et évolutif explicitant la façon dont vous obtenez, documentez, analysez et utilisez vos données à la fois au cours de votre recherche et une fois le projet terminé.
- Basé sur le cycle de vie des données
- Anticipe les problèmes
- Démontre que les données sont gérées selon des « bonnes pratiques »

Objectif Qualité !



En pratique...

Etape 1 : le PDG (DMP)



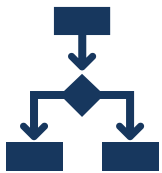
Description

- Projet (résumé, participants, financeurs)
- Identification responsable(s) gestion de données
- Ressources mobilisées pour PGD



Données

- Nature (collectées ou créées)
- Modalités de collecte/création
- Nature des métadonnées



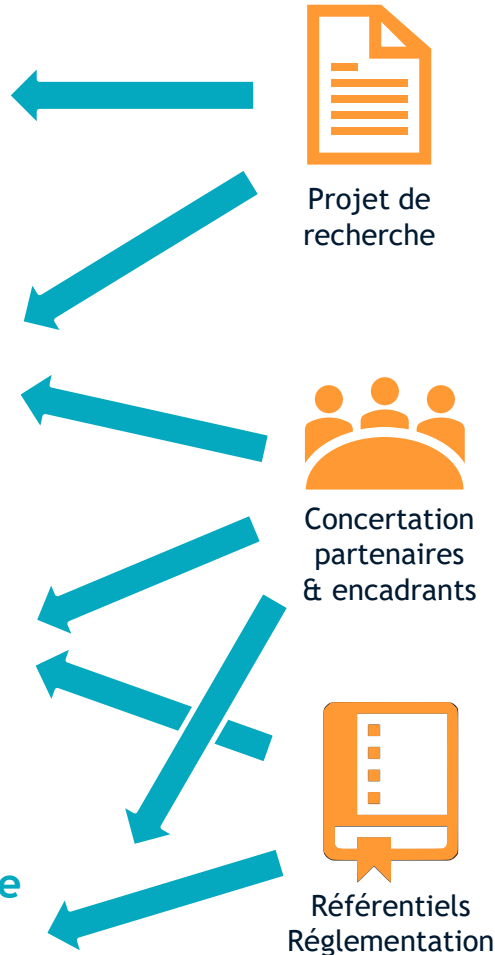
Infrastructures et procédures

- Modalités de stockage durant projet
- Modalités d'accès/partage/sécurité
- Gestion (conventions nommage, indexation)
- Enjeux éthiques (si données personnelles)



Avenir des données en fin de projet

- Identifications des données conservées/partagées
- Modalités de conservation/archivage
- Modalités de partage
- Restrictions éventuelles



Gestion des données par la qualité

- Première version en **amont du projet** (en parallèle au projet scientifique)
- Incite à ce poser les bonnes questions en matière de gestion des données
- Mise à jour au fil de l'eau



PGD : Examples



► PGD de projet

<https://dmp.opidor.fr/plans/23436/export.pdf>

[illegible]

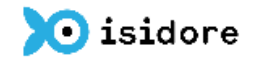
► PGD de thèse

<https://doi.org/10.3897/rio.2.e8664>

[illegible]

Etape 2 : La collecte de données

- ▶ Utiliser des **jeux de données existants**
 - ▶ Données publiques : accès et utilisation facilités
 - ▶ Autres données : accès et utilisation sous conditions
- ▶ **Créer ses jeux de données**
 - ▶ Compilation de données
 - ▶ Expérimentation
 - ▶ Enquêtes
 - ▶ Collecte de Données à Caractère Personnel (DCP) ?
→ RGPD ⇔ Mise en conformité du traitement
 - ▶ Sinon : secret statistique, secret des affaires, etc.
- ▶ **Organiser ses données : structurer les fichiers**



Utiliser les outils adaptés à la collecte et au traitement des données

Quels outils ? Comment bien les choisir ?

