Mieux gérer les données
De l'obtention à la diffusion

- Pourquoi?

- Comment?
  - Bonnes pratiques
  - Outils

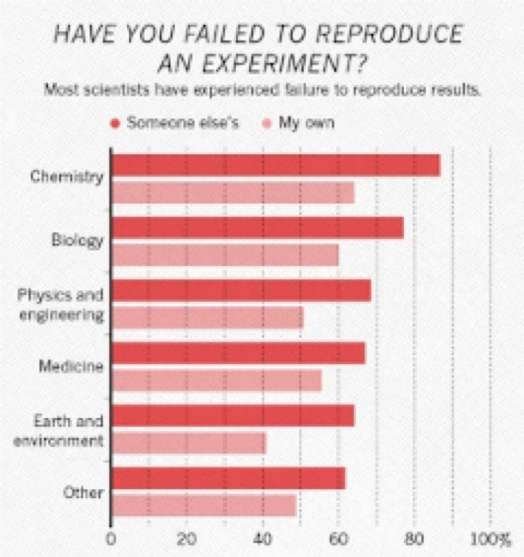
Pourquoi mieux gérer les données?

- **Recherche reproductible**

Augmentation du volume de données / difficulté à retrouver les données

Diminution de la qualité des données / problèmes de reproductibilité

➢ Crise de reproductibilité



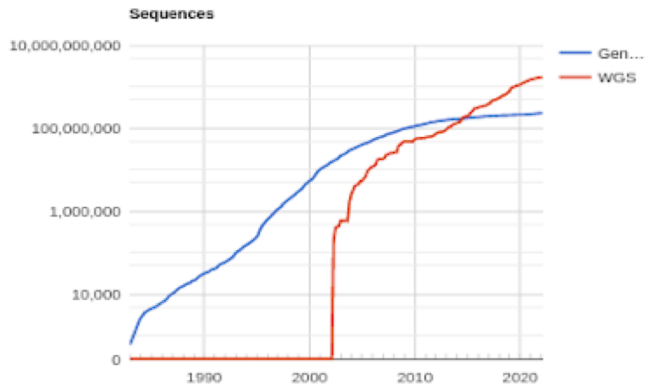
M. Baker *Nature* 533:452–454 (2016)
https://www.nature.com/articles/533452a

Parmi les causes : méthode non documentée ou mauvaise méthode d'analyse , non accès aux données (données brutes / format non utilisable), répétitions …

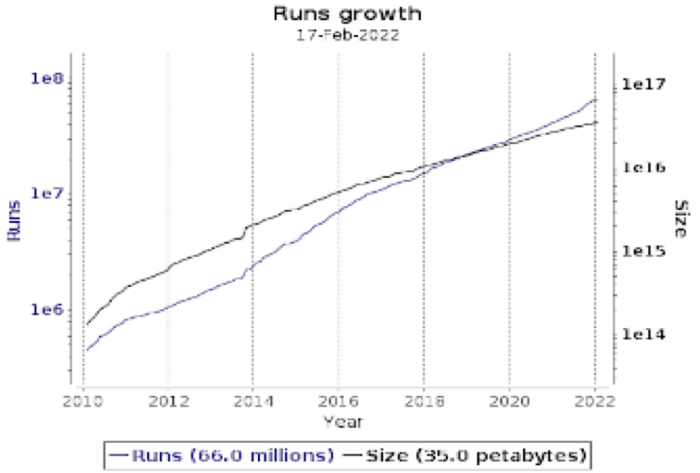
Pourquoi mieux gérer les données?

- Explosion du volume de données numériques (Haut débit) / diminution de la capacité d'analyses

  Coût financier et environnemental



https://www.ncbi.nlm.nih.gov/genbank/statistics/
GenBank = séquences nucléiques du NCBI
WGS = Whole Genome Shotgun



https://www.ebi.ac.uk/ena/browser/about/statistics
European Nucleotide Archive (NGS)

➢ favoriser le partage et la réutilisation des données. Conserver les données exploitables

Contexte de science ouverte:

Loi pour une République Numérique (libre accès aux données), plan national pour la science ouverte, exigences agences de financement)).

Data Sharing and Management Snafu in 3 Short Acts

by Karen Hanson, Alisa Surkis & Karen Yacobucci
NYU Health Sciences Libraries

August 3, 2012 (Last Update: December 12, 2012)

Quelles données?

