



# Module 1 Séquence 1

Fred de Lamotte - Montpellier  
<https://orcid.org/0000-0003-4234-1172>





# Bienvenue



**INRAE**

Fred de Lamotte - Montpellier  
<https://orcid.org/0000-0003-4234-1172>



# Et merci !



<https://ifb-elixirfr.github.io/IFB-FAIR-data-training>

# 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

Fred de Lamotte - Montpellier  
<https://orcid.org/0000-0003-4234-1172>





# Crise de reproductibilité

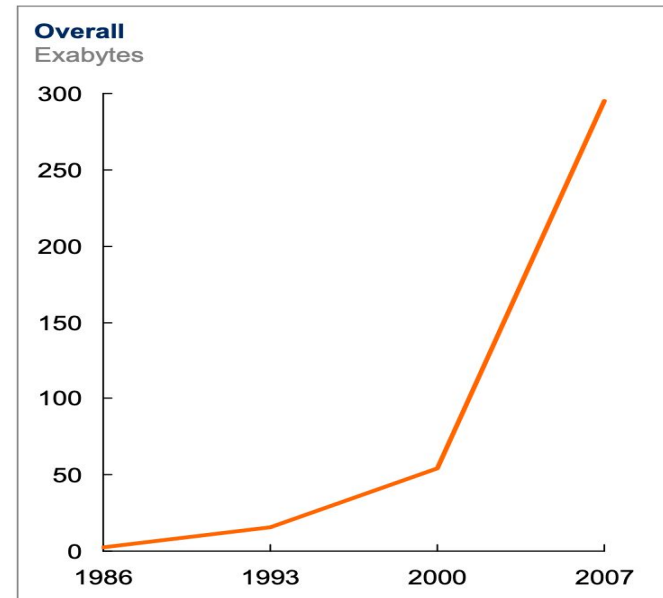
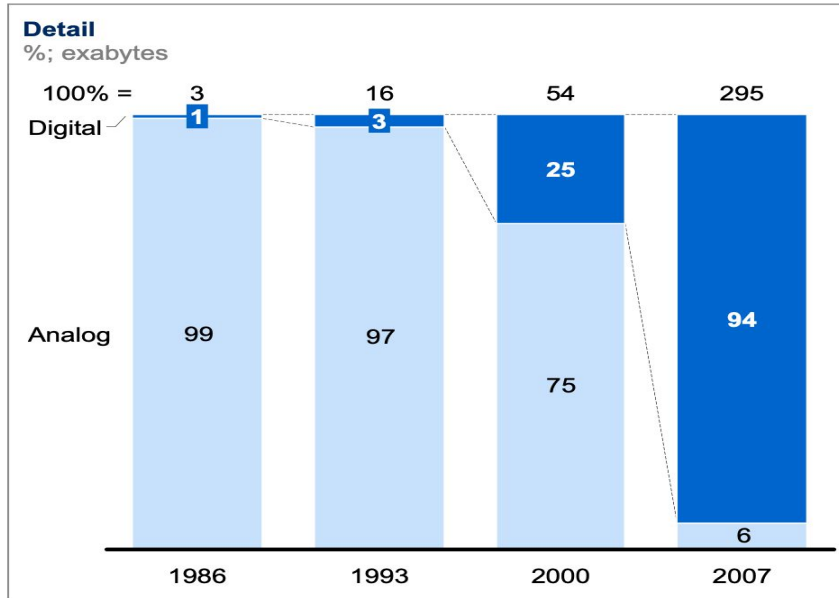
30 minutes

