

FAIR

DATA

by IFB

 ARTbio
IBPS

 migale

 INSTITUT
PASTEUR

 PLANT
BIOINFORMATICS
FACILITY


INSTITUT FRANÇAIS DE
BIOINFORMATIQUE

 elixir
FRANCE

 cnrs

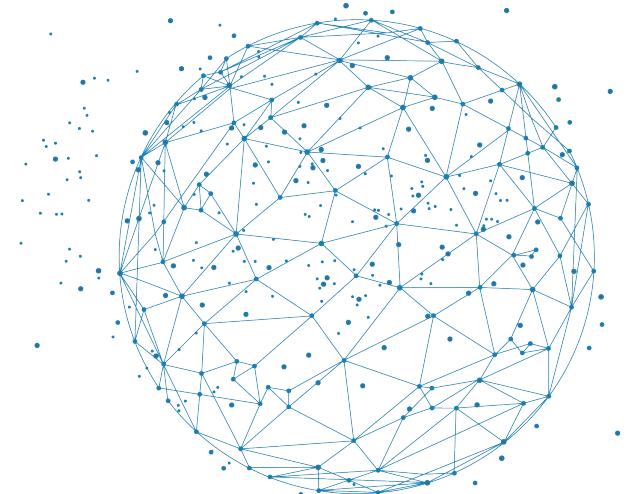
 INRAE

 Inserm

 cea

 INVESTIR
L'AVENIR

Introduction





- Montage de la formation par un Gdt IFB pendant le confinement en 2020



H. Chiapello



T. Denecker



J.-F. Dufayard



P. Lieby



L. Maurel



G. Sarah



J. Seiler



F. de Lamotte

- 2 éditions “nationales” en distanciel synchrone en 2021 : mars et octobre
- 2 éditions “régionales” en présentiel en 2022 : édition “francilienne” - Sorbonne Univ - 22 et 23 mars + édition “Grand Est” - Strasbourg les 23 et 24 mars



- ARTbio



Christophe Antoniewski



Naïra Naouar



Léa Bellenger (helper)

- Migale



Valentin Loux



Cédric Midoux

- IFB



Hélène Chiapello



Thomas Denecker



Frédéric de Lamotte

- PlantBioinfoPF



Célia Michotey



Cyril Pommier

- Institut Pasteur Paris



Anne-Caroline Delétoille



Fanny Sébire



Lucie Lamothe (helper)



- Horaires : 9h30 -12h30 & 13h30 - 17h30 les mardi et mercredi, 1 pause vers 10h15 et vers 15h15
- En présentiel
- ... donc 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
- Espace Slack
 - Inscrivez vous à <slack.com>
 - [Tutoriel de connexion](#)
 - Rejoignez le canal #fairdata-paris-mars2022
-



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)



- **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

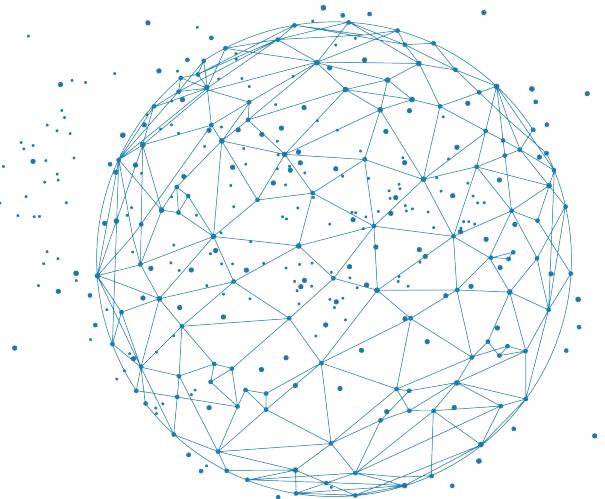
- **Module 4 : Partage et valorisation des données**

Cadre Juridique - Les licences sur les données - Quel entrepôt pour quelles données ?

- Supports ici : <https://moodle.france-bioinformatique.fr/course/view.php?id=8>



- 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
 - Décide de la proposition qu'on lui soumet



Module 1 : Les données de la Recherche





Christophe Antoniewski – <https://orcid.org/0000-0001-7709-2116>

Célia Michotey – <https://orcid.org/0000-0003-1877-1703>

Naïra Naouar – <https://orcid.org/0000-0003-2161-8603>

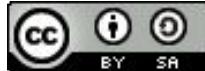
Fanny Sébire – <https://orcid.org/0000-0002-6301-7147>

Merci à :

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

Jean-François Dufayard - <https://orcid.org/0000-0002-7427-6822>

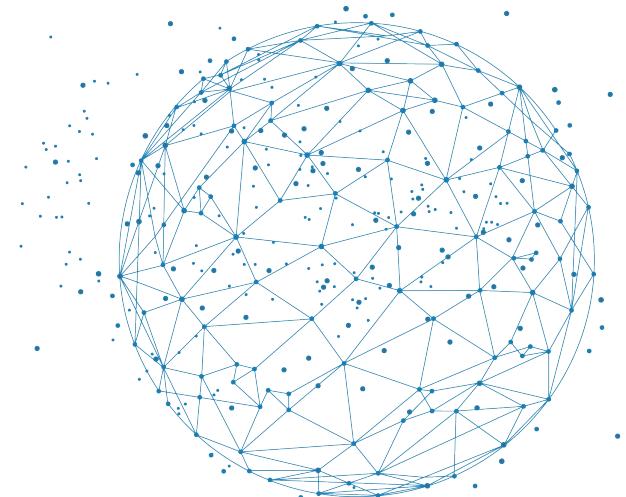
Paulette Lieby - <https://orcid.org/0000-0002-9289-9652>



<https://www.youtube.com/watch?v=YcuE54E9col>

FAIR pourquoi faire ?

