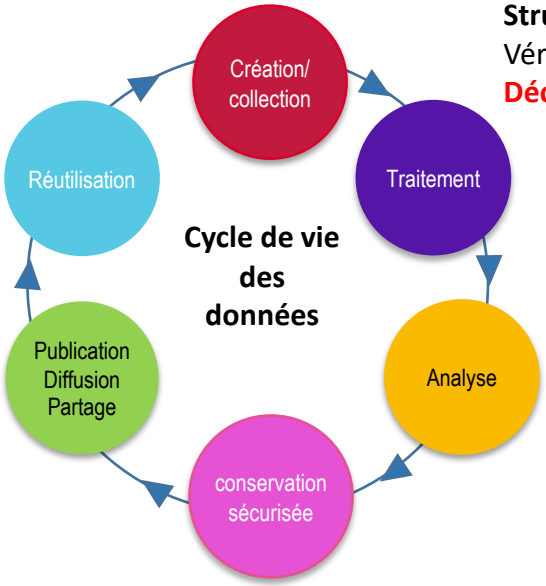
Gestion des données stockage, sauvegarde et Partage

Le cycle de vie de la donnée

Diffusion/Partage :
sous quelle forme?
data papers, datasets,
Par quel moyen?
En local, entrepôt national, disciplinaire, pluridisciplinaire.
Définir le niveau d'ouverture:
de l'ouverture interne à l'ouverture publique
Définir les conditions d'utilisation des données (licences)



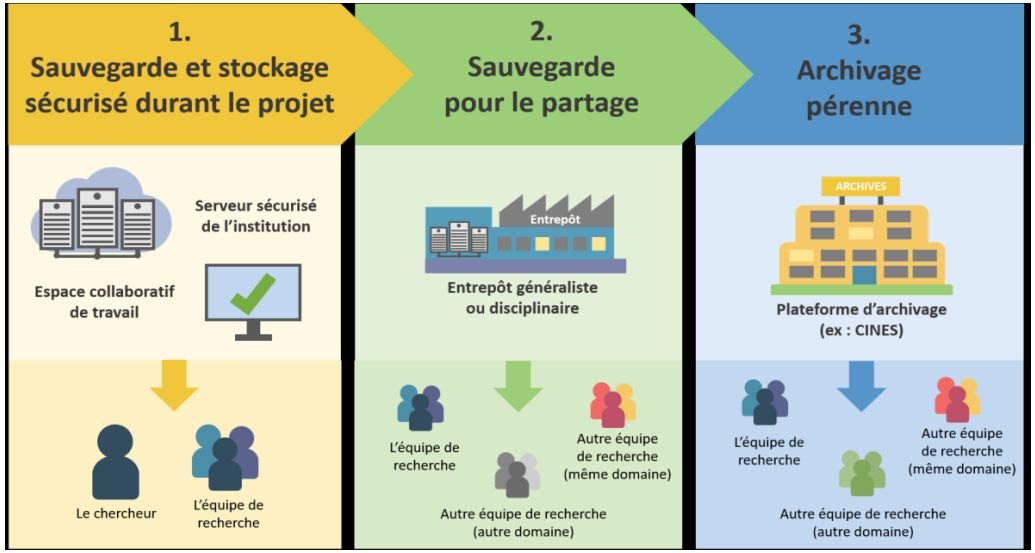
Structurer, organiser
Vérifier, nettoyer, valider
Décrire les données (métadonnées exp, contexte)

Éventuellement protéger (brevet, cryptage)
Préparer les données pour la préservation, la valorisation
Partager en local (centraliser)

sauvegarder les données (au meilleur format et un support adapté)
Compléter les métadonnées
Trier les données: données exploitables,
Estimer le potentiel de réutilisation des données

la sauvegarde et le stockage des données

source : DoRANum (Données de la Recherche Apprentissage Numérique), les 3 niveaux de sauvegarde



Environnement de travail sûr :

Poste de travail

Règle du 3-2-1 : 3 copies/ 2 Systèmes différents dont au moins un distant

Mise à jour régulière

Antivirus à jour

Chiffrage des données (en cas de vol)