Fred de Lamotte - Montpellier  
<https://orcid.org/0000-0003-4234-1172>





# 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

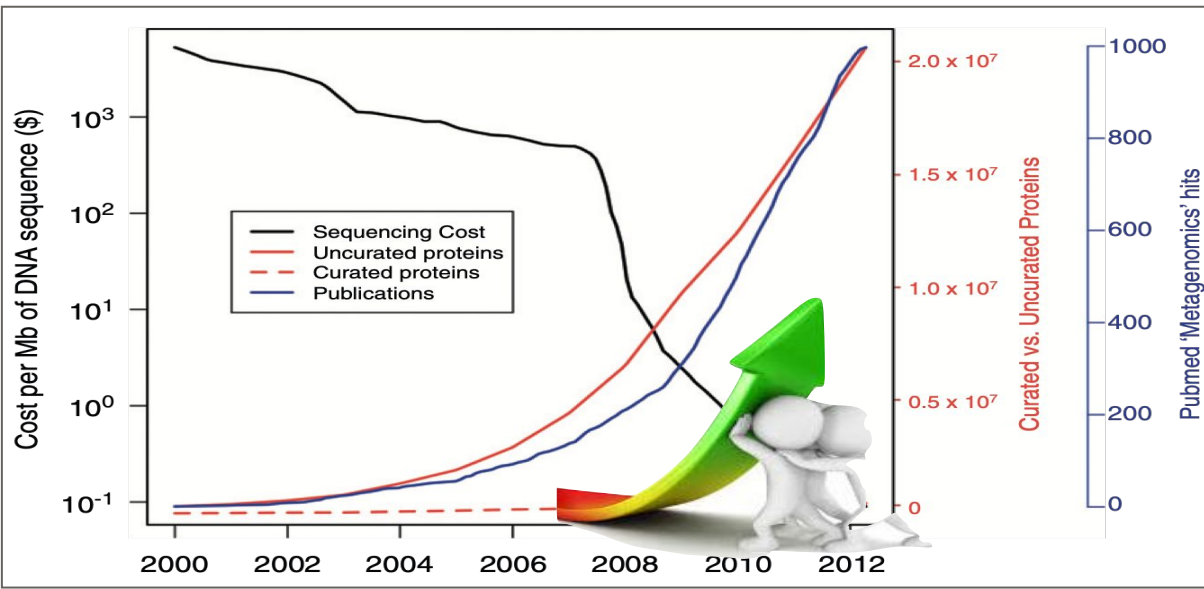


# Le déluge des données en Science

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 \$



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

---

# Répondre aux questions de la Science

---

**AVANT**

- 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

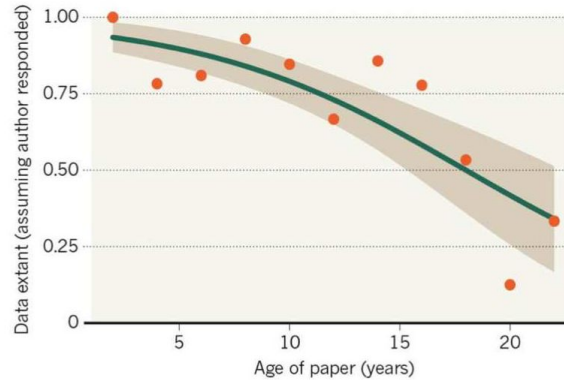
# Les ravages du temps

## Data Entropy



### MISSING DATA

As research articles age, the odds of their raw data being extant drop dramatically.



Vines, T. H. et al. Curr. Biol. <http://dx.doi.org/10.1016/j.cub.2013.11.014> (2013).

# Les défis de la reproductibilité

## RESEARCH ARTICLE

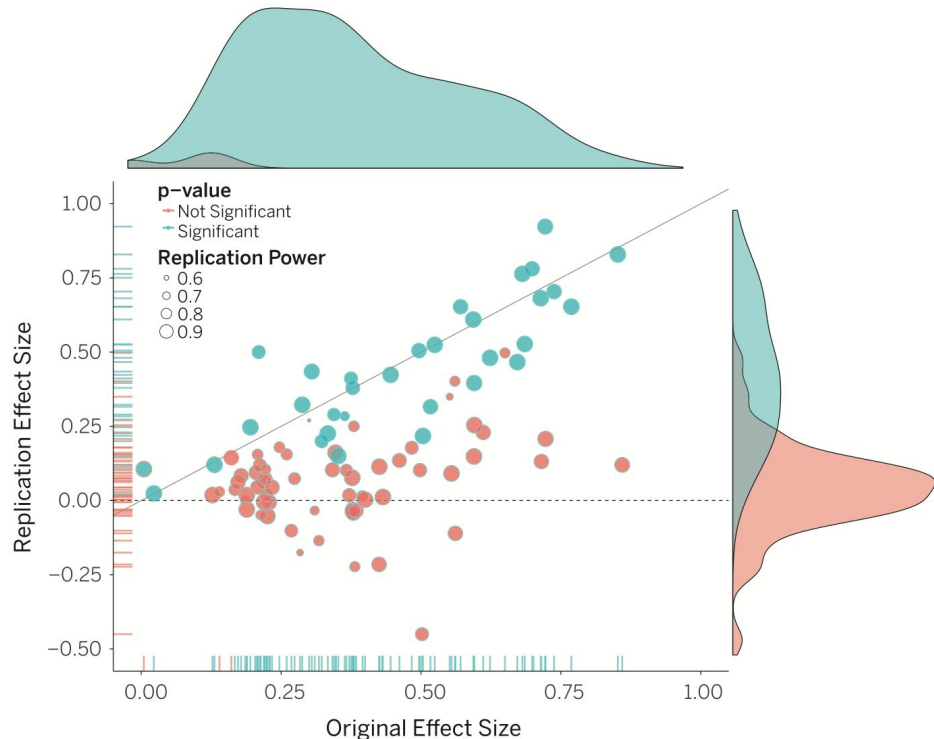
### Estimating the reproducibility of psychological science

Open Science Collaboration<sup>\*,†</sup>  
\* See all authors and affiliations

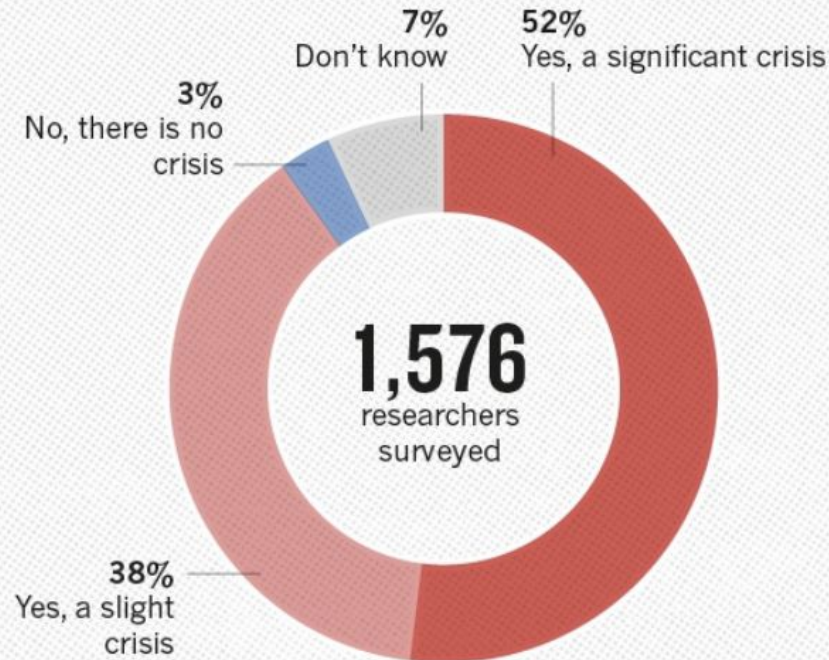
Science 28 Aug 2015;  
Vol. 349, Issue 6251, aac4716  
DOI: 10.1126/science.aac4716