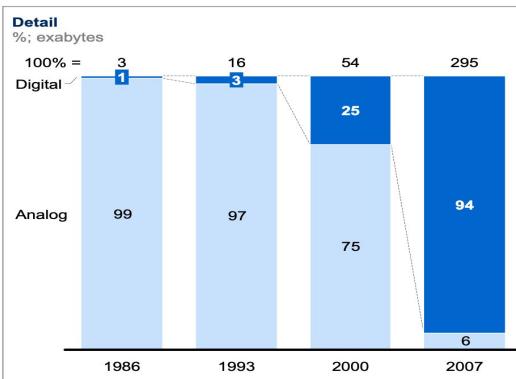
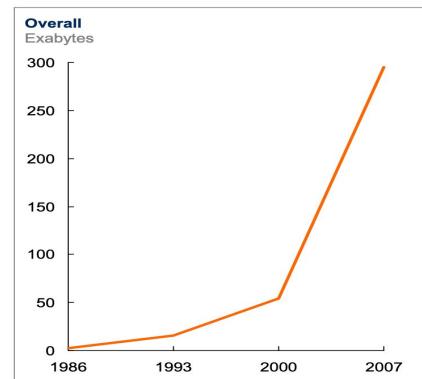
Une réponse aux crises des sciences biomédicales



La disruption numérique, une bascule brutale



1 000 000 000 000 000 000 bytes (octets) 1 exaoctet = 10^{18} octets, 1 zetaoctet = 10^{21} octets
eo po to go mo ko



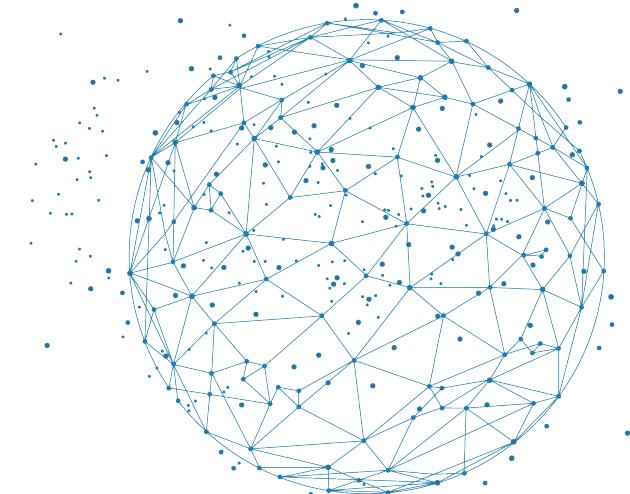
The World's Technological Capacity to Store, Communicate, and Compute Information. Science. 2011;332.



- 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)



Les technologies numériques ont transformé la biologie et médecine



Dites nous tout sur <http://scrumblr.ca/disruption> !

Le déluge des données en Science

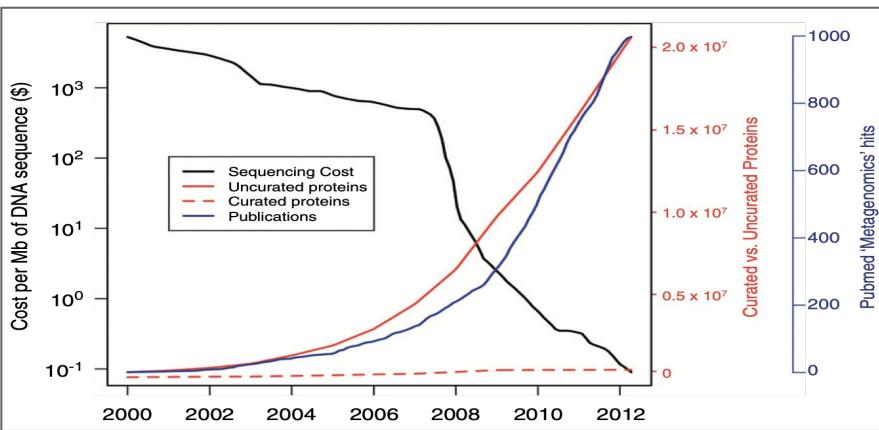


Les techniques à haut débit induisent une chute des coûts et une explosion de la production de données

Génome humain :

en 1990 = **13 ans et 3 Milliards \$**

en 2015 = **quelques heures et 1000 \$**



→ La quantité de données à stocker et analyser explose

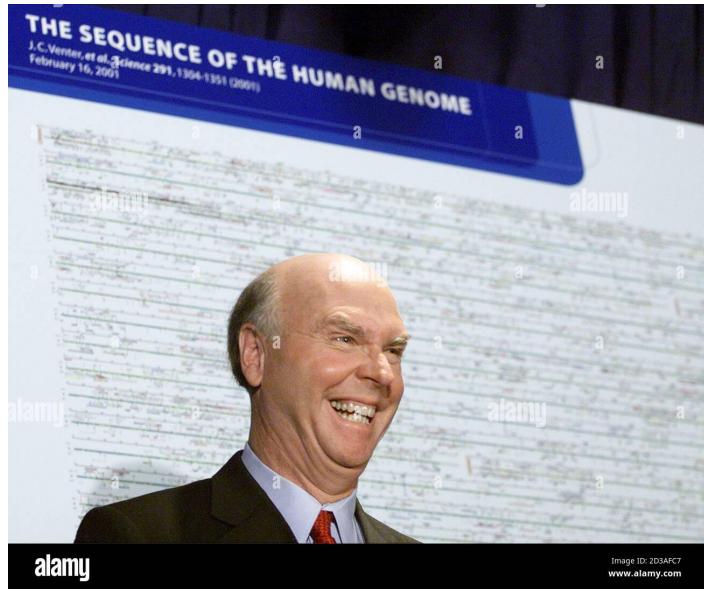
→ Le **rendement** d'analyse chute

Un changement d'échelle qui impacte profondément la science



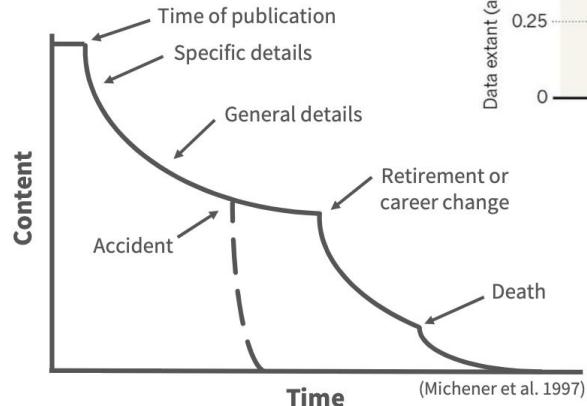
Les étapes	AVANT	MAINTENANT
Concevoir l'expérimentation	Une connaissance des dispositifs expérimentaux existants accessible à un ou quelques individus. Un volume d'observations attendues à taille humaine	Une matrice de technologies qui échappe à un expérimentateur individuel Possibilité d'utiliser comme de générer une quantité massive de données expérimentales
Collecter des résultats		
Analyser des résultats		
Diffuser le savoir		

