

Bienvenue















Formation Science Ouverte et PGD

Édition 2 - Chez vous - du 15 au 18 novembre 2021



Informations pratiques

- Horaires: 9h00 -12h30 de lundi à jeudi, 1 pause vers 10h30
- Lien zoom à enregistrer
 - https://inrae-fr.zoom.us/j/91269258165?pwd=TUtuV25tM1BMcWpLZUdobHRCNEU3dz09
 - o ID de réunion : 912 6925 8165
 - Code secret : 123
- En distanciel
 - Couper vos micros pendant les sessions de cours
 - Utiliser le chat ou l'icône dédiée pour lever la main
- ...mais interactif! Nous avons planifié
 - Beaucoup de (courtes) séances pratiques avec des outils dédiés comme :
 - Scrumblr http://scrumblr.ca
 - Mentimeter https://www.menti.com
 - Des séquences d'échanges : n'hésitez pas à partager vos expériences sur le sujet



Objectifs pédagogiques

Après cette formation, vous connaîtrez

- les différents points fondamentaux (théoriques, pratiques, juridiques) en lien avec la politique nationale d'ouverture des données de la recherche
- les ressources nationales et internationales accessibles à la communauté scientifique ainsi que les solutions proposées par l'IFB pour gérer les données d'un projet de recherche
- les principales notions et ressources concernant les métadonnées en biologie
- les outils et principes permettant de rédiger un Plan de Gestion de Données (PGD)

Contenu de la formation : 1 module par ½ journée

 Module 1 : Les données de la Recherche et leur centralité dans le processus de recherche

Crise de reproductibilité - Vers FAIR - Cycle de vie des données - Le plan de gestion des données

- Module 2 : La vie des données pendant le projet : guide de bonnes pratiques
 Introduction Le nommage des fichiers Format de fichier Stockage et accès -Outils et solutions
- Module 3 : Les Métadonnées : les standards en sciences de la Vie et retour d'expérience sur la soumission dans des entrepôts internationaux

Introduction aux métadonnées - Choix des métadonnées en sciences de la Vie - Retour d'expérience sur la soumission dans les dépôts internationaux

- Module 4 : Partage et valorisation des données
 Cadre Juridique Les licences sur les données Quel entrepôt pour quelles données ?
- > supports ici : https://moodle.france-bioinformatique.fr/course/view.php?id=3

Brise-glace

Chacun dépose une photo de son choix sur le site https://osf.io : répertoire
 Fair DATA Novembre 2021

(...toutes nos excuses, vous n'avez le droit d'y écrire que depuis peu...)

- Chacun se présente rapidement :
 - Nom, prénom, affiliation professionnelle
 - Vos activités professionnelles, votre lien aux données de la Recherche
 - Vos attentes vis à vis de la formation
 - Projection de votre photo "emblématique" (lieu ou objet de votre choix) et séance de questions/réponses pour deviner de quoi il s'agit

Et merci!













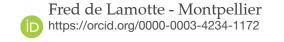




https://ifb-elixirfr.github.io/IFB-FAIR-data-training
Et bientôt : https://moodle.france-bioinformatique.fr/



Module 1 Séquence 1













Brise glace

Mieux se connaître et vérifiez que :

- Votre micro est fonctionnel
- Votre caméra est fonctionnelle
- Vous avez repéré l'espace de partage
- Vous êtes en mesure de partager votre écran





Votre photo emblématique

- ☐ A-t-elle été bien téléversée sur le site de partage de la formation ?
- ☐ A-t-elle été nommée d'une manière reconnaissable?
- Est-elle libre de droits ?





Introduction

Chaque participant se présente :

- Prénom
- Lieu d'activité
- Bref exposé des rôles & responsabilités
- Qu'attendez vous de cette formation ?
- Présentation de votre photo téléversée :
 - 3 questions autorisées pour déterminer le lieu où cette photo a été prise
 - La langue sera donnée au chat si pas de succès