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.



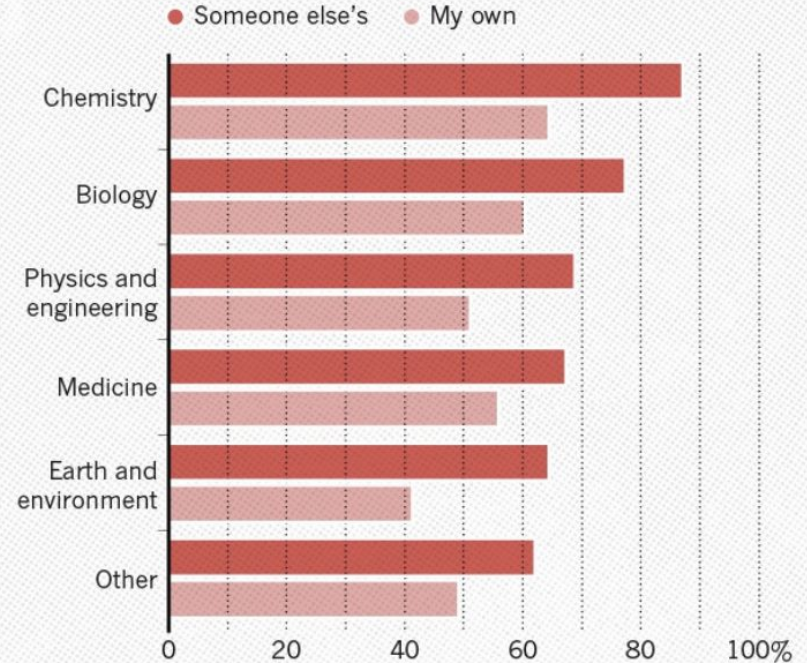
## IS THERE A REPRODUCIBILITY CRISIS?



©nature

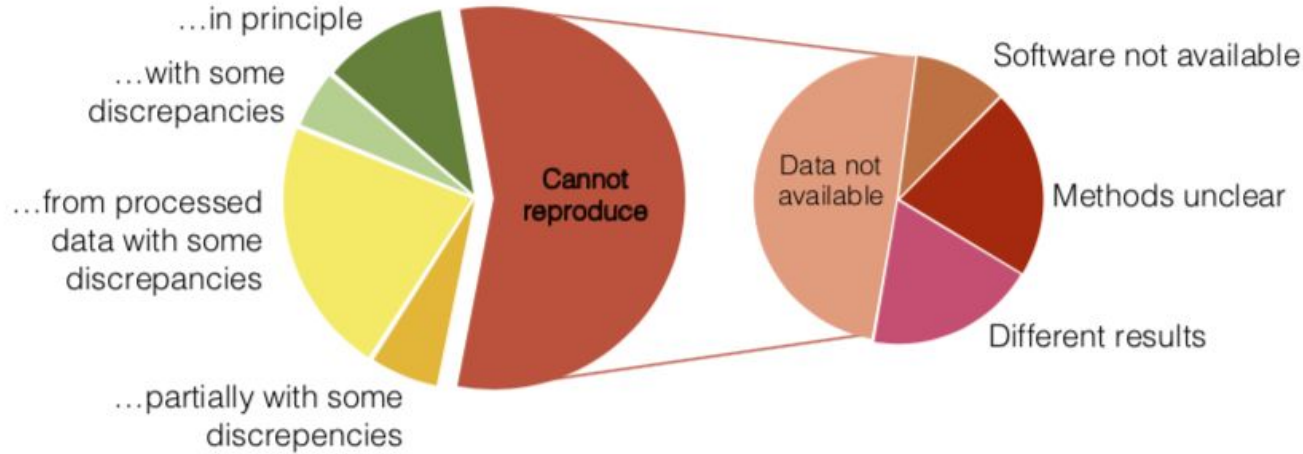
## HAVE YOU FAILED TO REPRODUCE AN EXPERIMENT?

Most scientists have experienced failure to reproduce results.



# 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: Ioannidis et al. Repeatability of published microarray gene expression analyses  
*Nature Genetics* **41** (2009) doi:10.1038/ng.295

## Open Science

Open Data

Open Source

Open Methodology

Open Peer Review

Open Access

Open Educational Resources



---

# Disruption + BigData + Crise

---

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



---

# Exercice 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 **5 minutes** pour rédiger sur le document partagé une première définition qui vous est personnelle.  
(<https://scrumblr.ethibox.fr/>)
2. Pendant les **5 minutes** suivantes, vous discuterez de vos définitions avec trois 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**
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 les droits peuvent être détenus par d'autres.