► Critères à prendre en compte :

- Faciliter le suivi des traitements : **script** spécifiant les données, fonctions et paramètres utilisés
- Limiter les risques et garantir l'intégrité des données
- **Sécurité** des données : outils européens, institutionnels

► Outils institutionnels à disposition :

- LimeSurvey :



<https://enquetes.univ-lorraine.fr/admin/> | <https://sondages.inrae.fr> |
<https://enquetes.agroparistech.fr> | <https://sondagesv3.unistra.fr/admin/index.php>

- Stockage durant traitement :

<https://bul.univ-lorraine.fr> (20 GB) et espaces collaboratifs | <https://mycore.core-cloud.net/> (100 GB)
<https://seafile.agroparistech.fr> (100 GB) | <https://seafile.unistra.fr> (500GB)



Focus : La collecte de données à caractère personnel

- ▶ Se mettre en conformité avec le **Règlement Général sur la Protection des Données (RGPD)** = déclarer votre traitement (collecte) auprès du Délégué à la Protection des Données (DPD ou DPO, Data Policy Officer) du BETA pour l'inscrire au Registre Informatique et liberté.
- ▶ Pourquoi ? Garantir le **droit des personnes** et une utilisation respectueuse de la vie privée, maîtriser les risques liés au traitement
- ▶ Principe : les données appartiennent aux personnes => elles doivent donner leur **accord** pour que vous puissiez les utiliser et être informées de leurs droits (information, accès, rectification, effacement, portabilité)
- ▶ Etapes :
 - AVANT la collecte : remplir le formulaire et l'envoyer à la DPD CNRS
 - PENDANT la collecte : informer les personnes (finalité du traitement, base légale choisie, droits, réutilisation éventuelle) + recueillir le consentement si c'est la base légale choisie (voir trames Fiche d'information et Consentement)
 - APRES le traitement : anonymisation, conservation sur la durée prévue, suppression

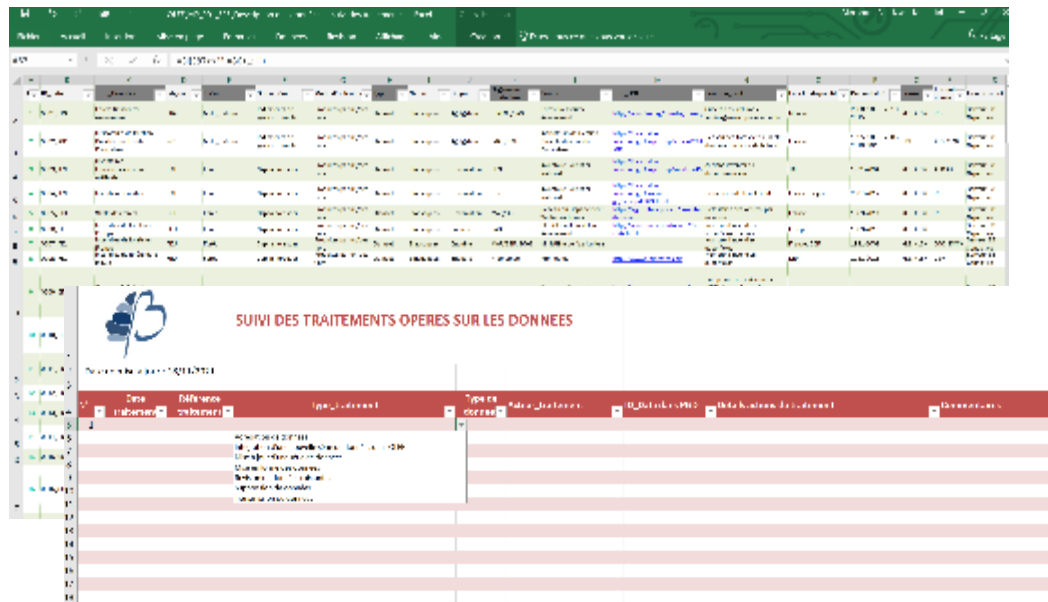
Etape 3 : Les métadonnées

- ▶ La documentation associée aux données permet de répondre aux questions suivantes :
 - Dans quel **but** ont été collectées les données ? Par qui ?
Auteurs et rôles de chacun ; institutions ; financement
 - **Comment** ont été collectées les données ?
*Base publique ou achetée ? Data mining ? Enquêtes ou entretiens ? Sources multiples ?
Protocole de collecte, méthode d'échantillonnage, logiciels utilisés*
 - **Description** du dataset
*Format, type de données
Quelles sont les variables utilisées ? Codes, libellés, unités, modalités*
 - Quels **traitements** avez-vous apporté ? Comment passer des données brutes aux données raffinées ?
*Nettoyage, codification, pseudonymisation, etc.
Logiciels utilisés*
 - **Contrôle** qualité
Procédures mises en œuvre pour assurer la qualité et l'intégrité des données
 - **Accès** aux données
DOI, embargo, licence, restrictions éventuelles, citation