Solutions de stockage :

Stratégie 3-2-1

Pérennité en phase avec les besoins

Mise à jour régulière

Les systèmes de sauvegarde individuels proposés par nos tutelles

OneDrive inrae (login + mot de passe LDAP national) partage-fichiers.inrae.fr
NextCloud inrae (login + mot de passe LDAP national) nextcloud.inrae.fr
MyCore cnrs (connexion via janus avec votre adresse mail – à activer via sesame) mycore.cnrs.fr
NextCloud supagro (prenom.nom@supagro.fr + mot de passe du LDAP local) nuage.supagro.fr

Tout les permanents, CDD & doctorants ont accès à ces espaces.

Les systèmes de sauvegarde proposés sur le campus

Sys. de fichiers	Taille	Utilisé	Dispo	Uti%	Données
U : //cubebe/bpmp	800G	590G	211G	74%	chaudes
Q : //backup1/backup-bpmp	16T	15T	1,8T	90%	perso (rsync)
//smbcampus/homes	3,0G	0	3,0G	0%	perso
S : //stocka2/bpmp-unites	11T	11T	719G	94%	froides

U : données sauvegardées grâce à des snapshots toutes les 4h (entre 8h et 20h du lundi au vendredi).
nb snapshot limité à 64 => restauration pouvant remonter à un peu plus de 4 semaines.
En plus ce répertoire est sauvegardé sur bandes magnétiques dans une salle informatique tiers avec une rétention de 6 mois.

S : données non sauvegardées. Cet espace peut être augmenté (Henrique)

Q (rsync) : données sauvegardées par Henrique sur les NAS dans son bureau

Coût pour le laboratoire : 40 euros / To / an

Sélection des données à sauvegarder

Critères à prendre en compte :

- quand et à quelle fréquence faire cette sélection (fin de contrat/projet, à date régulière)
- durée de conservation des données sélectionnées
- obligation administrative
- formats sous lesquels les données sont conservées (lisibilité dans le temps)
- sur quels supports
- niveau de sécurité et d'accès aux données
- coût

Penser à l'obsolescence matérielle, logicielle, des formats de fichiers et au manque de documentation qui font que des fichiers sauvegardés bien soigneusement sont impossibles à réutiliser

Les systèmes de partage de données : outils collaboratifs

Espace de travail où plusieurs personnes peuvent déposer/modifier des documents

- NextCloud inrae
- Sharepoint inrae
- Espace CoRe cnrs

Ces outils sont à activer par le laboratoire

Pour partager des données au sein de BPMP, utiliser le disque **U** (//cubebe/bpmp) situé physiquement à l'EIC

Avantages des outils collaboratifs

Tous les participants à un projet ont accès aux données de tous

Valorisation du travail de chacun

Favorise les discussions entre tous les acteurs du projet

Favorise la ré-utilisation des jeux de données

Sauvegardes à l'IPSiM

Les systèmes proposés par le campus et nos tutelles suffisent dans la majorité des cas

Il faut néamoins que le laboratoire active les outils collaboratifs

Pour la solution NextCloud d'inrae, attente de la livraison de matériel (combien de temps ?)

Cas des plateformes produisant de grosses données (imagerie, phénotypage, protéomique, autres ?) : plusieurs To

Quelle quantité de données ? Qu'est-ce qu'on garde ? Combien de temps ?

Protéomique : solution en interne

Imagerie : chaque utilisateur se débrouille

Stockage Meso@LR

https://meso-lr.umontpellier.fr/fonctionnement-du-centre/

MESO@LR

Demande d'accès au service Stockage

Vous devez compléter le document suivant et le retourner par email à l'adresse indiquée. N'hésitez pas à nous contacter si vous avez des questions.

Offre de service stockage

Document de demande de ressources stockage

Une analyse de votre besoin permettra de déterminer le tarif et les modalités qui correspondent le mieux à votre demande.