Un changement d'échelle qui modifie la place de la science dans la société





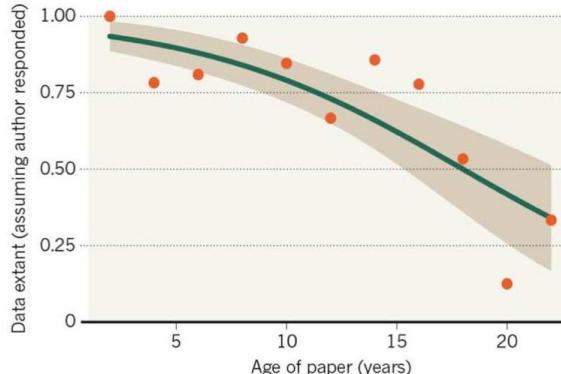
Data Entropy



3

MISSING DATA

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

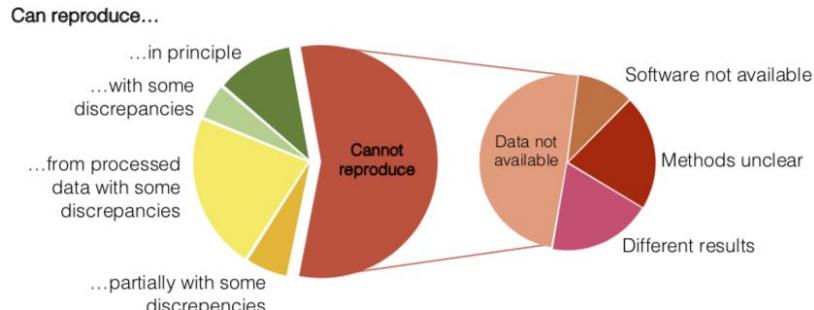


Vines, T. H. et al. Curr. Biol. <http://dx.doi.org/10.1016/j.cub.2013.11.014> (2013).

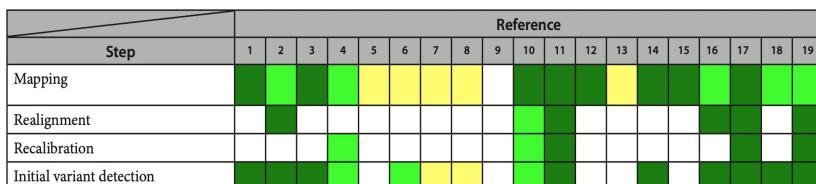
DataONE

La crise de reproductibilité

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



Repeatability of published microarray gene expression analyses.
Nat Genet. **2009**;41: 149–155. doi:10.1038/ng.295



Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. Nat Rev Genet. **2012**;13: 667–672.

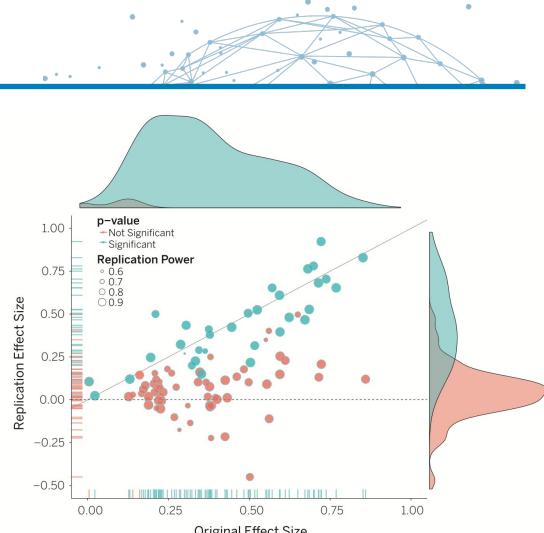
RESEARCH ARTICLE
Estimating the reproducibility of psychological science

Open Science Collaboration^{1,†}
★ all authors and affiliations

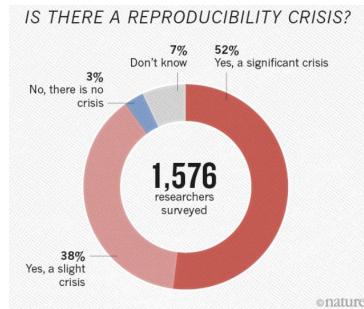
Science 28 Aug 2015;
Vol. 349, issue 6251, aac4716
DOI: 10.1126/science.aac4716

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.



Estimating the reproducibility of psychological science.
Science. **2015**;349: aac4716. doi:10.1126/science.aac4716



Is there a reproducibility crisis in science?
Nature. **2016**.
doi:10.1038/d41586-019-00067-3

Les premières formalisations du FAIR



- Accept that the computational component is becoming an integral component of biomedical research
- Always provide access to primary data
- Record versions of all auxiliary data sets used during the analysis
- Note the exact versions of software used
- Record all parameters even if defaults are used
- Provide all custom scripts
- Do not reinvent the wheel

Anton Nekrutenko and James Taylor
Nat Rev Genet. 2012;13: 667–672.