Module 1 Séquence 2



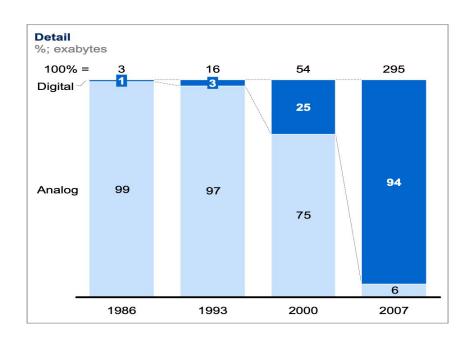


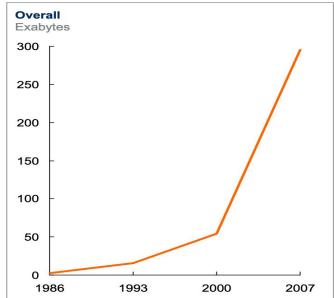
Crise de reproductibilité

30 minutes



La disruption numérique : une bascule brutale





Qui bouscule l'existant

- La première compagnie de taxi n'en possède aucun (Uber)
- Le premier fournisseur de logement n'en possède pas (AirBnB)
- La première compagnie de téléphonie ne possède pas de standard (Skype)
- Le premier fournisseur d'info ne crée pas de contenu (Facebook)
- Le premier diffuseur de film ne possède pas de salle de cinéma (Netflix)

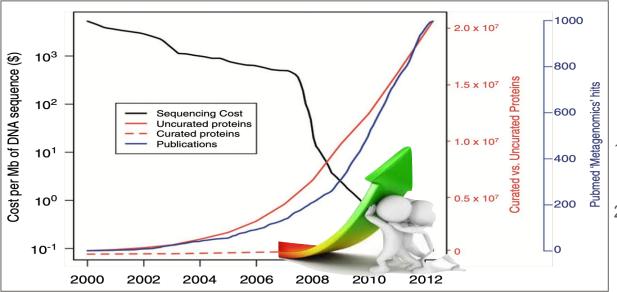
Waves of Digital Disruption



Le déluge des données en Science

Les techniques à haut débit, une révolution qui provoque un déluge de données Génome humain :

en 1990 = **13 ans** et **3 Milliards** \$ en 2015 = **quelques heures** et **1000** \$



- La quantité de données à stocker et analyser explose
- 2. Le **rendement** d'analyse chute

Répondre aux questions de la Science

PIVANT

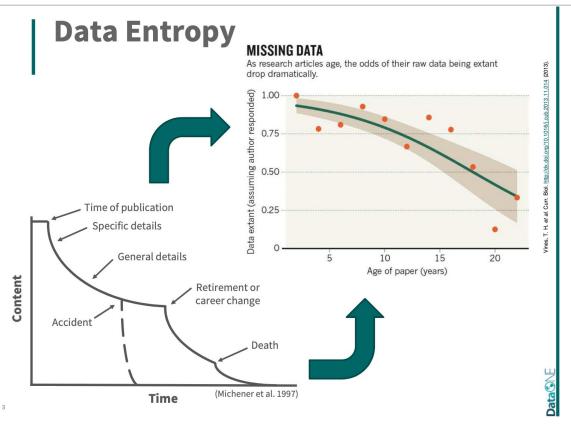
- 1 Concevoir l'expérimentation
- 2 Collecter des résultats
- 3 Analyser des résultats

Un changement de paradigme

MAINTENANT

- 1 Générer massivement des données
- 2 Organiser (stocker, documenter, annoter)
- 3 Analyser (extraire de l'information)
- 4 Diffuser l'information

Les ravages du temps



Les défis de la reproductibilité

RESEARCH ARTICLE

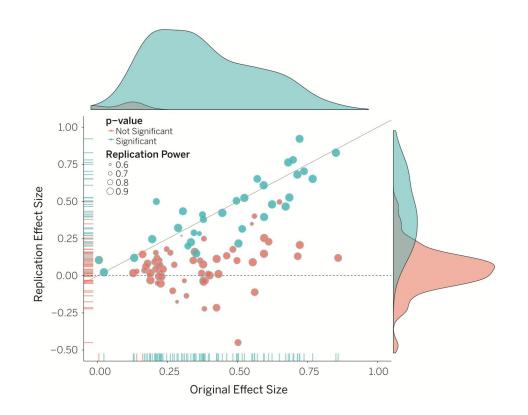
Estimating the reproducibility of psychological science