# Module 1 Séquence 3

Fred de Lamotte - Montpellier

<https://orcid.org/0000-0003-4234-1172>

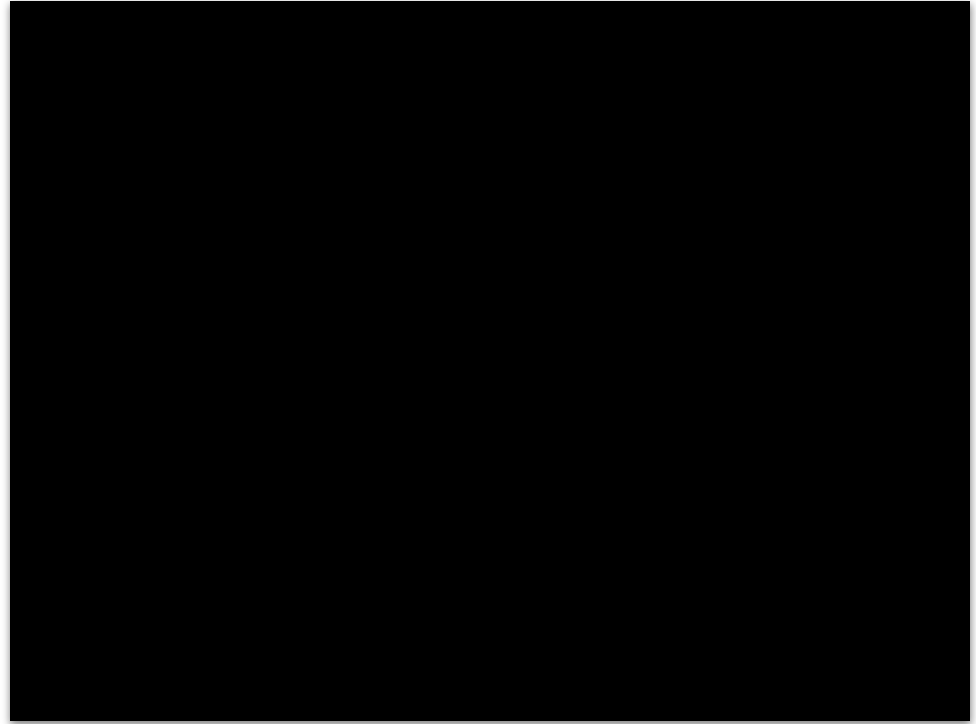
Paulette Lieby - IFB-core

<https://orcid.org/0000-0002-9289-9652>



# 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 !

---

**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/bo2sardvkz>

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 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.

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 auto évaluation

## FAIR self-assessment tool

Welcome to the ARDC FAIR Data self-assessment tool. Using this tool you will be able to assess the 'FAIRness' of a dataset and determine how to enhance its FAIRness (where applicable).

This self-assessment tool has been designed predominantly for data librarians and IT staff, but could be used by software engineers developing FAIR Data tools and services, and researchers provided they have assistance from research support staff.

You will be asked questions related to the principles underpinning Findable, Accessible, Interoperable and Reusable. Once you have answered all the questions in each section you will be given a 'green bar' indicator based on your answers in that section, and when all sections are completed, an overall 'FAIRness' indicator is provided.

Please be aware that additional explanatory information is provided within the tool. The (i) information button provides an overview of each of the FAIR high-level elements (Findable, Accessible, Interoperable and Reusable). Additionally, each question is hyperlinked, leading users to explanatory information and links to wider resources on related topics.

### Findable (i)

Does the dataset have any identifiers assigned?

Is the dataset identifier included in all metadata records/files describing the data?

How is the data described with metadata?

What type of repository or registry is the metadata record in?

---

### Accessible (i)

### Interoperable (i)

### Reusable (i)

Total across F.A.I.R

To learn more about making your data more FAIR visit: [www.ands.org.au/fair](http://www.ands.org.au/fair)

**Note:** Click on the linked heading text to expand or collapse accordion panels.

**Tool Disclaimer:** The ARDC FAIR data Self-Assessment Tool has been developed by ARDC. It is provided purely for educational and informational purposes. It is based on our interpretation of the FAIR Data Principles with the acknowledgement that there are other interpretations of the principles. Please also see other tools like the CSIRO 5 star data rating tool and the DANS FAIRdat tool which provided valuable inspiration in developing this tool.

The scores arising from this tool are intended for self assessment purposes only and to trigger thinking and discussion around possible ways of making data more FAIR.





# FAIR : évaluation automatique

FAIR CHECKS

Base Metrics

Usage statistics

BETA

Custom Metrics

## How FAIR is my resource

### Demo context and goal

### Monitoring progress in FAIRification through self-assessment of resources maturity indicators

This demo is based on the FAIRMetrics framework [Wilkinson, Dumontier et al., Scientific Data 6:174] [GitHub] that is composed of Maturity Indicators (MI), compliance tests and the evaluator application itself. For now, few efforts have been done so far to take advantage from their concrete implementation, in the process of improving FAIRness of users/community resources. Furthermore, existing interfaces do not provide concrete help or guidelines to developers for better sharing their published works. In this work we propose a web demonstrator, leveraging existing web APIs, aimed at i) evaluating FAIR maturity indicators and ii) providing hints to progress in the FAIRification process.