Box 3 | Guidelines for reproducibility

- **Accept that the computational component is becoming an integral component of biomedical research.** As the life sciences are becoming increasingly data-driven, there will be no escape from computation and data handling. Familiarize yourself with best practices of scientific computing using existing educational resources, such as the Software Carpentry project¹¹. Implementing good computational practices in your group will automatically take care of many of the points listed below.
- **Always provide access to primary data.** It is obvious that without access to the original data sets, any claims made in a publication cannot be verified. In situations in which the data cannot be made public (for example, clinical data sets under Institutional Review Board protection), they should be deposited in controlled access repositories (such as cBioCancerGrid¹²), where they can be retrieved by authorized users. One potential issue with this point is the fact that there is currently a debate on what constitutes primary data. Storing images generated by some next-generation sequencing (NGS) machines on a large scale has long been unfeasible. Public sequencing archives, such as those at the National Center for Biotechnology Information (NCBI) and the European Bioinformatics Institute (EBI), are still accepting sequencing reads as submissions and should be used. Going forward, other formats, such as aligned data in BAM format, are likely to be used (as is already done by the 1000 Genomes Project).
- **Record versions of all auxiliary data sets used during the analysis.** For example, in most NGS analyses, such as variant discovery detailed here, sequencing reads are compared against a reference genome. It is crucial to record which reference genome was used because, just as software has versions and cars have model years, genomes have build identifiers. For example, the latest human genome build distributed by the UCSC Genome Browser is called hg19 (it is derived from the GRCh37 build prepared by the Genome Reference Consortium) and has the highest number of functional annotations (7,330 annotation types) and should be the preferred version to be used. Note that the latest version may not always be the best choice. The latest mouse genome build (mm10) has only a fraction of annotations (258 tracks) compared with its predecessor (mm9, which has 2,096 tracks). Thus, it would be easier to interpret results of an NGS experiment mapped to the mm9 build even though mm10 has an additional 48 megabases of actual sequence.
- **Note the exact versions of software used.** Different versions of the same software often produce different results, and important bug fixes may have implications to results produced with a particular version of a tool.
- **Record all parameters, even if defaults are used.** Although the reason to record all parameters requires no explanation, we emphasize the importance of explaining default settings for reproducibility. A clause ‘software was used with default settings’ is found in many publications. However, the meaning of default settings often changes between versions of software and can be quite difficult to track down when a substantial amount of time has passed since publication. Thus, record what the default settings actually are.
- **Provide all custom scripts.** With the complexity of NGS analysis, it is often unavoidable to create simple scripts that carry out such straightforward tasks as, for example, changing data formats. Such scripts must be made accessible as any other part of the analysis.
- **Do not reinvent the wheel.** It pays to reuse existing software. Integrative frameworks and associated application stores already house hundreds of tools (for example, as of May 2012, Galaxy ToolShed contains ~1,700 tools). It is likely that a script for a particular problem has been already written. Ask around through existing resources such as SEQanswers¹³ and BioStar¹⁴.

Les racines du mouvement "#OpenScience"



- La crise de reproductibilité
- La crise de déontologie scientifique (P-Hacking)
P-hacking in clinical trials and how incentives shape the distribution of results across phases.
Proc Natl Acad Sci U S A. 2020;117: 13386–13392.)
- La crise politique (éthique/démocratique)
(indicateurs-ego metrics, producteurs de savoir biomedical privé, marché de l'édition scientifique)
- La crise de la "bioinformatique"

→Améliorer la science et l'innovation



démocratiser l'accès aux savoirs



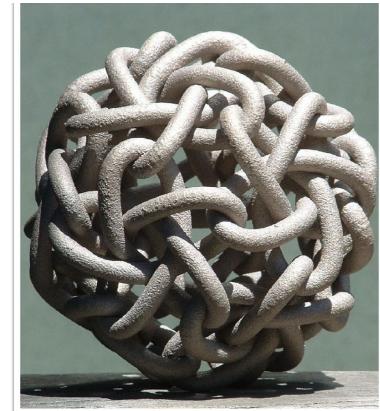
rendre la science plus cumulative, plus fortement étayée par les données, plus transparente



augmenter l'efficacité de la recherche en évitant de dupliquer les efforts, en réutilisant des données ou du matériel scientifique



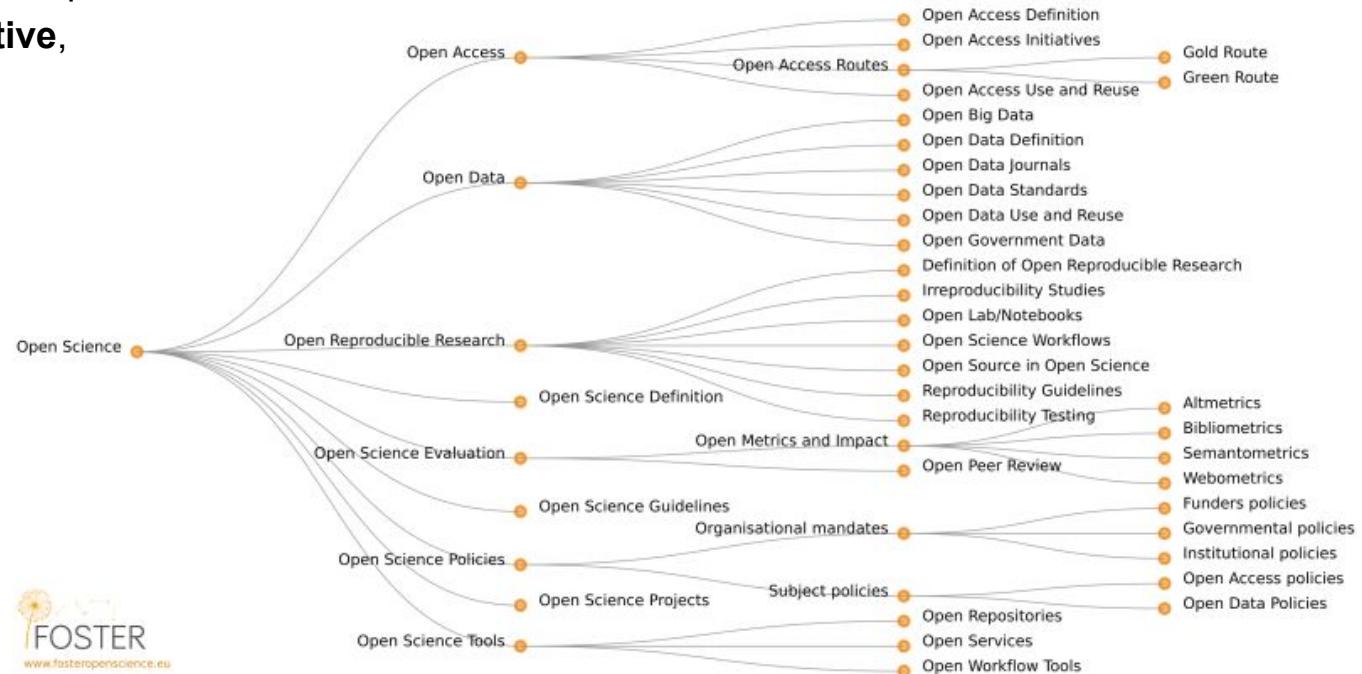
favoriser la confiance des citoyens dans la science