	A	B
1	Métadonnées	Format
2	Nom du projet	Texte libre
3	Titre développé	Texte libre
4	Nom du porteur du projet	Nom prénom
5	Institution	Texte libre
6	Titre du jeu de données	Texte libre. Le titre doit être explicite.
7	Identification du jeu de données	Nom du fichier
8	Description du jeu de données	Description simplifiée du contexte de production de données
9	Date de création	YYYY-MM-DD (Norme ISO8601, W3CDTF)
10	Responsable du fichier	Nom prénom
11	Laboratoire de rattachement du responsable de fichier	Texte libre ou liste contrôlée
12	Adresse mèl du responsable de fichier	
13	Producteur des données	Nom prénom
14	Laboratoire de rattachement du producteur de données	Texte libre ou liste contrôlée
15	Adresse mèl du producteur de données	
16	Langue des métadonnées	fr ou en (Norme ISO639-1)
17	Thématique scientifique	Catégorie issue des Thèmes INSPIRE (cf. Annexe I) Mots-clés issus de thésaurus ou classifications
18	Nom du point de collecte ou d'observation	Texte libre
19	Géolocalisation du point de collecte/prélèvement	Coordonnées GPS
20	Système de coordonnées utilisé	Lambert
21	Date/heure de collecte	YYYY-MM-DD ou YYYY-MM-DDThh:mmTZD (TZD : désigne zone Horaire. Pour la France +01 :00 en hiver, +02 :00 en été) Norme ISO8601, W3CDTF
22	Nom du protocole*	
23	Version du protocole*	
24	Description du protocole*	Description ou référence bibliographique
25	Paramètres du protocole*	Paramètres appliqués dans la méthode
26	Composants du protocole*	Instruments, logiciels ou scripts
27		
28	*Les champs concernant les protocoles peuvent être répétés si plusieurs protocoles ont été successivement	

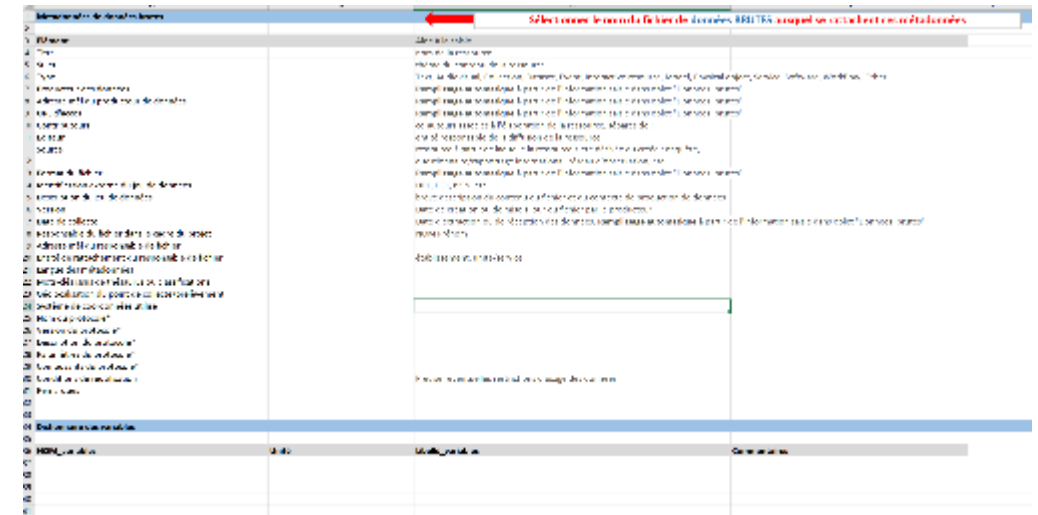
Pour une mise en œuvre complète

- Faciliter la gestion des données et la constitution de dataverse complet (données + métadonnées)



The screenshot shows a data management interface. The top part is a table listing various data files with columns for file name, date, and other metadata. Below this, there is a section titled "SUIVI DES TRAITEMENTS OPERES SUR LES DONNEES" (Tracking of operations performed on the data). This section contains a table with columns for "Date", "Différence", "Type de donnée", "Date de création", "Date de mise à jour", "Date de suppression", and "Date de validation". The table lists several data processing operations, including "Importation de données", "Exportation de données", "Mise à jour de données", and "Suppression de données".