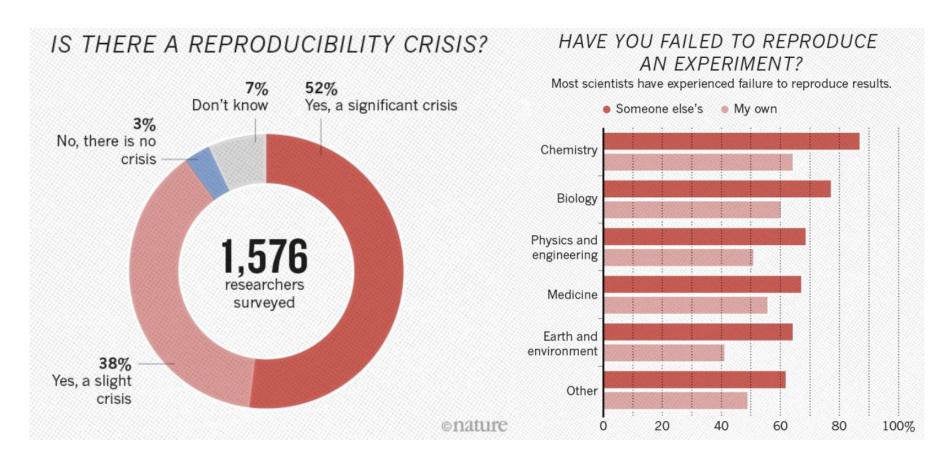
Open Science Collaboration*,†
+ See all authors and affiliations

Science 28 Aug 2015: Vol. 349, Issue 6251, aac4716 DOI: 10.1126/science.aac4716

The *Reproducibility project* set out to replicate 100 experiments published in high-impact psychology journals.

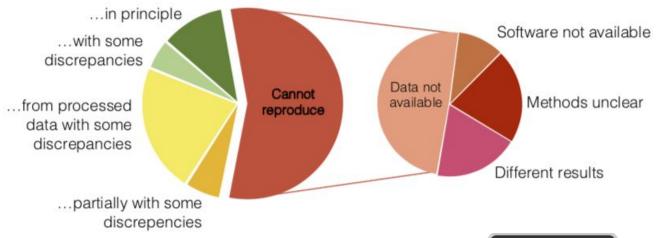
About one-half to two-thirds of the original findings could not be observed in the replication study.





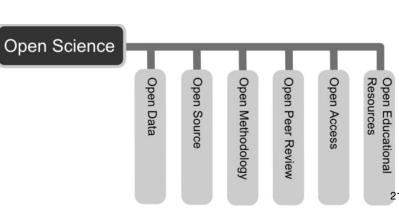
Replication of data analyses in 18 articles on microarray-based gene expression profiling published in Nature Genetics in 2005–2006:

Can reproduce...



Summary of the efforts to replicate the published analyses.

Adopted from: loannidis et al. Repeatability of published microarray gene expression analyses Nature Genetics 41 (2009) doi:10.1038/ng.295



Disruption + BigData + Crise

Le traitement de l'information (scientifique) sera notre Noeud Gordien



Exercise 1.1!

Quelle définition pour les données de la recherche ?

Pour tenter d'aborder cette question, nous allons procéder en 4 étapes :

- 1. Vous allez prendre un temps de réflexion **individuelle** de **3 minutes** pour rédiger une première définition qui vous est personnelle.
- 2. Pendant les **5 minutes** suivantes, vous discuterez de vos définitions avec 4 autres participants et proposerez une définition combinée, en gardant la trace des divergences s'il y a lieu. Donc vous finissez ces 5 minutes avec **1 définition**-(https://scrumblr.ethibox.fr/) (... JE VAIS VOUS MONTRER ...)
- 3. Votre groupe prendra connaissance de l'ensemble des définitions consolidées proposées puis une discussion portera sur les divergences entre ces définitions (temps prévu **10 minutes**)
- 4. L'activité se terminera par la présentation des définitions les plus courantes des données de la Recherche





Définition OCDE

Les données de recherche sont les **preuves** qui sous-tendent la réponse à la question de recherche et peuvent être utilisées pour **valider** les **résultats**, quelle que soit leur forme (i.e. imprimée, numérique ou physique).

Il peut s'agir de **renseignements quantitatifs** ou d'**énoncés qualitatifs** recueillis par les chercheurs dans le cadre de leurs travaux par **expérimentation**, **observation**, **modélisation**, **entrevue** ou autres méthodes, ou de renseignements tirés de preuves existantes.