Nouvelle forme de pratique de la science : **collaborative, participative et interdisciplinaire**

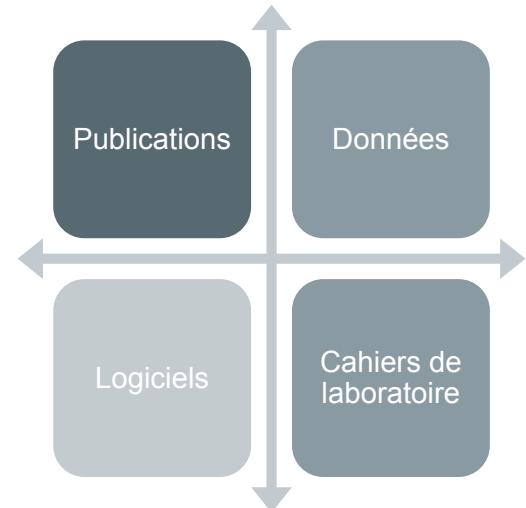
Open Science Taxonomy



www.fosteropenscience.eu



- Rendre les **résultats de la recherche scientifique** accessibles à tous (un des aspects de la Science Ouverte)
- En permettant une diffusion et une réutilisation sans entrave des données et résultats de la recherche



Contexte national



2018



2021

Mesures

4

Mettre en œuvre l'obligation de diffusion des données de recherche financées sur fonds publics

5

Créer Recherche Data Gouv, la plateforme nationale fédérée des données de la recherche

6

Promouvoir l'adoption d'une politique de données sur l'ensemble du cycle des données de la recherche, pour les rendre faciles à trouver, accessibles, interopérables et réutilisables (FAIR)



Citez les types de données de la recherche que vous connaissez

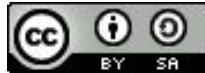
En utilisant le tableau à Post-It disponible ici <http://scrumblr.ca/typesdedonnees>

Définition OCDE

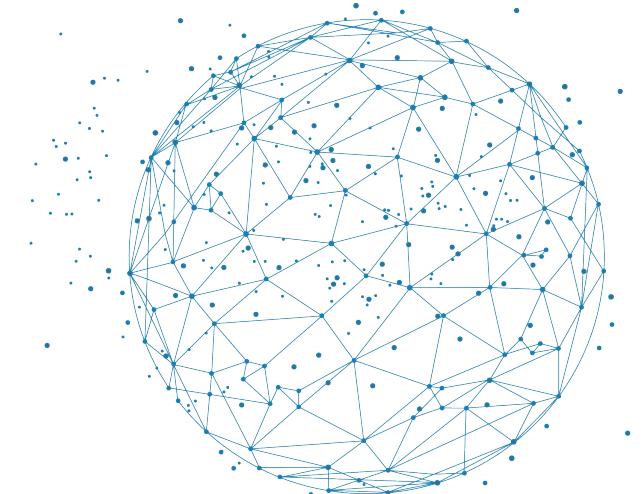


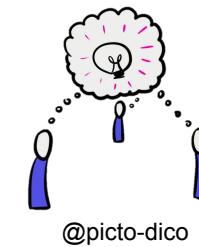
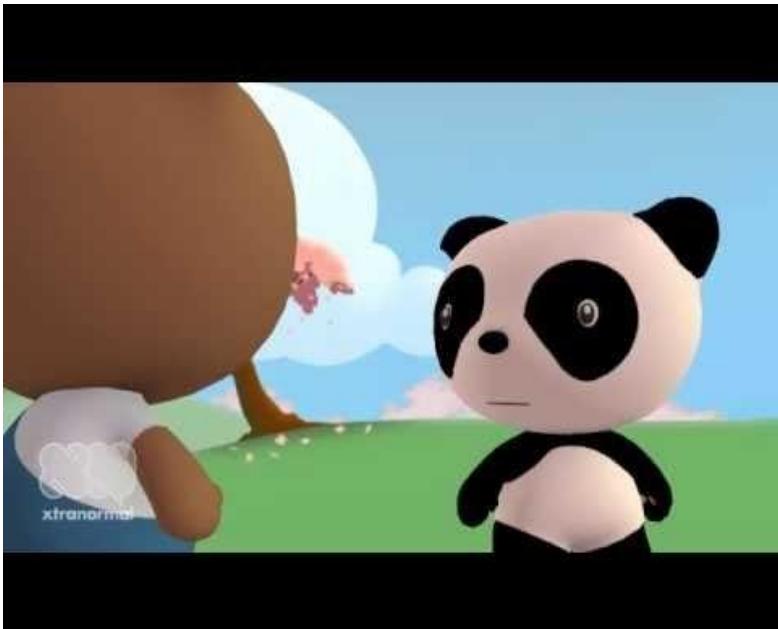
L'Organisation de coopération et de développement économiques (OCDE) est une organisation internationale qui œuvre pour la mise en place de politiques meilleures pour **une vie meilleure**. Notre objectif est de promouvoir des politiques publiques qui favorisent la prospérité, l'égalité des chances et le bien-être pour tous.

- 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.



Vers le FAIR





@picto-dico

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

https://youtu.be/66oNv_DJuPc



<https://www.menti.com/xtqhkq1kf4>

- Notez 5 conditions nécessaires
 - 1 seul mot à la fois
 - en français
 - sans majuscule
- A partir du nuage de mots créé collectivement quels regroupements pouvons-nous faire ?

Les principes FAIR sont un ensemble de lignes directrices visant à rendre les données **Facilement trouvables, Accessibles, Interopérables et Réutilisables.**

- Ils 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
- Ils mettent l'accent sur la capacité des machines à gérer des données de façon automatique, avec le minimum d'interventions humaines

Références :

- Wilkinson et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). <https://doi.org/10.1038/sdata.2016.18>
- Les principes FAIR. DORANum. <http://doi.org/10.13143/z7s6-ed26>



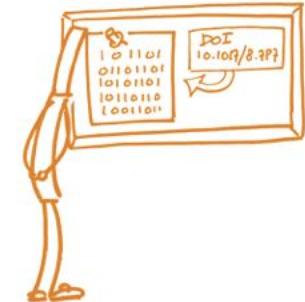
Faciliter la découverte des données (et de leurs métadonnées)
tant pour les humains que pour les machines



- Les données ont un **PID** (Persistent Identifier, identifiant unique et pérenne)
- Les données sont décrites par des **métadonnées**
- Ces métadonnées incluent le PID des données qu'elles décrivent
- Les données sont déposées dans un **entrepôt de données**



Permettre l'accès aux données et leur téléchargement,
ce qui peut inclure l'authentification et l'autorisation.