+



The screenshot shows a metadata management interface. The top part is a table listing various metadata records with columns for record name, date, and other metadata. Below this, there is a section titled "MÉTADONNÉES" (Metadata). This section contains a table with columns for "Date", "Différence", "Type de donnée", "Date de création", "Date de mise à jour", "Date de suppression", and "Date de validation". The table lists several metadata records, including "Importation de données", "Exportation de données", "Mise à jour de données", and "Suppression de données".

Métadonnées

Listing des données brutes et travaillées du projet (cf. PDG)
+ Suivi des traitements (script ou manuel)

- (1) **Qualité** des données et des résultats
 = (2) **Rédaction** rapports/articles/data papers facilitée
 (3) **(Ré)Utilisation** facilitée des données

Etape 4 : Stockage et conservation des données

- ▶ Données « chaudes » [édition et **stockage** au cours du projet]
 - Règle des 3 2 1** = 3 copies identiques sur 2 supports différents (USB, disque dur externe, serveurs institutionnels - adaptés pour données sensibles, clouds) dont 1 dans une localisation différente (hors de son bureau).
 - ▶ Serveur permettant sauvegarde, partage / collaboration
 - ▶ **Structurer**, hiérarchiser les dossiers et fichiers
 - ▶ Règles de nommage
 - => **Rendre accessibles et lisibles par tous (humains et machines)**
- ▶ Données « tièdes » ou « froides » [**conservation** après la fin du projet]
 - ▶ Serveurs institutionnels dédiés (data-beta sur Nextcloud INRAE, PETA à l'UL pour données volumineuses)
 - ▶ Mieux : **entrepôts** de données (thématiques, généralistes)
- ▶ **Archivage sécurisé à (très) long terme** : CINES (coûteux car maintenance)

Pourquoi choisir les espaces institutionnels ?



Stockage adapté

- Dossier dédié au projet
- A organiser selon nos choix (règles de nommage)
- Paramétrage selon nos besoins

Stockage sécurisé

- Accès connu et sécurisé (LDAP)
- Partage simple et contrôlé (cercle d'utilisateurs)
- Synchronisation / Sauvegarde régulière

Gestion facilitée des données

- Centralisation
- Standardisation

Limiter l'empreinte carbone des données

Sobriété
numérique

- ▶ Ne générer que des **données utiles**
- ▶ **Trier** les données à conserver !
 - Réduire le volume de données : inutile de conserver toutes les données intermédiaires (données non-essentiels)
 - Ne conserver que les données froides : données brutes + résultats
- ▶ Partager via des **liens**
 - Evite l'envoi d'emails (plus ou moins sécurisés)
 - Maîtrise des droits d'accès (lecture/téléchargement/modification, durée de partage)
- ▶ **Compresser** vos fichiers
- ▶ **Limiter les duplications** superflues de données
- ▶ **Archivage** sur bandes magnétiques comme au CINES : moins coûteux à grande échelle que le stockage sur disques, plus fiables sur le long terme et moins énergivore



Etape 5 : Partage des données & résultats

► HAL pour les articles/communications Entrepôts pour les données

- DOI pour le jeu de données
- DOI pour le **data paper**
- DOI pour l'article scientifique