Les données peuvent être **brutes** ou **primaires** (par exemple, directement issues de mesures ou de collectes) ou **dérivées** de données primaires par analyse ou interprétation (e.g. nettoyées ou extraites d'un ensemble de données plus vaste), ou encore dérivées de sources existantes dont <u>les droits peuvent</u>



Module 1 Séquence 3



Vers FAIR 30 minutes



Pendant la vidéo suivante

Notez les points marquants (bon ou mauvais) en gestion des données

Réutiliser les données? Où est le problème?



Exercice 1.2!

Quelles conditions pour que les données soient réutilisables ?

Pour tenter d'aborder cette question, nous allons procéder en 2 étapes :

1 - Réfléchissez à cinq conditions nécessaires et notez les (Mentimeter : lien dans le chat - Attention : notez un seul mot à la fois et en français, non composé et sans majuscule).

https://www.menti.com/upo8aj5tcw

2 - A partir du nuage de mots créé collectivement quels regroupements pouvons-nous faire ?



Un autre exemple...

Les principes FAIR



Les principes FAIR Data sont un *ensemble de principes directeurs* visant à rendre les données <u>trouvables</u>, <u>accessibles</u>, <u>interopérables et réutilisables</u>.

Ces principes fournissent des orientations pour la gestion des données scientifiques et sont pertinents pour toutes les parties prenantes de l'écosystème numérique.

Ils s'adressent directement aux producteurs et aux éditeurs de données afin de promouvoir une utilisation maximale des données de recherche.



Vos données sont elles FAIR?

Findable -- Faciliter la découverte des données

- Les données ont un **PID** (Persistent IDentifier ou identifiant pérenne en français)
- Les données sont décrites par des **métadonnées**
- Ces métadonnées doivent être liées aux PIDs des données
- Les données sont déposées dans un entrepôt de données





Vos données sont elles FAIR?

Accessible -- Permettre l'accès aux données et leur téléchargement

- Les données sont accessibles à travers un protocole de communication standard
- Ce protocole est **libre et ouvert**
- Ce protocole permet un accès par **authentification** si besoin
- Les **métadonnées restent accessibles** même si les données ne le sont plus



Vos données sont elles FAIR?

Interoperable -- Permettre l'exploitation des données quel que soit l'environnement informatique utilisé

- Les données sont décrites avec un vocabulaire contrôlé
- Le vocabulaire utilisé respecte les principes FAIR
- Les métadonnées sont reliées à d'autres données





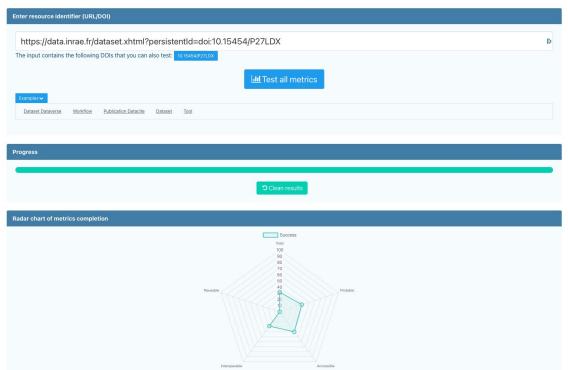
Vos données sont elles FAIR?

Reusable -- Permettre la réutilisation des données pour de futures recherches

- Les métadonnées ont une pluralité d'attributs
- Une licence de réutilisation est attribuée aux données
- La description des données indique leur provenance
- Le partage des données suit les **standards de la communauté scientifique**

FAIR Checker

How FAIR is my resource?



https://fair-checker.france-bioinformatique.fr/base_metrics



Session 1 Séquence 4

Fred de Lamotte - Montpellier https://orcid.org/0000-0003-4234-1172 Jean-François Dufayard https://orcid.org/0000-0002-7427-6822