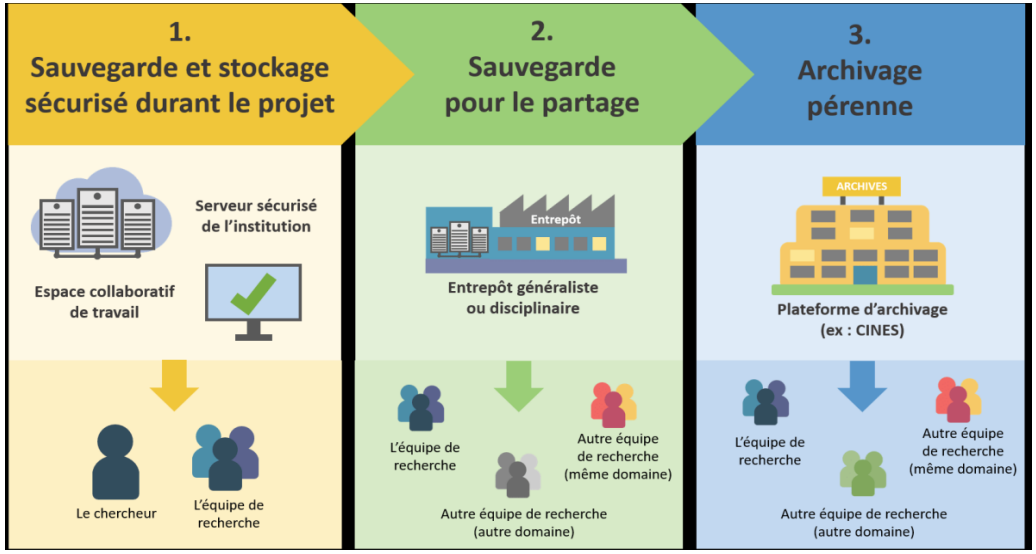
Chaque utilisateur doit accepter la Charte informatique DONNEES avant d'avoir accès aux ressources du mésocentre.

Tarif Stockage

Tarifs stockage / To / an (2)	Laboratoires de l'Université de Montpellier & Partenaires académiques MUSE	Académiques Autres	Privé
To Associé au calcul	40,00 €HT	45,00 €HT	120,00 €HT
To par an (sans minimum)	50,00 €HT	55,00 €HT	150,00 €HT
To par an > 100 To	45,00 €HT	50,00 €HT	135,00 €HT
To par an > 1000 To (soit 1 Po) (1)	30,00 €HT	35,00 €HT	90,00 €HT

Le partage des données

source : DoRANum (Données de la Recherche Apprentissage Numérique), les 3 niveaux de sauvegarde



Quelles obligations de partage des données ?

- *a priori**, les données issues d'une activité de la recherche sont soumises à une obligation de partage

  Les données sont soumises à un principe d'ouverture par défaut et de libre utilisation *(Loi République numérique 2016 LPRN)*

Aspects juridiques et éthiques

https://youtu.be/AVOMdmMQjb4

Autres ressources:
MOOC Sciences Ouverte: Aspects juridiques (page intranet)
Warusfel, Bertrand , “Propriété intellectuelle et droits sur les données de la recherche,” *Bibliothèque numérique Paris 8*, https://octaviana.fr/document/VUN0040_06.

Ce qu'il faut retenir

- En principe, les données produites par les chercheurs dans leur activité de recherche sont **des données publiques,** si elles sont en majorité financées sur des fonds publics
  (et doivent être conservées , afin de « permettre leur vérification » (nouvel art. L211-2 CRech)
- **Doivent être diffusées** et pouvoir **être réutilisées gratuitement**

mais il existe des **exceptions**:

- Droit d'auteur, pour les œuvres de l'esprit
  - publications scientifiques : Les chercheurs sont considérés comme des tiers par rapport à l'administration. Ils décident de la manière de diffuser, de partager leur œuvre.
  - image, qui sous certaines conditions d'originalité, peut aussi être considérée comme une œuvre
- RGPD (Données personnelles, données sensibles),
- secrets protégés (partenariat), (brevets)

➢ Données « Ouvertes autant que possible et fermées autant que nécessaires »