► Choix de la **licence**



Etalab 2.0 : Réutilisation libre sous réserve de sa source/auteur et la date

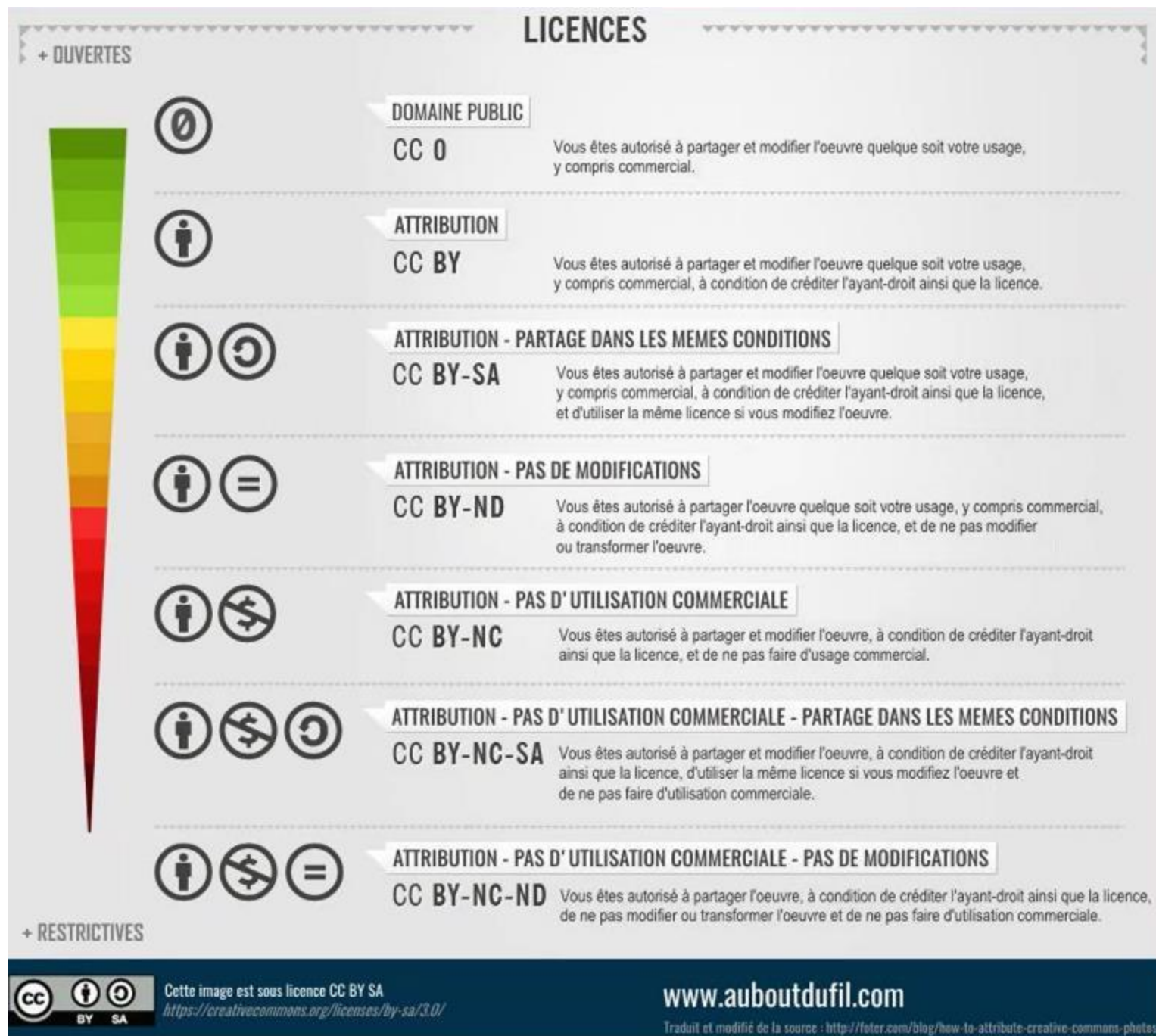


Licences Creative Commons



3 fois plus de visibilité
Liens entre données et productions scientifiques

Data paper =
Métadonnées
mises en forme,
publiées dans une
revue dédiée



Publier un data paper

- ▶ Pratique encouragée par le Plan national pour la Science Ouverte
- ▶ Un « *data paper* (ou *data article*) est une **publication scientifique**, soumise à la validation des pairs, dont le but principal est de décrire un ou plusieurs jeux de données, afin de permettre leur réutilisation. » (Dataacc)
- ▶ Possibilité de générer une ébauche de data paper directement à partir de Recherche Data Gouv (RDG).
- ▶ Même processus que pour les articles scientifiques standards :