Les données de la recherche 1.

- définition «standard» : « *des enregistrements factuels (chiffres, textes, images et sons), qui sont utilisés comme sources principales pour la recherche scientifique et sont généralement reconnus par la communauté scientifique comme nécessaires pour valider les résultats de la recherche* » (OCDE 2007)
- néanmoins, définitions moins restrictives :
  - données d'observation, données expérimentales, données computationnelles ou de simulation, données dérivées ou compilées, et données de référence (au delà du «factuel», INIST)
  - aussi, les données ne sont pas seulement des données «*nécessaires à la validation des résultats*» :
    - typiquement, en bioinformatique, les données produites sont plus nombreuses que celles strictement nécessaires pour valider un résultat
    - ces données gagnent en valeur précisément si elles peuvent être partagées

- Plan données de la Recherche (CNRS) : Données numériques, brutes ou reraitées, documents, logiciels, algorithmes, protocoles, ressources biologiques

Quelles données?

Plusieurs personnes

Plusieurs techniques

Plusieurs lieux

Plusieurs années

Ne rien perdre

Pouvoir retrouver

Pouvoir réanalyser

Pouvoir partager

Donc, dans la "vraie vie", gérer quoi ?

- **Le passé**
  - Le leg (du doctorant précédent …)
  - La biblio à T0
  - Les méthodes pré existantes
- **Le présent**
  - Les manipes
  - La création de connaissance (méthodes, posters …)
- **Le futur**
  - Le manuscrit
  - Les publications

- **Des échantillons**
  - dans les frigos
  - dans les tiroirs
- **Des fichiers**
  - des petits, des gros
  - un peu partout (PC, cloud, cluster)
  - des données brutes, du code, des résultats
- **De la connaissance**
  - des méthodes, du code
  - des systèmes d'information
  - des publications

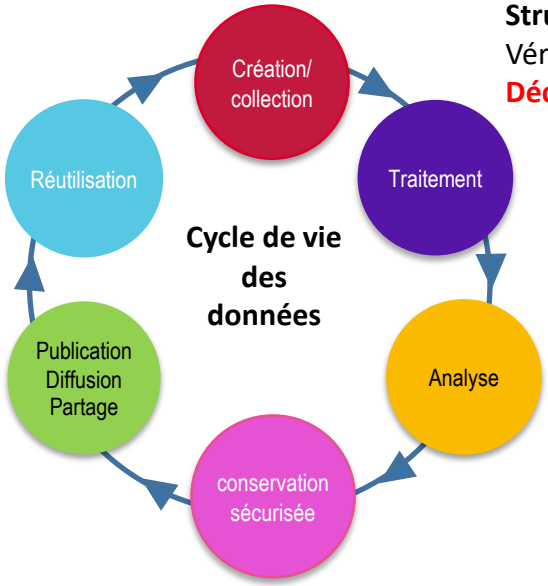
IFB
INSTITUT FRANÇAIS DE BIOINFORMATIQUE

Evaluer le potentiel de réutilisation
Trier et conserver ce qui est suffisamment documenté pour être exploitable

Comment faire?

Adopter de bonnes pratiques tout au long du 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)

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

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

Structurer les données

Objectif : Retrouver, partager entre collaborateurs du projet (éviter la dispersion, confusions de fichiers, perte d'information)

- Structurer les dossiers/fichiers
  - Convenir d'une structuration (arborescence prédéfinie) et la décrire

    Séparer données brutes et données traitées

    Décrire les données et leur traitement (fichier sauvegardé associé aux données, lisez-moi) : métadonnées expérimentales indispensables pour l'interprétation et la reproduction des données, méthodes d'analyse.
- Nommer correctement les fichiers
  - 5 règles de nommage:

- **nom court et explicite** (30 caractères max, plutôt abréviation compréhensible que nom complet)
- **uniquement caractères** alphanumériques sans espace mais avec - ou _ ou Majuscule pour séparer les mots
- **la date au bon format** AAAA-MM-DD (ISO 8601) ou AAAAMMDD
- **Placer l'élément le plus important en premier**
- **Indiquer le N° de version si nécessaire**

Expliquer la convention de nommage dans un document associé aux données

Associer les données à des métadonnées

- pourquoi les données ont été produites,
- liste des échantillons utilisés (âge de la plante, accession, ID de mutant, tissu …),
- liste des protocoles suivis (procédures expérimentales & d'analyses des données) avec les ref associées, (Concernant les analyses bioinfo effectuées, noter le nom de chaque programme utilisé avec son n° de version, les options utilisées)
- nomenclature choisie,
- liste des personnes contact pour chaque étape du projet, les prestataires choisis, ...