Cycle de vie des données

15 minutes





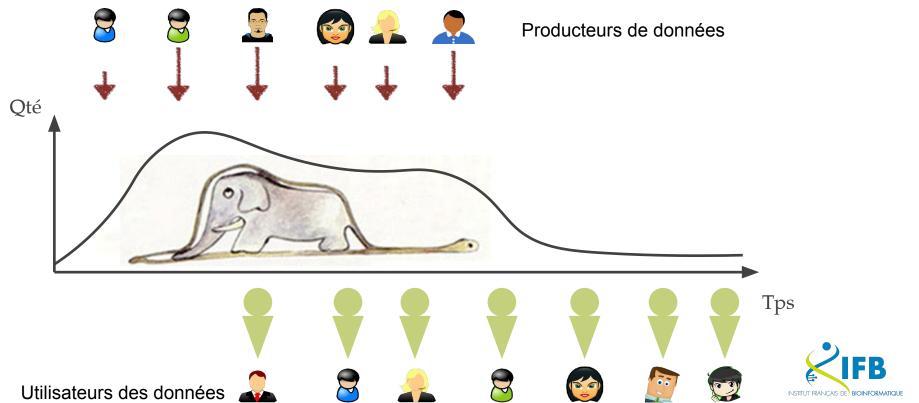
La vie des données

- Plusieurs temporalités
 - Le temps d'une thèse
 - Le temps d'un projet de recherche
 - Le temps de vie de la thématique dans le labo
 - Le temps de vie de la thématique dans l'institution
 - Le temps de vie de la thématique ...





Un projet sur la durée



Exercice 1.3!

En sous-groupe

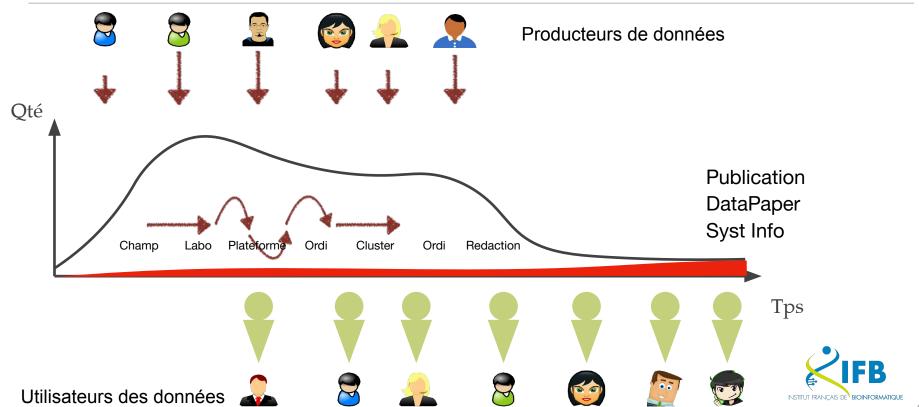
Connectez vous sur le scrumblr (https://scrumblr.ethibox.fr/pointsProjets1121)

Rédigez et positionnez des post-it concernant tous les points d'attention à avoir le long d'un projet, de sa conception jusqu'à sa valorisation

Fred(void



Un projet sur la durée





Les étapes

Le <u>modèle</u> de UK Data Archive définit les six étapes suivantes :

- Création ou collecte des données (creating data) ;
- Traitement des données (processing data);
- Analyse des données (analysing data);
- Conservation des données (preserving data);
- Accès aux données (giving access to data / data discovery);
- Réutilisation des données (reusing data).





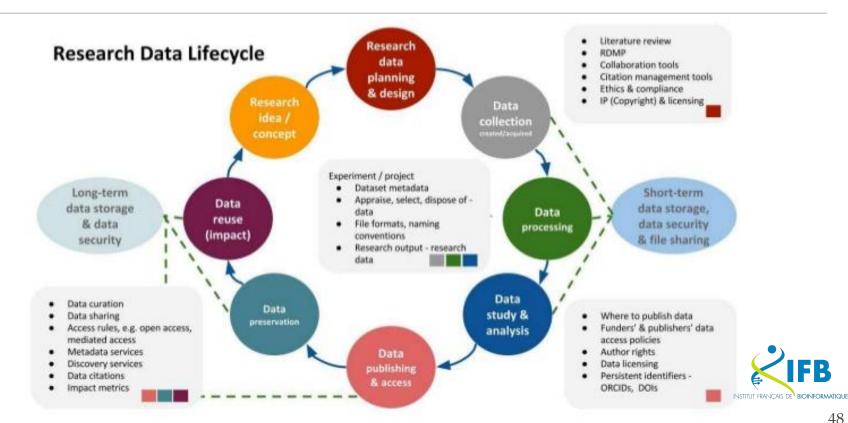
Les étapes



Une introduction à la gestion et au partage des données de la recherche - Le cycle de vie des données



Une vue plus détaillée





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





Session 1 Séquence 5

Fred de Lamotte - Montpellier
https://orcid.org/0000-0003-4234-1172
Jean-François Dufayard
https://orcid.org/0000-0002-7427-6822
Paulette Lieby
https://orcid.org/0000-0002-9289-9652





Donc, un PGD

Un plan de gestion de données

Ou un DMP: data management plan





Le plan de gestion des données

25 minutes



Pour quoi faire?