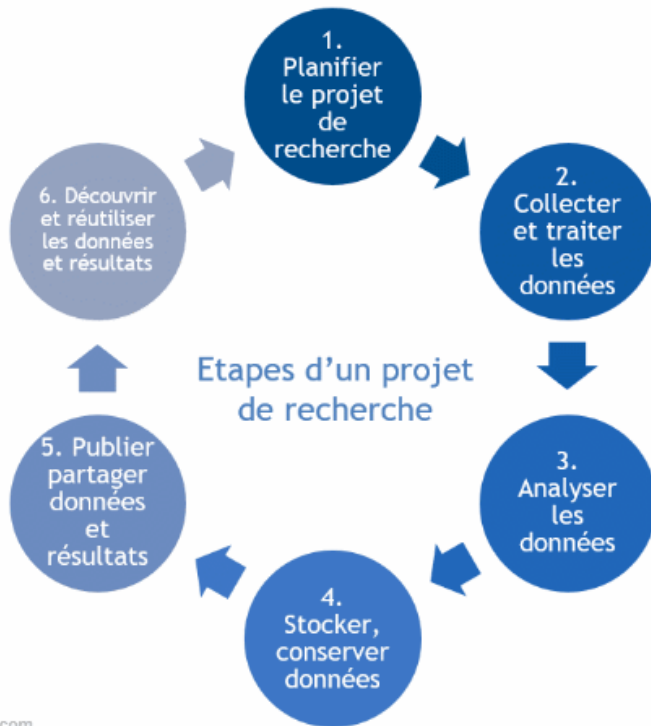
- ▶ « **Ouvert à tous types de données**, sous réserve que le jeu de données ait une valeur scientifique, ou un impact environnemental, sociétal, économique, etc. » (CoopIST-Cirad)
- ▶ « Informe la communauté scientifique de l'existence, de l'originalité, de la qualité et de la disponibilité d'un jeu de données => **valorise le travail des auteurs** en expliquant l'importance des données produites et leur potentiel de réutilisation pour de futures recherches. » (CoopIST-Cirad)

Etape 6 : Fin du projet



Objectif en vue !

- Mettre à jour la version finale du **PGD**
- **Initier de nouveaux projets** en réutilisant vos données
=> nouveau PGD, nouveaux traitements, nouveaux résultats à valoriser, etc.



A vous de jouer !

Cette photo par Auteur inconnu est soumise à la licence [CC BY-SA](#)

**Pour vous
accompagner...**

La plateforme OLEF

- ▶ Plateforme d'appui du BETA, dédiée à la forêt et au bois... mais pas que !
- ▶ Missions :
 - ▶ **Collecter et pérenniser** les données pour la recherche, l'expertise et l'enseignement
 - ▶ **Valoriser les données**
 - ▶ **Appuyer les chercheurs au quotidien**
 - ▶ Appui technique : informatique, logiciel, stockage
 - ▶ Conseils pour respecter la réglementation en vigueur : RGPD, plan de gestion de données
 - ▶ Bonnes pratiques liées aux données : rendre FAIR les données de la recherche dans le cadre de l'Open Science, nommage, archivage, data paper, etc.
- ▶ Labellisé StAR dans le cadre du programme INFRA+ LUE
- ▶ Atouts :
 - ▶ **Expérience de la plateforme** dans la gestion des données
 - ▶ **Outils** pour assurer la qualité des processus
 - ▶ **Intégration** dans les réseaux de Référents #Données, #ScienceOuverte



Contact : beta-olef@inrae.fr

Groupe « Sciences ouvertes » du BETA

► Les volontaires constituant le groupe :

Stefano Bianchini , Sandrine Wolff

Cécile Bourreau-Dubois, Sébastien Massoni, Julie Mansuy,
Agnès Gramain (en disponibilité jusqu'au 31/08/2024)

Claire Montagné-Huck, Alexandra Niedzwiedz

► Objectifs :

Accompagner dans la gestion de données (de l'idée à l'archivage)

Sensibiliser à l'ouverture de la recherche (aussi ouvert que possible mais aussi fermé que nécessaire), dans une perspective de reproductibilité

=> Définir la politique « Science Ouverte » du BETA

=> **Valoriser** l'activité de production de données et faciliter le partage maîtrisé des données

Ateliers de la Donnée ou guichets uniques en région

► **ADOC** - Accompagner aux Données les Chercheurs et chercheuses en Lorraine

Établissements partenaires : Université de Lorraine, CNRS, Inria, AgroParisTech

Taille de l'équipe (nombre de personnes mobilisées) : 8

Coordinateur/pilote : Bracco, Laetitia

Contact : donnees-recherche@univ-lorraine.fr

► **ADELE** - Atelier Donnée aLsacE Helpdesk

Établissements partenaires : UNISTRA, UHA, INSA, BNU, MISHA...