Un récapitulatif (non exhaustif) de la législation applicable

- en France
  - Loi Lemaire - LPRN 2016 *(ouverture des données)*
  - Loi Valter 2015 *(gratuité)*
  - RGPD 2018 *(protection des données sensibles)*
- en Europe
  - Directive on open data and the re-use of public sector information, **Open Data Directive**, 16 July 2019
  - Regulation on a framework for the free flow of non-personal data in the EU (EU) 2018/1807 (*free movement of non-personal data*)
  - GDPR 2018
- notons la précédence de la LPRN

Choix d’un entrepôt

https://doranum.fr/depot-entrepots/formation

Autres outils pour trouver des entrepôts

DataCite Repository Selector

Cat OPIDoR

Catalogue services français (entrepôts)

Questions diverses gestion données

RDM kit

Modalités de partage, licences, métadonnées, standards…

Les entrepôts de données

- **Institutionnels** Dataverse
- **Europe** Zenodo, B2Share
- **Généralistes** Figshare, Dryad
- **Editeurs** Oxford Univ Press (GigaDB) ; Ubiquity Press (Dataverse)
- Recherche Data Gouv, the federated national research data platform
- **Thématiques**
  - GBIF (Global Biodiversity Information Facility)
  - KNB (Knowledge Network for biocomplexity), OBIS (Ocean Biogeographic Information System), BioFresh (Freshwater Biodiversity Data Initiative)
  - Pangaea, SEANOE
  - Movebank, WormBase, ViPR, MycoBank, ComBase, FLOW
  - GenBank, Barcode of Life Data Systems, UniProt, Intact
  - Dataverse, ICPSR, DataFirst, Quetelet, beQuali

Critères de choix d'un entrepôt

Les caractéristiques, fonctions et exigences de l'entrepôt conviennent-ils à ma situation?

Caractéristiques

- Pérennité de l'entrepôt
- Facilité de dépôt
- Facilité de recherche / découverte des données
- Localisation du serveur
- Qualité de la description des données

Fonctions

- Préservation des données
- Identifiant pérenne
- Gestion des versions
- Traçabilité, provenance
- Statistiques d'usage
- Contrôle d'accès aux données
- Interopérabilité
- Pré-publication (accès reviewers via lien privé)

Exigences

- Limites de la discipline
- Coûts
- Types de données acceptées
- Formats acceptés
- Licences possibles
- Limites volumétrie (Zenodo 2GB, figshare 5GB/fichier)

E. Dzalé Yeumo, DIST INRA, 2016
Les entrepôts de données : pierre angulaire du partage des données de la recherche

Entrepôts certifiés:

Certification de type Data Seal of Approval (DSA) ou CoreTrustSeal : 16 exigences

- Entrepôts FAIR : Garantie d'un bon niveau de compatibilité des données avec les principes FAIR (ex Data INRAE)

Entrepôts non certifiés :

Critères à prendre en compte

- Des identifiants uniques et pérennes (par exemple DOI) associés aux jeux de données
- possibilité d'associer des métadonnées
- L'entrepôt permet-il de mentionner clairement la licence (licence ouverte, CC BY, etc.)
- accessibilité publique des citations et des métadonnées même si données à accès restreint
- métadonnées compatibles avec des standards de métadonnées reconnus
- plan de préservation à long terme des données

Entrepôt institutionnel

(Dataverse)

Accès aux agents INRAE (LDAP)

- Gestion des données : stockage, liaison aux métadonnées,
- Partage : attribution DOI, possibilité de générer template de Datapapers
- Recherche de données (visibilité)

(présentation https://youtu.be/g4fVc5wu0-U)

Dépôts : jeux de données (< 15 Go)
données d'un projet (Dataverse)

Connexion à Nextcloud
Définition des conditions d'accès (restreint ou ouvert) aux fichiers de données
Possibilité de donner accès à collaborateurs non INRAE

- Open class « déposer des données » régulièrement organisées

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

recherche.**data.gouv**.*fr*

La plateforme nationale fédérée des données de la recherche

LANCEMENT
Printemps 2022

DEUX AXES FORTS, CINQ MODULES

Accompagner les équipes de recherche

ATELIERS DE LA DONNÉE
Des experts de la donnée à disposition et à proximité des chercheurs