=> doit fournir toutes les informations utiles à l'analyse et à la réutilisation des données

La sauvegarde et le stockage des données: des points essentiels

Enjeux du stockage :

- Assurer un accès facilité (qui ?, quand ?, modalités ?)
- Éviter la perte,
- Assurer l'intégrité,
- Garantir la confidentialité

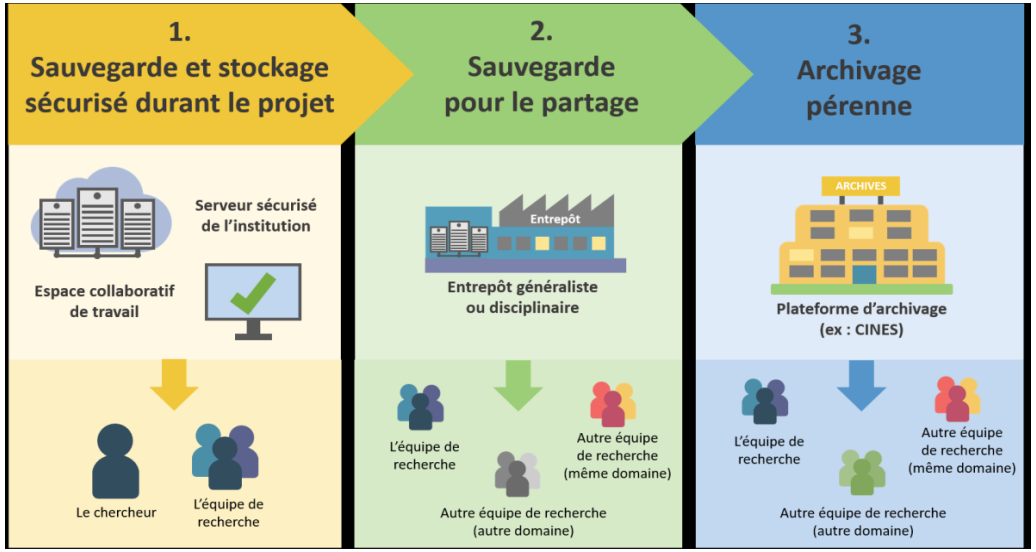
A Réfléchir bien en amont

Choix du support en fonction

- du type et volume des données
- de l'environnement de travail (vigilance pour les ordinateurs portables (pas de sauvegarde automatique), disques externes, clés USB)
- de la fréquence d'accès aux données

la sauvegarde et le stockage des données : des points essentiels

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



Dans l'idéal, dupliquer, stocker les données à différents endroits sur différents supports

Règle du 3-2-1:

- Garder 3 exemplaires de données
- sur 2 supports ou technologies différentes
- dont 1 se trouve hors site

Les possibilités de stockage

Possibilités de sauvegarde proposées par l'EIC sur le campus :

Sys. de fichiers	Taille	Utilisé	Dispo	Uti%	Monté sur	
//cubebe/bpmp	800G	569G	232G	72%	U	données chaudes
//stocka2/bpmp-unites	11T	11T	732G	94%	S	données froides
//backup1/backup-bpmp	16T	15T	2,0T	89%	Q	données personnelles

U : données sauvegardées grâce à des snapshots toutes les 4h (entre 8h et 20h du lun au ven).
restauration pouvant remonter à un peu plus de 4 semaines.
En plus, sauvegarde sur bande magnétique (rétention de 6 mois)

S : données non sauvegardées

Q : destination de rsync. Données sauvegardées tous les soirs par Henrique sur les NAS dans son bureau

Possibilités de sauvegarde proposées par nos tutelles :

OneDrive inrae : 30 Go
NextCloud inrae : 100 Go
MyCore cnrs : 100 Go
NextCloud supagro : 5 Go

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

Les outils collaboratifs

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

(quelque soit l'organisme d'appartenance)

- 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

Partage et diffusion

Une fois le projet terminé, partage des données avec la communauté : publication(s) & dépôt des données et des métadonnées du projet dans un entrepôt de données