### Enter or select a resource identifier example

OR pick an [Example](#)

### Start basic or advanced test

[Test all metrics](#)



# Session 1 Séquence 4

Fred de Lamotte - Montpellier  
<https://orcid.org/0000-0003-4234-1172>





# Cycle de vie des données

15 minutes

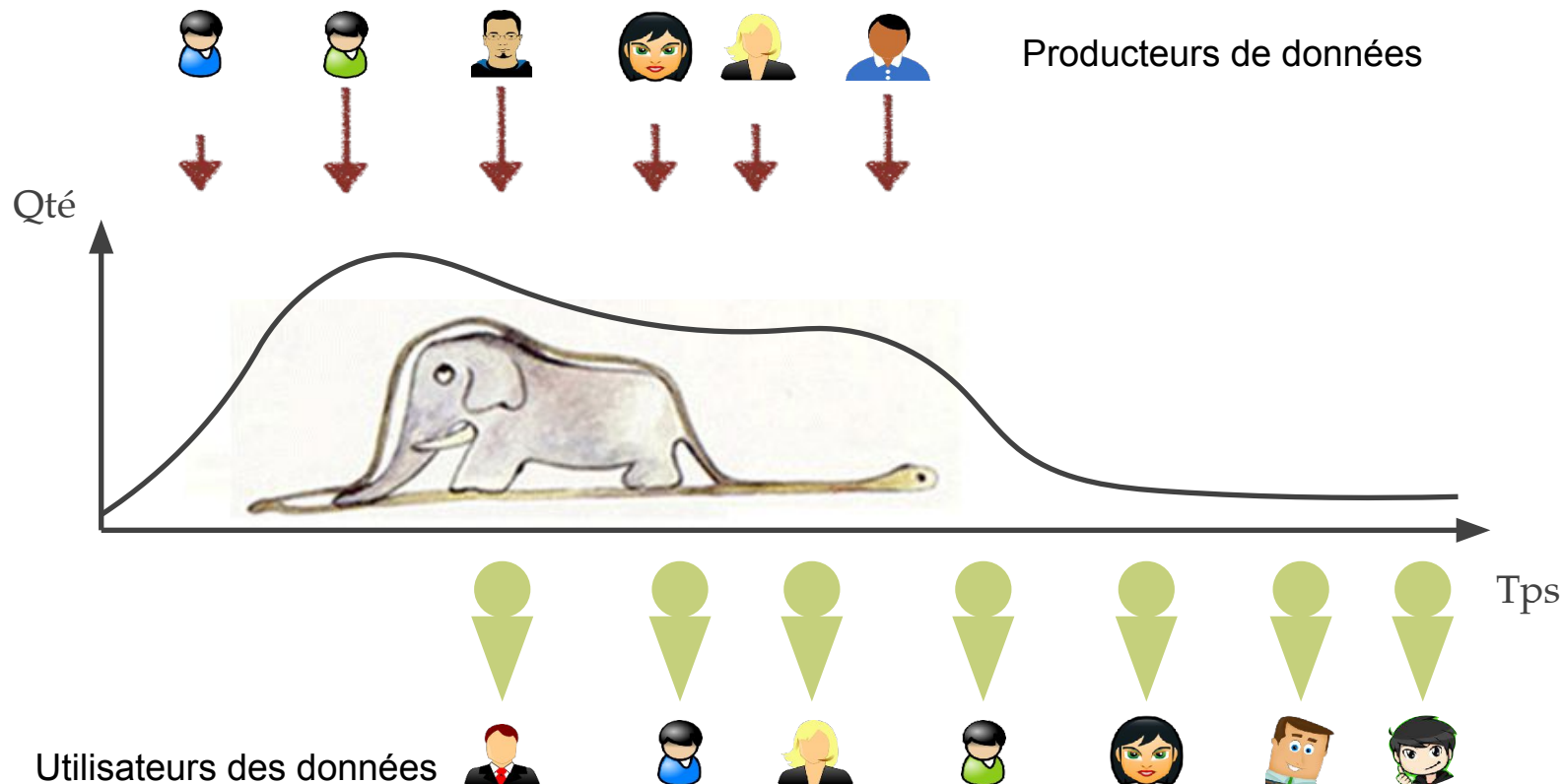
Fred de Lamotte - Montpellier  
<https://orcid.org/0000-0003-4234-1172>



# 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

En sous-groupe

Connectez vous sur le scrumblr ([https://scrumblr.ethibox.fr/pgdonline2\\_laviedesdonnees\\_exo](https://scrumblr.ethibox.fr/pgdonline2_laviedesdonnees_exo))