CENTRES DE RÉFÉRENCES THÉMATIQUES
Un spectre de compétences fondamentales relatives à la donnée et propre à chaque discipline

CENTRES DE RESSOURCES
Un accompagnement technique sur le dépôt de données sur Recherche Data Gouv

Ouvrir et partager les données

ENTREPÔT
Une interface web où déposer ses données avec un espace de modération

CATALOGUE
Un outil pour repérer et moissonner des données issues d'entrepôts externes

DÈS LE PRIMPTEMPS 2022 :

- **ÊTRE ACCOMPAGNÉ** tout au long du cycle de la donnée
- **DÉPOSER & PUBLIER** ses données
- **RECHERCHER** des (méta) données
- **IDENTIFIER** l'atelier de la donnée pour vous accompagner
- **CONSULTER** la liste des entrepôts thématiques de l'écosystème français

CHIFFRES CLÉS

- **4** Millions d'€ alloués aux modules entrepôt et catalogue
- **3** Années consacrées au projet
- **8** Etablissements partenaires
- **44** Personnes mobilisées

Plus d'infos sur
projet-recherchedatagv.ouvrirlascience.fr

@RechercheDataGv

Quelques standards de métadonnées

MIAME: Minimum Information about a Microarray Experiment (Brazma et al. 2001) - MIAME/Plant (TAIR 2005)
MIAPPE: Minimal Information About Plant Phenotyping Experiment - miappe.org
MIAPE: Minimum Information about a Proteomic Experiment (Taylor et al. 2007)
MINSEQE: Minimum Information about a high-throughput SEQuencing Experiment (Brazma et al. 2012)
REMBI: Recommended Metadata for Biological Images - enabling reuse of microscopy data in biology (2021).

- **Exemple de licence Creative Commons (CC)**

BY – **attribution** : vous demandez à être cité si on utilise votre travail
NC – **usage non-commercial** : vous autorisez à utiliser votre travail à des fins uniquement non commerciales
SA – partage **dans les mêmes conditions** : vous autorisez les modifications, corrections et diffusion de votre oeuvre, à condition que l'oeuvre dérivée soit proposée sous la même licence.

Autres licences CC

CC-BY : Licence très permissive - seule obligation : mention de l'auteur
CC-0 : CC-0 et CC-BY ont les mêmes valeur en France et dans les pays où le droit moral est reconnu. Dans les autres pays (États-Unis), la licence CC-0 n'exige même pas de mentionner l'auteur.
CC-BY-ND : interdit toute modification