- 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 pas (disparues ou inaccessibles)



Permettre l'exploitation et l'intégration des données
quel que soit l'environnement informatique utilisé



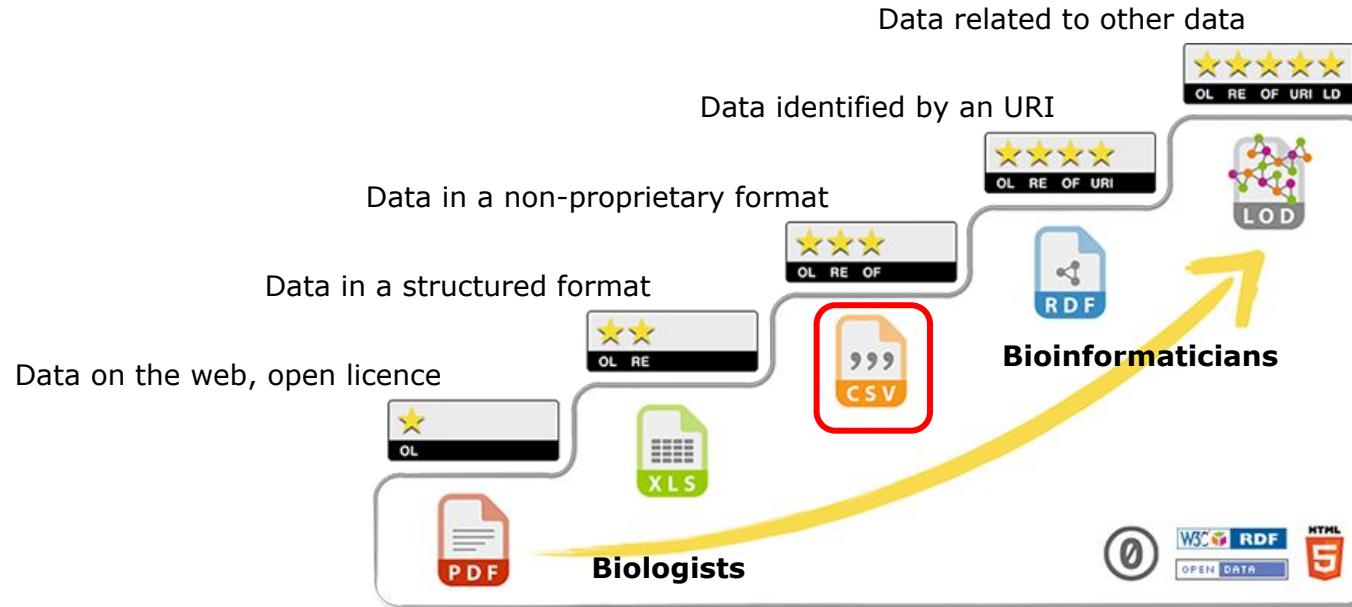
- Les données sont **décrivées avec un vocabulaire contrôlé**
- Le vocabulaire utilisé **respecte les principes FAIR**
- Les **métadonnées sont contextualisées** avec des liens vers d'autres données



Permettre la réutilisation des données pour de futures recherches



- Les métadonnées contiennent toutes les informations qui peuvent être utiles (**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**



La progression vers FAIR et l'Open Data nécessite une coopération multidisciplinaire :

- Biologistes
- Bioinformaticiens
- Spécialistes des ontologies/sémantiques

Evaluation de la FAIRness des données



■ Grille SHARC (SHAring Reward & Credit)

- RDA (Research Data Alliance)

- Grille simplifiée :

https://zenodo.org/record/2551500#.X4hc_-re70

- Grille complète :