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

Petite démo de prise en main

### DEBUT

Budget

Ressorts

Volumétrie

Transfert

identification

Format

Classement

Intégrité

### MILIEU

Analyse Nettoyage,  
vérification, sélection

stockage

Partage

Versionning

### FIN

Ethique

Sécurité

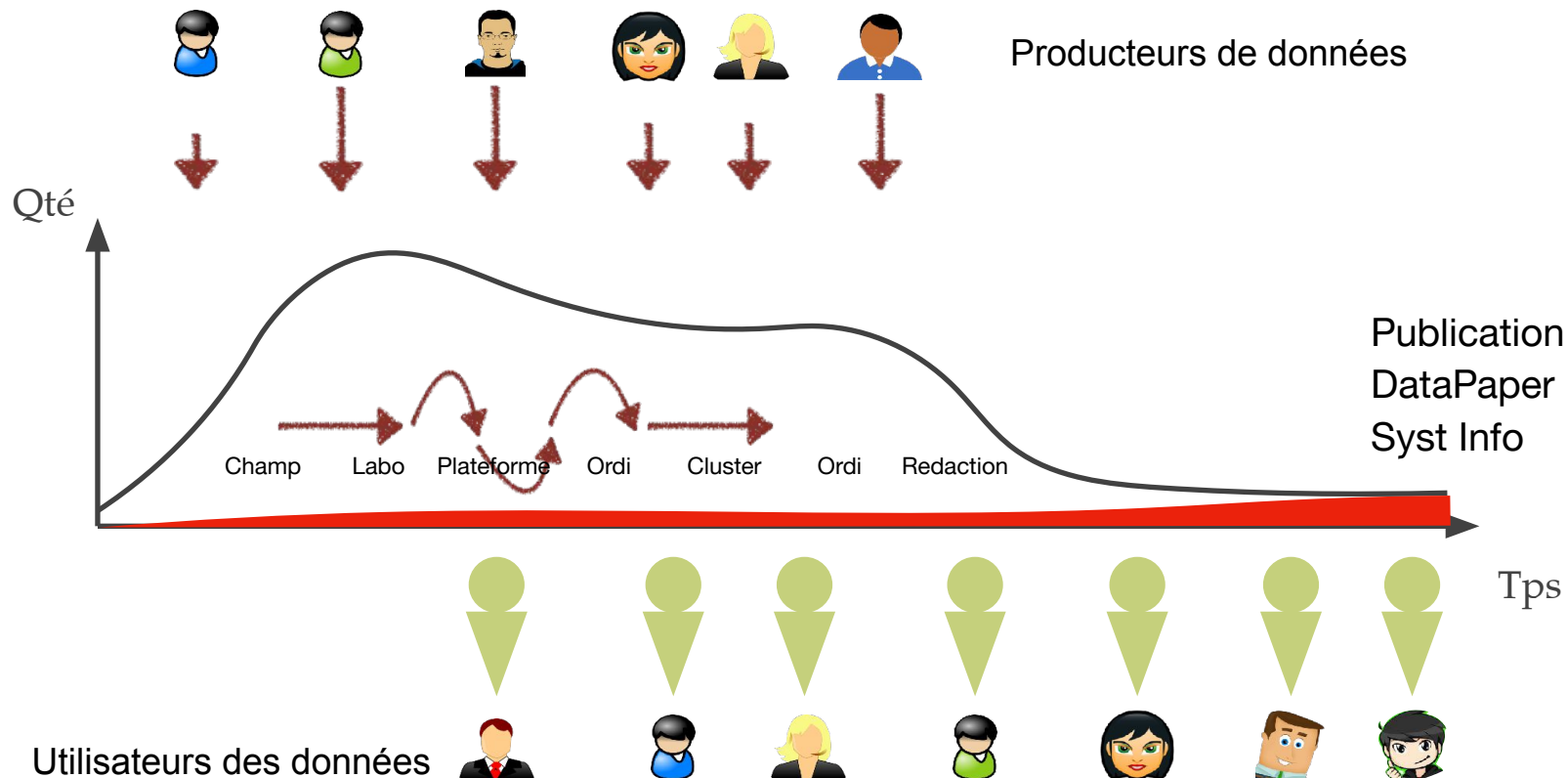
Publication

Confidentialité

Suppression



# Un projet sur la durée



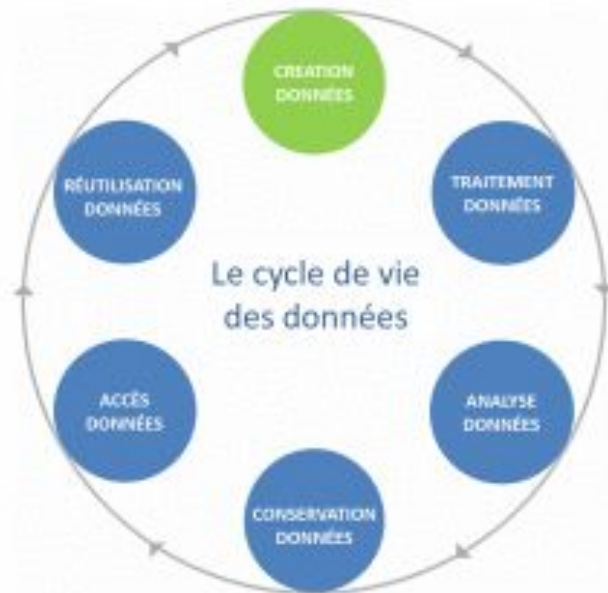


# Les étapes

Le [modèle](#) de UK Data Archive définit les six étapes suivantes :