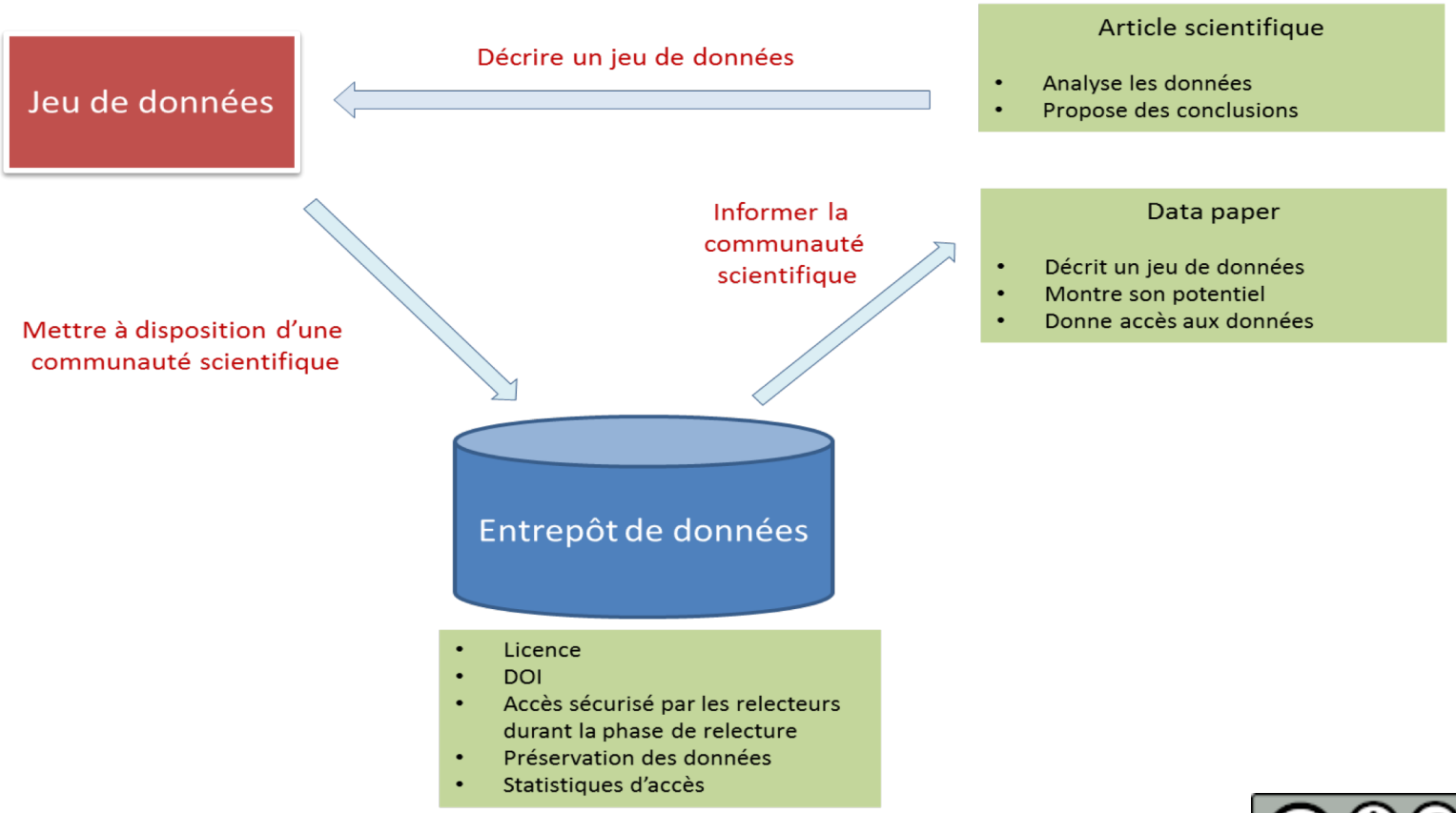
Voir la page sur les licences CC du CoopIST (Cirad)

Licences Etalab

Licence ouverte, libre et gratuite
Autorise la reproduction, la redistribution, l'adaptation et l'exploitation commerciale des données
compatible avec les standards des licences Open Data développées à l'étranger
Obligation de mentionner la paternité

En savoir plus : https://www.etalab.gouv.fr/licence-ouverte-open-licence/

Intérêt du dépôt : Faire connaître les données



D'après Laurence Dedieu, éditrice scientifique, Cirad - Dist

E. Dzalé Yeumo, DIST INRA, 2016

Les entrepôts de données : pierre angulaire du partage des données de la recherche

DATApapers

Décrit un jeu de données (dataset),
La méthode ayant permis de l'obtenir et le potentiel de réutilisation de ce jeu.

Publier un Data Paper permet
•d'informer la communauté scientifique de l'existence et la disponibilité du jeu de données
•de valoriser les données, avec un objectif d'ouverture :
- en leur apportant une bonne visibilité
- en facilitant leur réutilisation par une rédaction soignée des métadonnées
- en explicitant leur potentiel de réutilisation

•d'obtenir une reconnaissance du travail réalisé (crédit aux auteurs), grâce à sa citabilité

Conclusion

Des solutions multiples pour :

- le stockage des données
- le partage des données (articles scientiques, Datapapers, jeux de données)

➢ trouver les plus adaptées à sa situation en fonction:

- du type de données,
- du moment dans le cycle de vie de la donnée
- du contexte de travail

A discuter :

- Besoins et souhaits pour stockage des données volumineuses (images)
  - Solutions en interne ou autres (Meso@LR?)

- Besoins en accompagnement :
  - PGD : exemples validés à disposition? atelier de discussion des questions soulevées par la rédaction?
  - Métadonnées
  - Autres questions?