EBI https://www.ebi.ac.uk/submission/

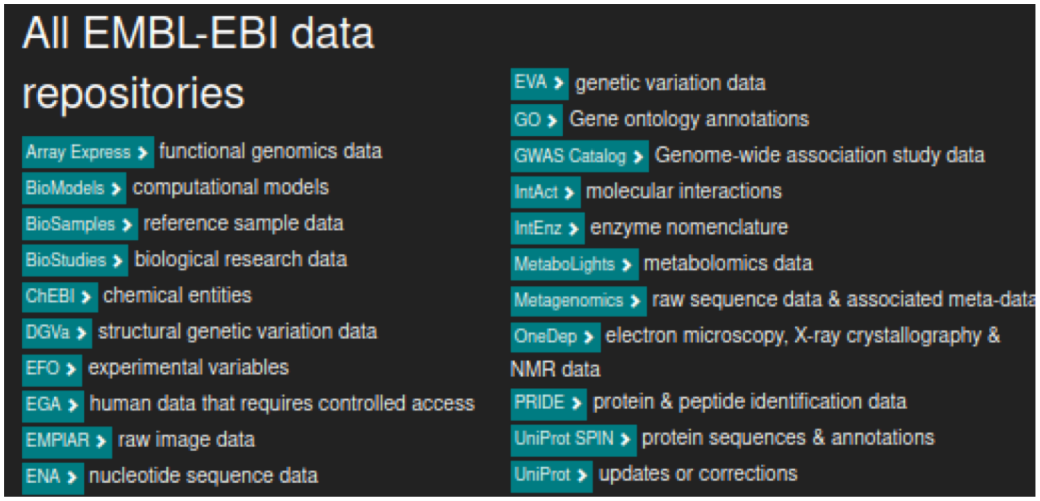
NCBI https://submit.ncbi.nlm.nih.gov/

data INRAe https://data.inrae.fr/

IDR (imagerie) https://idr.openmicroscopy.org/

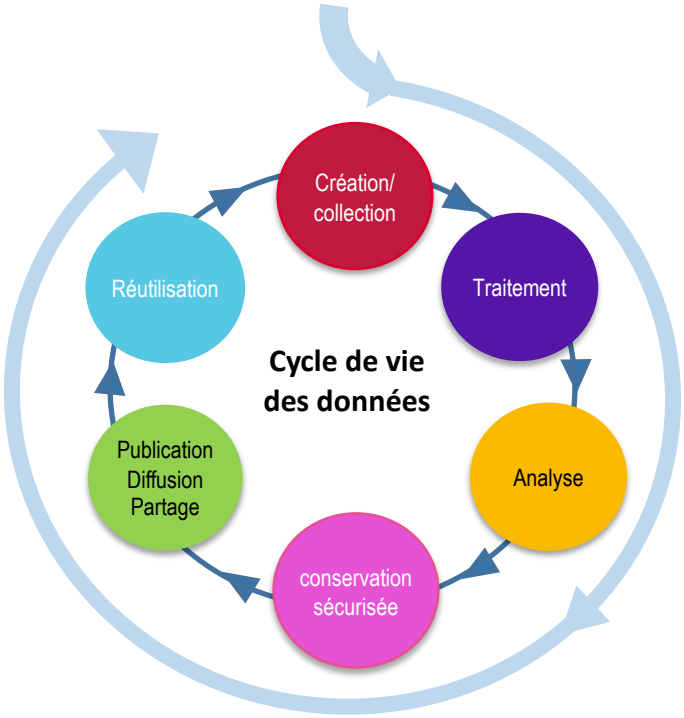
IFB /FBI (France Bio Imaging) collaborent pour produire un PGD de structure pour toutes les infras d'imagerie de FBI.