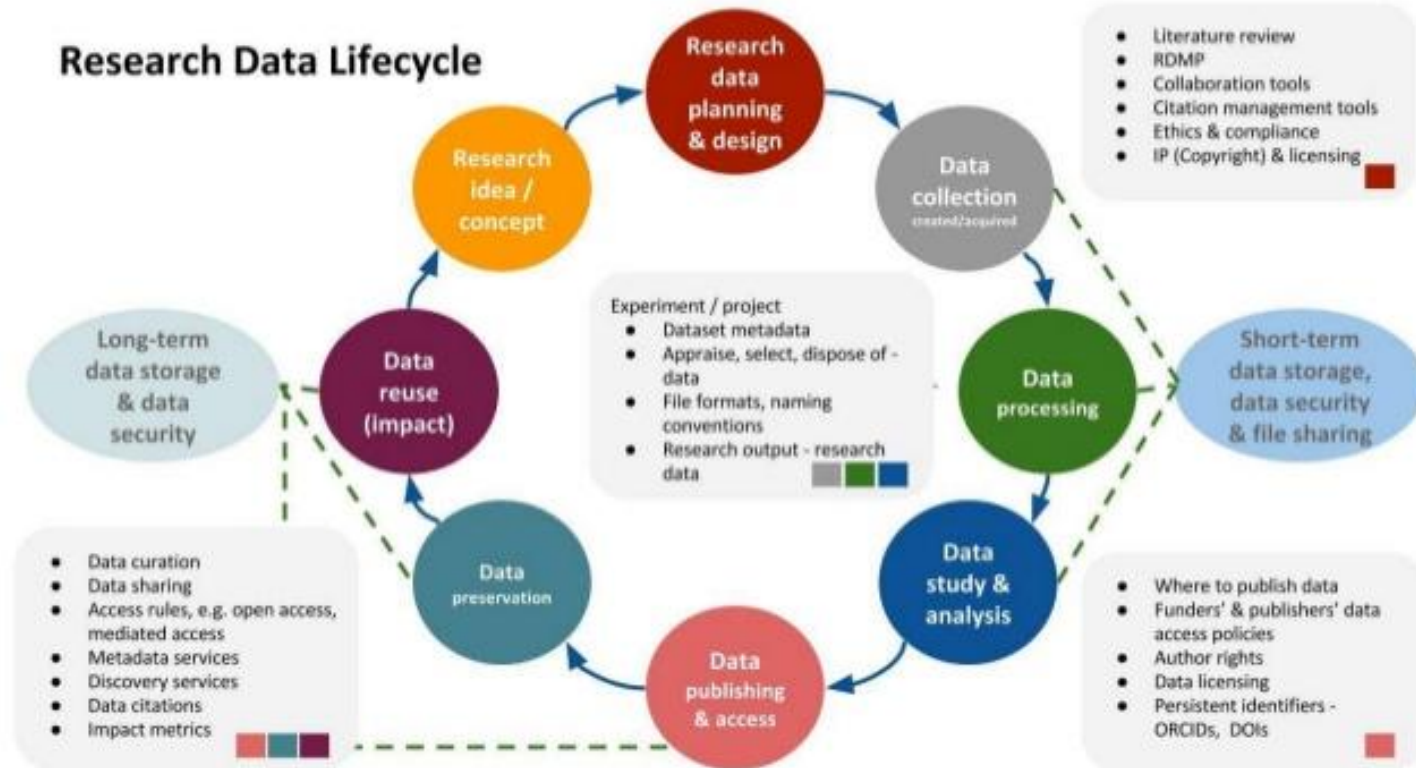
- **Création ou collecte** des données (creating data) ;
- **Traitement** des données (processing data) ;
- **Analyse** des données (analysing data) ;
- **Conservation** des données (preserving data) ;
- **Accès** aux données (giving access to data / data discovery) ;
- **Réutilisation** des données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



---

# 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)

# 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

# Les objectifs du PGD

- Assurer la reproductibilité des expériences
  - Décrire les données

<https://dmp.opidor.fr/>

IFB\_training WGS RNAseq Variant\_calling

**Quels méthodes et outils sont utilisés pour acquérir et traiter les données ? Précisez les différents formats dans lesquels les données seront disponibles aux différentes phases de la recherche**

**B I** [List] [List] [Link] [Table]

- Format : fastq, outil : séquenceur Illumina HiSeq 3500
- Format : fasta, outil : assembleur SOAP de novo
- Format : gff3, outil : Eugene

Enregistrer

Répondu 1 month ago par [helene.chiapello@inrae.fr](mailto:helene.chiapello@inrae.fr)

Commentaires (1)

Vérifier avec Gautier si le mode opératoire de l'Illumina est stabilisé

[frederic.de-lamotte@inrae.fr](mailto:frederic.de-lamotte@inrae.fr) 1 hour ago  
Modifier Supprimer

**Ajouter un commentaire à partager avec les collaborateurs**

**B I** [List] [List] [Link] [Table]

# Les objectifs du PGD





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

IFB\_training WGS RNAseq Variant\_calling

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

**Qui détiendra les droits sur les données et les autres informations créées lors du projet ?**

*Faire attention quand un partenaire privé amène des données dans le projet.*

**B** *I*    

par défaut INRAE sinon voir convention

Enregistrer

Répondu 1 month ago par [helene.chiapello@inrae.fr](mailto:helene.chiapello@inrae.fr)

Commentaires (1)