Taille de l'équipe (nombre de personnes mobilisées) : 29

Coordinateur/pilote : Cheviron, Stéphanie

Contact : contact@adele.help



=> **Accompagnement** au plan de gestion des données ; conseil sur la gestion FAIR et la diffusion des données et du code, sur le stockage et le calcul ; accompagnement au RGPD ; accompagnement au dépôt de jeux de données en entrepôt et curation ; formations aux données de la recherche, etc.

Délégués à la protection des données

DPD du BETA : *(en cours d'officialisation)*

CNRS - Service Protection des Données

2 rue Jean Zay 54519 Vandoeuvre-lès-Nancy Cedex

Téléphone : +33 3 83 85 64 26

Courriel : dpd.demandes@cnrs.fr

Intranet : https://intranet.cnrs.fr/protection_donnees/donnees/Pages/default.aspx



Autres tutelles :



Merci pour votre attention !

Contact

Observatoire pour L'Economie de la Forêt :

beta-olef@inrae.fr

Site Web :

<https://beta-economics.fr/plateformes/olef/>

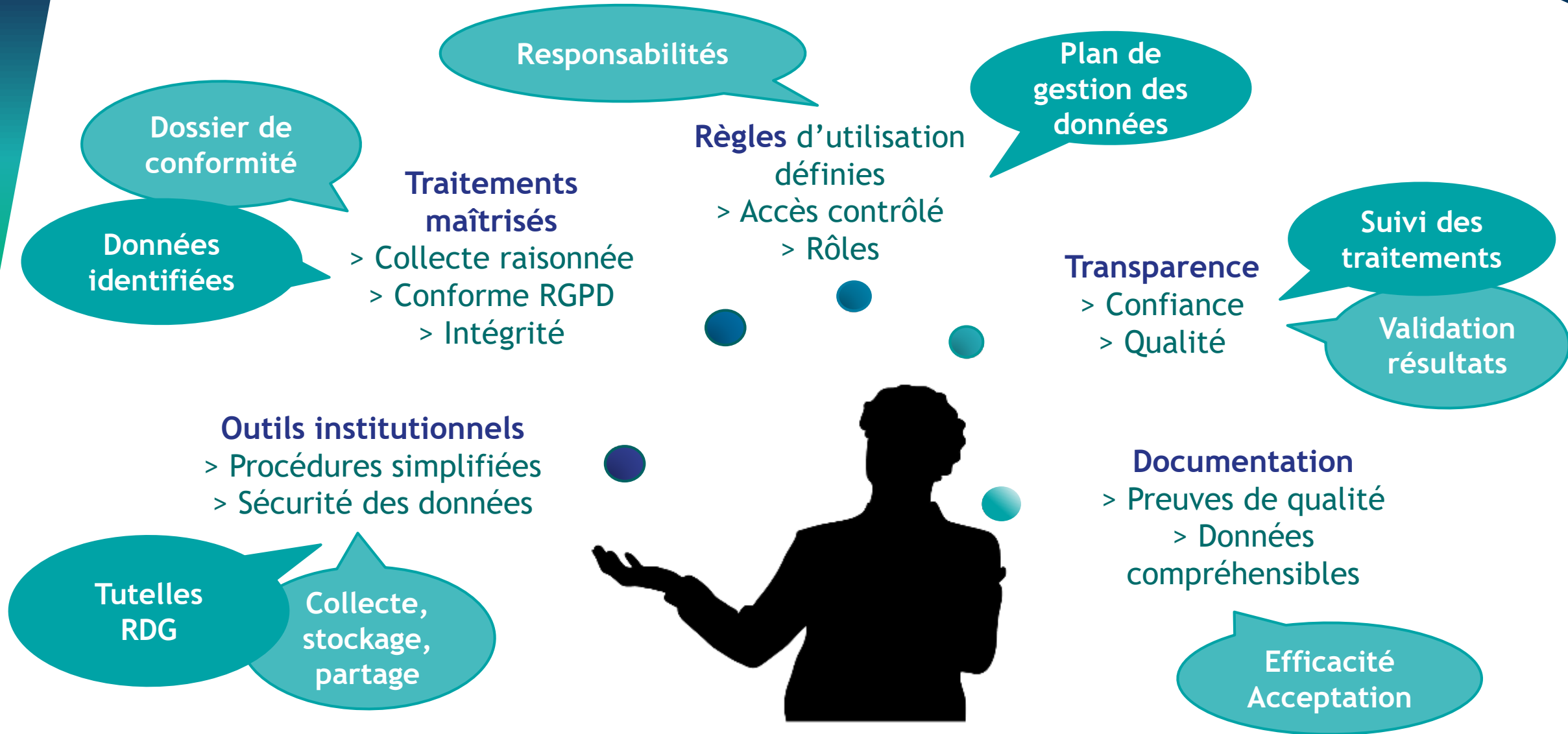


Etalab 2.0 : Réutilisation libre sous réserve de mentionner sa source/auteur et la date du présent support.