Ces dépôts ne sont pas éternels et peuvent disparaître du jour au lendemain



Les outils

Plan de gestion de données
PGD / DMP



PGD :
Document formalisé décrivant la gestion des données (Datasets) au cours du cycle de vie

Evolutif : plusieurs versions
feuille de route, en ajustement constant.
Décrit et rassemble les informations indispensables à l'interprétation et reproduction des données auparavant dispersées entre différents acteurs

Permet de se poser les bonnes questions pour prendre les bonnes décisions tout au long du cycle de vie des données

Se réfléchit et rédige à plusieurs

livrable du projet
obligatoire depuis décembre 2021 (Décret n° 2021-1572 du 3 décembre 2021 , Respect des exigences de l'intégrité scientifique par les établissements publics)

Les objectifs du PGD

- Assurer la reproductibilité des expériences
  - Décrire comment les données sont obtenues
- Respecter le droit et les personnes
  - Clarifier le cadre juridique et éthique
- Permettre la réutilisation des données
  - Garantir la compréhension des données
- Éviter les pertes de données
  - Assurer un stockage adapté
- Établir le rôle de chacun
  - Définir les responsabilités
- Clarifier les droits de réutilisation
  - Spécifier les modalités de partage

« aussi ouvert que possible ; aussi fermé que nécessaire » : les principes FAIR

- les principes FAIR sont un *ensemble de principes directeurs* visant à rendre les données **trouvables, accessibles, interopérables et réutilisables**
- ces principes fournissent des orientations pour la gestion des données scientifiques et sont pertinents pour toutes les parties prenantes de l'écosystème numérique

Comment rédiger un Plan de gestion de données ?
Outils pour Créer et rédiger un PGD/DMP

- 

Optimiser le Partage et l'Interopérabilité des Données de la Recherche

Inist/CNRS : Portail d'outils et de service d'aide à la mise en application des principes FAIR

➢ Trames de PGD/DMP : DMP-OPIDOR

- Questionnaire concernant tous les aspects du cycle de vie de la donnée de sa création jusqu'à la valorisation.
- Plusieurs trames disponibles (ANR, institutionnelles, …)

Rédaction en ligne (aide)
Nouvelle version DMP V3 disponible : plus structuré, import facilité d'informations

Classes virtuelles Plan gestion de données organisées régulièrement par l'INRAE

- Datastewardship Wizard DSW https://ds-wizard.org
- RDM kit https://rdmkit.elixir-europe.org/

Bienvenue !

DMP OPIDoR vous accompagne à travers l'élaboration et la mise en pratique de plans de gestion de données et de logiciels.

Accessible à la communauté scientifique de l'ESR et à ses partenaires français ou étrangers

Personnalisable par tout organisme de recherche pour la mise en place de sa politique de données

Enrichi par des exemples et des recommandations adaptés à l'environnement de recherche

Collaboratif : il facilite les échanges entre les partenaires d'un même projet et les services d'accompagnement

DMP OPIDoR évolue grâce à vos retours. Les développements s'inscrivent dans le cadre d'une collaboration internationale autour du logiciel open source DMPRoadmap

Rejoignez la communauté des utilisateurs de DMP OPIDoR

Créez un compte, connectez-vous et laissez-vous guider !

Découvrez DMP OPIDoR

Accès institutionnel

Connexion

Accès individuel

* **Courriel**

* **Mot de passe**

Mot de passe oublié ?

Se souvenir de moi

Connexion


DMP OPIDoR
Tableau de bord
Créer des plans
DMPs publics
Modèles de DMP
Aide
Plus
Français
Colette TOURNAIRE-ROUX
Tableau de bord
Dans le tableau ci-dessous figurent les plans que vous avez créés, ainsi que ceux que vous partagez avec d'autres. Vous pouvez à tout moment les modifier, les partager, les télécharger, en faire une copie ou les supprimer.
Titre du plan
Modèle
Modifié
Rôle
Propriétaire
Test
Visibilité
Partagé
DMP du projet "Colette's Plan"
INRAE - Trame Structure
17/11/2021
Propriétaire
Vous
N/D
Non
Actions
DMP du projet "Colette's Plan : ABAQUA - ABA-dependent contr ...
ANR - Modèle de PGD (français)
15/11/2021
Propriétaire
Vous
N/D
Oui
Actions
Modifier
Partager
Télécharger
Copier
Supprimer
DMP du projet "ABAQUA - ABA-dependent control of plant hydra ...
ANR - Modèle de PGD (français)
15/11/2021
Propriétaire
Vous
N/D
Oui
DMP du projet "Targeting Root Hydraulic Architecture to impr ...
ERC DMP
08/10/2019
Propriétaire
Vous
N/D
Oui
Créer un plan

Ce plan est basé sur le modèle "ANR - Modèle de PGD (français)" fourni par Agence nationale de la recherche (ANR). **(version: 2, publiée: 26 septembre 2019)**

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

tout développer | tout réduire

expression de gènes | Ressources génétique | transport d'eau | protéomique | reconstitution

☐ Les réponses de cette section sont communes à tous les produits de recherche

1a. Comment de nouvelles données seront-elles recueillies ou produites et/ou comment des données préexistantes seront-elles réutilisées ?