- j'ai un souci : depuis la loi LRN les données sont ouvertes par défaut, sauf exceptions (données personnelles, sensibles, code, autres...). Est-il encore opportun de parler de "propriété intellectuelle" dans ce cas
- je me réfère là à un commentaire de Lionel Maurel "La question de la propriété des données n'est pas forcément le bon angle d'attaque depuis LPRN (et principe d'ouverture par défaut). Aujourd'hui savoir qui est le propriétaire des données n'est plus souvent. Il s'agit de savoir si on est dans le principe de l'ouverture par défaut ou bien dans un cas d'exceptions." -- compte rendu atelier <https://mate-shs.cnrs.fr/actions/tutomate/tuto25-propriete-donnees-lionel-maurel/>
- en fait, je crois qu'ici cela demandera à terme une re-

# Les objectifs du PGD

- Permettre la réutilisation des données
  - Garantir la compréhension des données

Brève présentation des données générées, collectées ou réutilisées :


- Mode d'obtention, origine, type, nature et périmètre thématique des données
- Publications associées

<b>B</b>	<i>I</i>	☰	☰	🔗	📄
Mode d'obtention : sortie d'un workflow de variant calling					
Origine : vitis vinifera (la vigne)					
Type : variants au format VCF					

Enregistrer

INRAE Exemple de réponse +

Répondu 1 month ago par helene.chiapello@inrae.fr

Commentaires					
Ajouter un commentaire à partager avec les collaborateurs					
<b>B</b>	<i>I</i>	☰	☰	🔗	📄
Merci de rajouter un lien vers le workflow					
					
Enregistrer					

# Les objectifs du PGD

- Éviter les pertes de données
  - Assurer un stockage adapté

IFB\_training WGS RNAseq Variant\_calling

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

**Stockage : Quels seront les supports utilisés pour les données au cours du projet ?**

**B I** [Liste] [Liste] [Liste] [Liste] [Liste] [Liste]

Préciser l'infrastructure et la plateforme bioinformatique utilisée (IFB, Southgreen, migale,...)

Enregistrer

INRAE Exemple de réponse +

Répondu 1 month ago par helene.chiapello@inrae.fr

Commentaires

**Ajouter un commentaire à partager avec les collaborateurs**

**B I** [Liste] [Liste] [Liste] [Liste] [Liste] [Liste]

Jeff doit réserver 2 To avant le début du projet !





Enregistrer

# Les objectifs du PGD

- Établir le rôle de chacun
  - Définir 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  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

### \* Permissions

- Co-proprétaire: peut modifier les détails du projet, changer la visibilité et ajouter des collaborateurs.
- Éditeur: peut commenter et effectuer des changements
- Lecture seule: peut voir et commenter, mais ne peut pas faire de modifications

# Les objectifs du PGD

- 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.***

<b>B</b>	<i>I</i>	☰	☰	🔗	📄
En biologie ce sont jusqu'à maintenant les banques internationales (ENA, ArrayExpress,...) qui assurent l'archivage pérenne des données					
Pour les données de phénotypes : quels sont les usages ? Plutôt accès via des bases de données dédiées ? Pérennité des données peut être variable					

Commentaires
<b>Ajouter un commentaire à partager avec les collaborateurs</b>
<b>B</b> <i>I</i> ☰ ☰ 🔗 📄
<b>Enregistrer</b>



---

# 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)



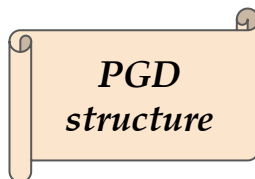
# Et si nous sortions de la préhistoire ?

## Dichotomie

### Structure

Plateforme  
Infrastructure  
Service...

PGD à durée indéterminé  
PGD plus générique



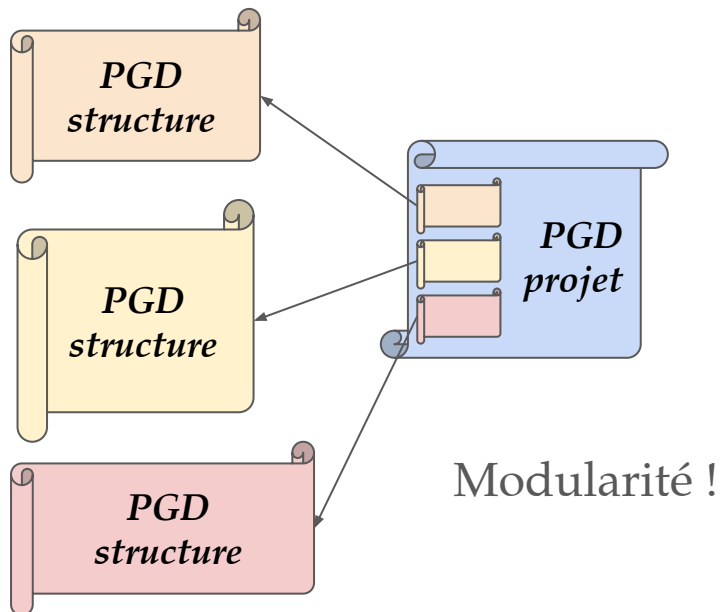
### Projet

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

PGD à durée déterminé  
PGD plus spécifique



# Et si nous sortions de la préhistoire ?



**Machine actionable DMP:** un plan de gestion lisible par les machines

**Objectif :** faire du Plan de gestion des Données un outil de configuration des environnements des infrastructures