- Plan : on planifie (donc on anticipe)
- **Gestion** : on gère, on fait fructifier (on commence déjà par ne plus perdre)
- **Données** : Data is the new oil, the new soil



Gérer les données : comment ?

- Une approche pragmatique
 - Simple à comprendre
 - Simple à mettre en place
 - Simple à évaluer
 - Simple à faire évoluer

- Donc le Plan de Gestion des Données
 - Actuellement... un document texte (en attendant mieux)



- 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





- Assurer la reproductibilité des expériences
 - Décrire les données

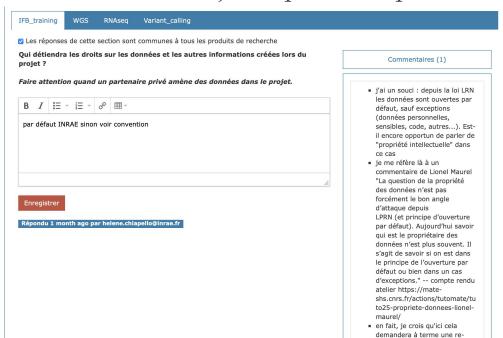
Modifier Supprimer

https://dmp.opidor.fr/

IFB training WGS RNAsea Variant calling Quels méthodes et outils sont utilisés pour acquérir et traiter les données ? Précisez Commentaires (1) les différents formats dans lesquels les données seront disponibles aux différentes phases de la recherche Vérifer avec Gautier si le mode opératoire de l'Illumina est stabilisé - Format : fastq, outil : séquenceur Illumina HiSeq 3500 frederic.de-lamotte@inrae.fr 1 hour ago - Format : fasta, outil : assembleur SOAP de novo - Format : gff3, outil : Eugene Ajouter un commentaire à partager avec les collaborateurs Enregistrer Répondu 1 month ago par helene.chiapello@inrae.fr



- Respecter le droit et les personnes
 - Clarifier le cadre juridique et éthique



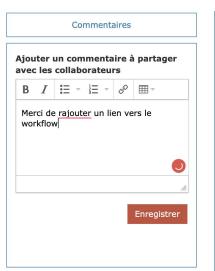




Permettre la réutilisation des données

Brève présentation des données générées, collectées ou réutilisées :

Garantir la compréhension des données







- Éviter les pertes de données
 - Assurer un stockage adapté







- Établir le rôle de chacun
 - Definir les responsabilités

Gérer les collaborateurs

Inviter des personnes à lire, modifier ou administrer votre plan. Les invités recevront une notification par courriel indiquant qu'ils ont accès à ce plan.

Adresse courriel	Permissions	
thomas.denecker@france-bioinformatique.fr	Éditeur Supprimer	
seilerj@igbmc.fr	Éditeur 6 Supprimer	
paulette.lieby@france-bioinformatique.fr	Éditeur Supprimer	
gautier.sarah@inrae.fr	Copropriétaire Supprimer	
frederic.de-lamotte@inrae.fr	Copropriétaire	
helene.chiapello@inrae.fr	Propriétaire	

Inviter des collaborateurs

* Courriel	ei	
* Permissions Co-propriétaire: peut modifier les détails du projet, changer la visibilité et ajouter des collat Editeur: peut commenter et effectuer des changements Lecture seule: peut voir et commenter, mais ne peut pas faire de modifications	oorateurs.	





- Clarifier les droits de réutilisation
 - Spécifier les modalités de partage

Sur quelle plateforme d'archivage pérenne seront archivées les données à conserver sur le long terme ? Sinon, quelles procédures seront mises en place pour la conservation à long terme ?

Il s'agit ici de plateformes d'archivage pérennes destinées à pérenniser les données, comme le C.I.N.E.S. Les entrepôts de données ne possèdent, à quelques exceptions près, pas cette possibilité.