1b. Quelles données (types, formats et volumes par ex.) seront collectées ou produites ?

2. Documentation et qualité des données

tout développer | tout réduire

expression de gènes | Ressources génétique | transport d'eau | protéomique | reconstitution

☐ Les réponses de cette section sont communes à tous les produits de recherche

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

2b. Quelles mesures de contrôle de la qualité des données seront mises en œuvre ?

3. Stockage et sauvegarde pendant le processus de recherche

- On n’est pas obligé de remplir toutes les cases, du moins dans les premières versions, on peut (doit) modifier au fur et à mesure de l’évolution du projet .

➢ En général 3 versions : début, milieu et fin de projet

➢La première version :

Décrire la structuration des données prévues

- ➢Arborescence fichiers
- ➢Règles de nommage (à formaliser)
- ➢Collecte des métadonnées
- ➢Règles de sauvegarde

Chronologie pour la rédaction d'un DMP

1. **Identifier les différents jeux** de données (A discuter entre collaborateurs du projet)

2. **Définir une organisation** pour structurer les données

   a. Se mettre d'accord sur les règles de nommage /classement

   b. Déterminer les métadonnées techniques/ expérimentales à associer aux données et prévoir leur collecte et enregistrement associé aux données (fichier associé au données).

   c. Définir une procédure de sauvegarde sécurisée , si possible une copie sur 1 espace collaboratif pour faciliter le partage.

3. **Créer un compte sur DMP OpiDoR**

   . Choisir la trame de DMP (ANR, ..)

   . Répondre aux questions

   V1 : décrire l'organisation prévue

   V2 : actualiser

   V3 : Description des données finales , modalités d'ouverture et partage

Exemple de structuration des données

Objectif :

- Organiser les données :
- Séparer les données brutes et traitées,
- Centraliser les données/partager
- Les documenter
- Ne rien perdre

HyArchi files NAMING and ORGANIZATION

Abreviation to use : Lpr, scan, archi for root architecture , Syn for synthesis, RNA, Exp for experiment, etc…

A folder per Experimenter :

Folder SN : (SN : SurnameName experimenter initials)

Subfolder :YYYY-MM-SN-Exp01 (subfolder per experiment)

File MetaD-Exp01

Fill « MetaD-Exp00 » with general experimental conditions and if necessary experimental design for scans (sheet2) and Lpr (sheet3) to create a new file MetaD-Exp01.

SN-LPR-Exp01 (subfolder per technic)

Subfolder SN-LprDataExp01 (subfolder for raw data)

Save here .csv files from pressure chambers (ddmmyyyy-hhmm ChannelIX.csv)

Subfolder :SN-LprAnalyzedExp01 (subfolder for analysed data)

Files made from the use of « analyse chambre automatique.xls » go here

File YYYYMMDD-SN-Lpr-Exp01-nn (for example 2019-05-13-LR-LPR-Exp04-1.1)
File YYYYMMDD-SN-Lpr-Exp01-nn (for example 2019-05-13-LR-LPR-Exp04-1.2)
File YYYYMMDD-SN-Lprxx …

SN-RootArchi-Exp01

Subfolder :SN-RootScan

File YYYYMMDD-SN-Scan01
File YYYYMMDD-SN-Scan02
File YYYYMMDD-SN-Scanxx

Subfolder :SN-Archi-Analysis (analysis)

File YYYYMMDD-SN-WinR

Subfolder for synthesis files : SN-SYN-Exp01 (necessary only if several files)

File YYYYMMDD-SN-SYN-EXP01

Exemple de Trame pour collecte des métadonnées expérimentales projetHyArchi

MetaD-ExpXX

Experiment (SN-xx)					
Date					
Species	Maize				
Experimental plan file					
Genotype(s)					
SeedUSoon stock(s)					
Sowing date					
Germination lengh (d)					
preculture lengh (d)					
treatment (medium)					
Culture Chamber N°					
Hygrometry (%)					
Day length (h)					
Chamber temp (°C)					
Plant age on the first day of experiment					
Remarks (anomalies during culture, ...					
Measures/analyses	Lpr ☐	Root Scan ☐	RNA seq ☐	DNA ☐	.... ☐
Corresponding files					
Details in notebook N°					
page					