**En annexe :
Mémo et références**

L'importance de la maîtrise des traitements



En bref...

À éviter !

- ✗ Se lancer tête baissée dans la collecte de données
- ✗ Stocker les données (uniquement) sur le disque dur de son pc
- ✗ Variables et/méthode de production non renseignée
- ✗ Conserver uniquement ses données sous un format nécessitant un logiciel payant
- ✗ Diffuser des données non anonymisées sans l'accord de leur propriétaire
- ✗ Déposer ses données dans des entrepôts privés ou d'éditeurs (⇔ cession des données ou transfert de propriété)



À privilégier

- ✓ Rédiger un Plan de Gestion des Données
- ✓ Collecter les données de manière raisonnée
- ✓ Stocker les données sur au moins deux supports différents, dont un sera un serveur institutionnel
- ✓ Documenter ses données pour faciliter leur réutilisation (par soi-même et par d'autres)
- ✓ Formats lisibles à long terme par tous les logiciels (txt, csv)
- ✓ Valoriser le travail de production de données : entreposer dans un espace dédié (rendre visibles ses données), rédiger des data papers
- ✓ Utiliser une licence Etalab ou CC-BY



Guides et templates mise à disposition sur :
<https://nextcloud.inrae.fr/s/FkB2RFkfSZr2Qx7>



Références et liens utiles

- ▶ Deboin, M.C. 2018. Se familiariser avec les plans de gestion de données de la recherche, en 6 points. Montpellier (FRA) : CIRAD, 6 p.
<https://doi.org/10.18167/coopist/0056>
- ▶ Hadrossek, C., Janik, J., Libes, M., Louvet, V., Quidoz, M-C. et al. 2021. Guide de bonnes pratiques sur la gestion des données de la Recherche. <https://hal.archives-ouvertes.fr/hal-03152732>
- ▶ Reymonet, N., Moysan, M., Cartier, A., Délémontez, R. 20108. Réaliser un plan de gestion de données “ FAIR ” : modèle.
https://archivesic.ccsd.cnrs.fr/sic_01690547v2
- ▶ Cat OPIDoR, wiki des services dédiés aux données de la recherche <https://cat.opidor.fr/index.php/>
- ▶ Digital Curation Centre (DCC) : <https://www.dcc.ac.uk/resources/data-management-plans>
- ▶ DMP OPIDOR : <https://dmp.opidor.fr/>
- ▶ DORANum : <https://doranum.fr/metadonnees-standards-formats/outils-creation-metadonnees/>
- ▶ Fiches AgroParisTech « **Gestion et Ouverture des Données de la Recherche** »
<https://zenodo.org/communities/agroparistechdr?q=&l=list&p=1&s=10&sort=newest>
- ▶ Guide en ligne « Données de la recherche » <https://sciencespo.libguides.com/donnees-de-la-recherche>
- ▶ Mise en œuvre d'un traitement de DCP dans le cadre d'une thèse <https://donnees-personnelles.intranet.inrae.fr/avant-de-collecter-ou-de-re-utiliser-des-donnees/bonnes-pratiques-et-situations-frequentes-faq/mise-en-aeuvre-d-un-traitement-de-dcp-dans-le-cadre-d-une-these>
- ▶ Outil d'anonymisation mis à disposition des chercheurs <https://amnesia.openaire.eu/>
- ▶ Ouvrir la science > Passeport pour la Science Ouverte - **Guide pratique à l'usage des doctorantes et doctorants (2022)**
<https://www.ouvrirlascience.fr/passeport-pour-la-science-ouverte-guide-pratique-a-lusage-des-doctorants/>
- ▶ R - Epi Handbook <https://epirhandbook.com/fr/>
- ▶ Intranet BETA : <https://labs.core-cloud.net/ou/BETA-Lab/ESPACECOLLECTIF/SitePages/OpenScience-DataManagement.aspx>