<https://zenodo.org/record/3922069>

1) FINDABLE (8 essential criteria)	
Indexed identifier ? Identification	<input type="checkbox"/> Never/NA <input type="checkbox"/> If Mandatory <input type="checkbox"/> Sometimes <input type="checkbox"/> Always Are each data/dataset identified by an indexed and independent identifier?
Unique, global, persistent ID? Identification	<input type="checkbox"/> Never/NA <input type="checkbox"/> If Mandatory <input type="checkbox"/> Sometimes <input type="checkbox"/> Always Are the data identifiers unique, global and persistent?
ID scheme? Identification	<input type="checkbox"/> Never/NA <input type="checkbox"/> If Mandatory <input type="checkbox"/> Sometimes <input type="checkbox"/> Always Has any identifying schema been used for data (e.g. DOI)?
Persistent metadata / data link ? Metadata traceability	
<input type="checkbox"/> Never/NA <input type="checkbox"/> If Mandatory <input type="checkbox"/> Sometimes <input type="checkbox"/> Always Are the metadata linked to the dataset through a persistent identifier?	
Metadata & authority linked ? Metadata traceability	<input type="checkbox"/> Never/NA <input type="checkbox"/> If Mandatory <input type="checkbox"/> Sometimes <input type="checkbox"/> Always Are the metadata of each dataset linked to a unique authority (responsible for the datasets at a given time)?
Datasets linked to authority? Metadata traceability	<input type="checkbox"/> Never/NA <input type="checkbox"/> If Mandatory <input type="checkbox"/> Sometimes <input type="checkbox"/> Always Are all datasets linked to an authority (e.g. institution, organization)?
3) INTEROPERABLE (2 essential criteria)	
Standard vocabularies, thesaurus, ontologies or data dictionary ? Identification	<input type="checkbox"/> Never/NA <input type="checkbox"/> If Mandatory <input type="checkbox"/> Sometimes <input type="checkbox"/> Always Are standard vocabularies, thesaurus or ontologies used for all data types present in datasets, to enable interoperability between well defined domains? If not, is a well-defined open data dictionary provided?
Interoperability criteria explained? Identification	<input type="checkbox"/> Never/NA <input type="checkbox"/> If Mandatory <input type="checkbox"/> Sometimes <input type="checkbox"/> Always Are the interoperability criteria explained?
Result for Interoperability: .../2 Never/NA .../2 If Mandatory .../2 Sometimes .../2 Always	
Result for Findable:	
2) ACCESSIBLE (3 essential criteria)	
Data repositories? Repository	<input type="checkbox"/> Never/NA <input type="checkbox"/> If Mandatory <input type="checkbox"/> Sometimes <input type="checkbox"/> Always Does the researcher use data repositories for the storage of data?
Efficient and rich services for data security and services Data security and services	<input type="checkbox"/> Never/NA <input type="checkbox"/> If Mandatory <input type="checkbox"/> Sometimes <input type="checkbox"/> Always Does the researcher use efficient and rich services to protect data?
Data access restriction Access restriction	<input type="checkbox"/> Never/NA <input type="checkbox"/> If Mandatory <input type="checkbox"/> Sometimes <input type="checkbox"/> Always In case of a non legal restricted access
Result for Accessible:	
4) REUSABLE (5 essential criteria)	
Relevant actions for data reuse potential? Data potential	<input type="checkbox"/> Never/NA <input type="checkbox"/> If Mandatory <input type="checkbox"/> Sometimes <input type="checkbox"/> Always Which relevant actions have been undertaken by the researcher to enhance the data reuse potential?
Provenance for raw and transformed data? Data traceability	<input type="checkbox"/> Never/NA <input type="checkbox"/> If Mandatory <input type="checkbox"/> Sometimes <input type="checkbox"/> Always Are the provenance and type of all data properly specified (origin of raw, primary, transformed, secondary...)?
Information on methods and tools that permit the understanding, integrity of data? Reusability tools	<input type="checkbox"/> Never/NA <input type="checkbox"/> If Mandatory <input type="checkbox"/> Sometimes <input type="checkbox"/> Always Does the researcher provide information on methods and tools that permit the understandability, integrity, value and readability of data intended to be kept on the long-term? (e.g. versioning, archival and long term reuse issue for protocols, software, required methods and contexts to create, read and understand data)
Data sharing arrangements meet data ethics and protection? Reusability right	<input type="checkbox"/> Never/NA <input type="checkbox"/> If Mandatory <input type="checkbox"/> Sometimes <input type="checkbox"/> Always Do the data reuse control and data sharing arrangements meet the data protection and 'local/national' ethics requirements?
Legal reuse restriction properly justified? Reusability right	<input type="checkbox"/> Never/NA <input type="checkbox"/> If Mandatory <input type="checkbox"/> Sometimes <input type="checkbox"/> Always In case of a legal reuse restriction (such as personal data, state and public security, national defense secret, confidentiality of external relations, information systems security, secrets in industrial and commercial matters), is the restriction properly justified?
Result for Reusable: .../5 Never/NA .../5 If Mandatory .../5 Sometimes .../5 Always	
TOTAL FAIR simple criteria evaluation results: .../18 'Never/NA' .../18 'If Mandatory' .../18 'Sometimes' .../18 'Always'	
*advice will be provided according to the criteria predominantly obtained	

Evaluation de la FAIRness des données

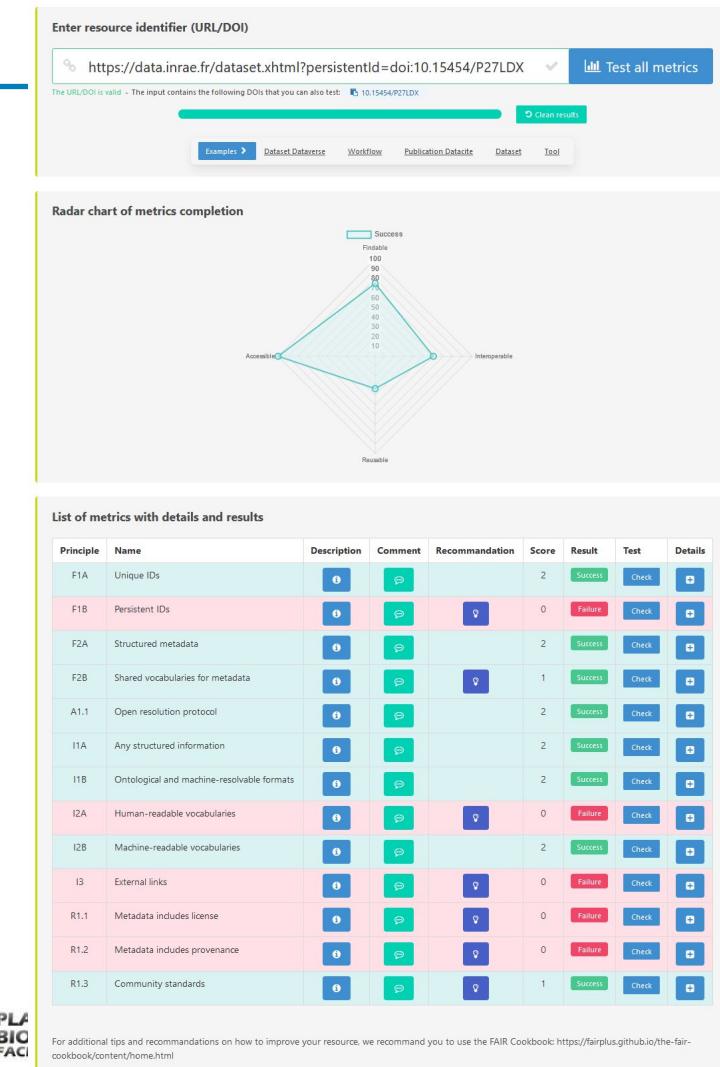
- Grille SHARC (SHAring Reward & Credit)
 - RDA (Research Data Alliance)
 - https://zenodo.org/record/2551500#.X4hC_-2re70
- FAIR self assessment tool
 - ARDC (Australian Research Data Commons)
 - <https://ardc.edu.au/resources/aboutdata/fair-data/fair-self-assessment-tool/>