Enregistrer

Un PGD - plusieurs versions

- C'est un document évolutif
 - au moins 3 versions :
- V1 à 6 mois
 - C'est un livrable pour H2020 et ANR
- V2 à mi-parcours
 - Quand la connaissance des données progresse
- V3 en fin de projet
 - Quand le flux de données se calme et qu'on a une bonne vision globale
 - Le confinement est un moment idéal pour ce genre de ménage!



Modèles pour les PGD

Modèle de PGD = une liste de sections, questions, et informations à remplir pour rédiger un PGD

Modèles existants:

- RDA Research Data Alliance un "modèle originel"
- Des modèles spécifiques par tutelles (INRAe, CEA, CIRAD, universités, grandes écoles...)
- Des modèles par grands guichets (HORIZON 2020 UE, ANR…)
- Des modèles pour certains centres de calcul/stockage (IN2P3)

Peu / pas de modèles spécifiques à des types de données



Outils pour les PGD

Systèmes pour rédiger des DMPs

• DMP OPIDOR - solution nationale





DSW - Data Stewardship Wizard - solution européenne (ELIXIR)









Et si nous sortions de la préhistoire?

Dichotomie

Structure

Plateforme Infrastructure Service...

PGD à durée indéterminé PGD plus générique



Projet

Projet d'imagerie Projet multi-omique Projet plus complexe ...

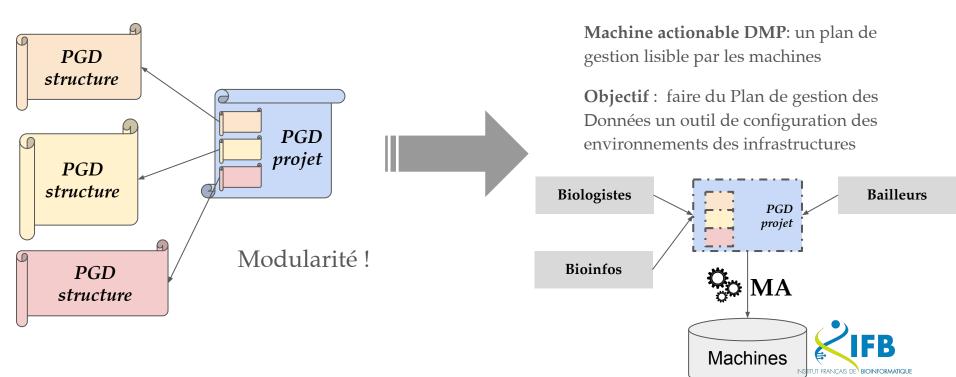
PGD à durée déterminé PGD plus spécifique





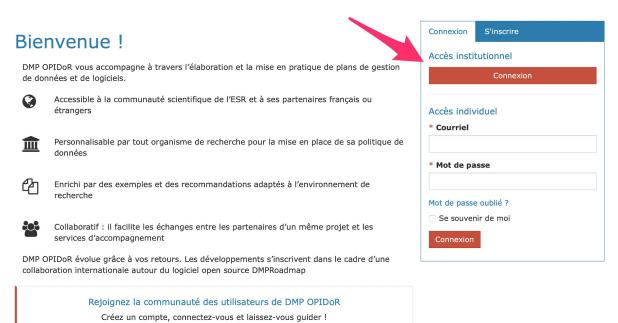


Et si nous sortions de la préhistoire?



Pour la suite

Aller sur https://dmp.opidor.fr





projet de test, d'entrainement ou à des fins de formation

Choisissez un modèle

Vous pouvez choisir soit un modèle fourni par votre organisme soit par un autre organisme, ou un modèle financeur. Le modèle par défaut est **Science Europe - DMP template (english).**

Retrouvez la liste des modèles disponibles







- Gestion et partage des données scientifiques
 - S Contact (Pôle Données de la recherche de l'IST)

Portail Data INRAE

IFB_Training_plant



Adresse courriel

lionel.maurel@cnrs.fr

Éditeur

Supprimer

nviter des collaborateurs	
* Courriel	
* Permissions	
Co-propriétaire: peut modifier les détails du projet, changer la visibilité et ajouter des collabora Editeur: peut commenter et effectuer des changements	ateurs.
Lecture seule: peut voir et commenter, mais ne peut pas faire de modifications	
Sauvegarder	