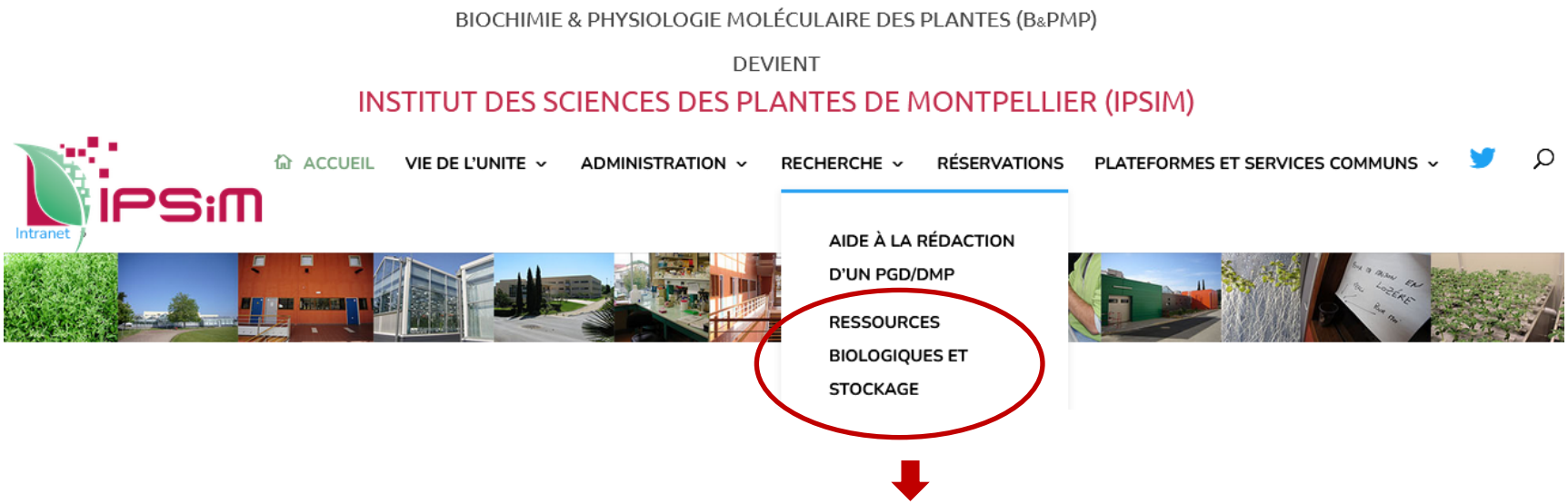
Pour vous aider…

Référent données opérationnels (RDO):

Information/Accompagnement à la gestion des données/ PGD / remonter les besoins

- Page sur l'intranet de l'unité

Pour vous aider…



Informations pour utiliser le logiciel de gestion des graines (SeedUSoon) :
Gérer les données concernant les ressources biologiques (données génétiques, données de génotypage, ID KO, …) et les partager.

Pour vous aider…

- Documentation

Doranum : Données de la Recherche, Apprentissage Numérique

➢ information et modules d'autoformation :

- Structuration/ Nommage
- Métadonnées
- Principes FAIR

Cahier de laboratoire électronique (recherche reproductible)

- Rapport GT cahier de labo électroniques (2021)

➢ Comparaison de différents outils : **recommandations sur les aspects d'interopérabilité entre différents outils** (accès aux données, sauvegardes, de l'archivage, etc.)

➢ **choix d'un outil** (interopérabilité, sécurité,….)

ELN open source: eLabFTW

➢ Séminaire 5 mai 2022 (visioconférence)

Conclusion

Bonne gestion de données :

- Pratiques qui favorisent le partage et la réutilisation :

Sauvegarde sécurisée
Description des données /méthodes d'obtention et d'analyse
Usage d'outils collaboratifs.

Le Plan de gestion de données est avant tout un outil d'aide à la gestion de données pour se poser les bonnes questions et anticiper
- Pour la gestion à l'échelle d'un projet (recherche, thèse, structure)
- pour le partage à terme.

Besoin d'échanges et de communication pour améliorer et faire évoluer un mode de gestion adapté à nos besoins

- volontaires pour échanger sur le sujet ?
  - Partager les informations, les expériences
  - Les besoins