Total across F.A.I.R

Findable	
Does the dataset have any identifiers assigned?	No identifier
Is the dataset identifier included in all metadata records/files describing the data?	No
How is the data described with metadata?	The data is not described
What type of repository or registry is the metadata record in?	The data is not described in any repository
Accessible	
How accessible is the data?	No access to data or metadata
Is the data available online without requiring specialised protocols or tools once access has been approved?	No access to data
Will the metadata record be available even if the data is no longer available?	Unsure
Interoperable	
What (file) format(s) is the data available in?	Mostly in a proprietary format
What best describes the types of vocabularies/ontologies/tagging schemas used to define the data elements?	Data elements not described
How is the metadata linked to other data and metadata (to enhance context and clearly indicate relationships)?	There are no links to other metadata
Reusable	
Which of the following best describes the license/usage rights attached to the data?	No license
How much provenance information has been captured to facilitate data reuse?	No provenance information is recorded

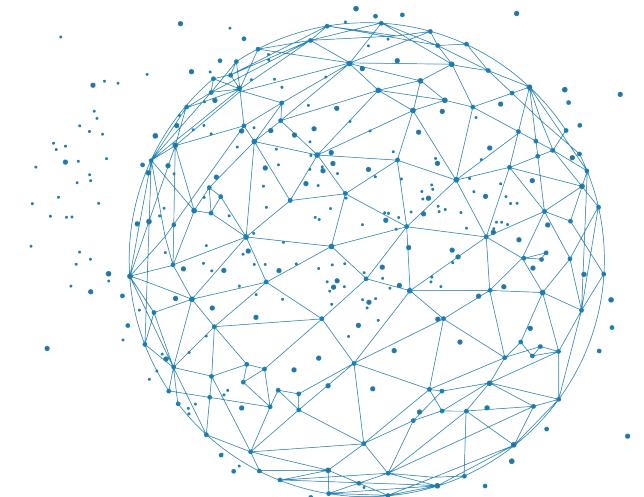
Evaluation de la FAIRness des données

- Grille SHARC (SHAring Reward & Credit)
 - RDA (Research Data Alliance)
 - https://zenodo.org/record/2551500#.X4hC_-2re70
- FAIR self assessment tool
 - ARDC (Australian Research Data Commons)
 - <https://ardc.edu.au/resources/aboutdata/fair-data/fair-self-assessment-tool/>
- FAIR checker
 - IFB
 - https://fair-checker.france-bioinformatique.fr/base_metrics





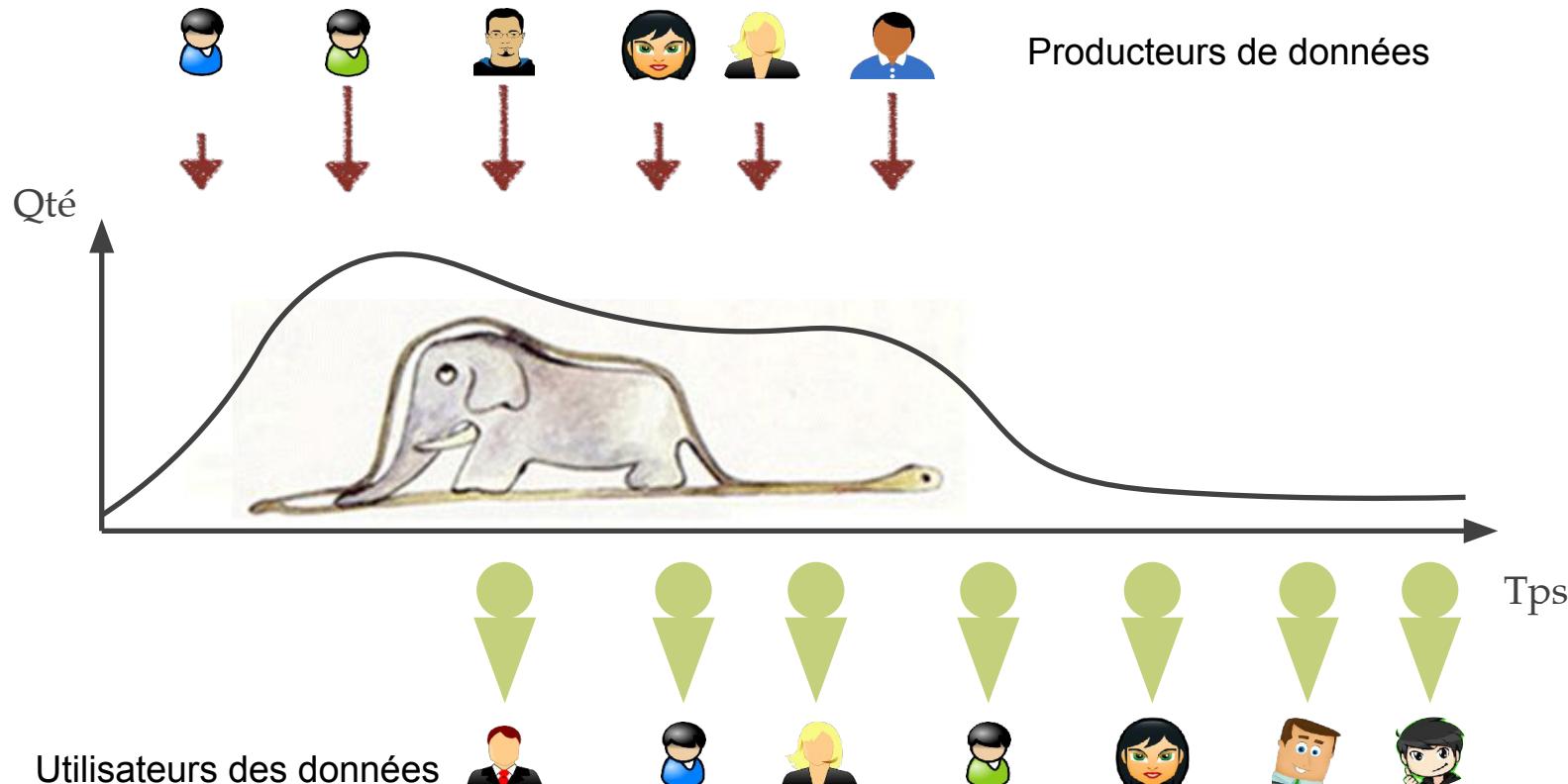
Cycle de vie des données





- 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 ...





Connectez vous sur le scrumblr : <http://scrumblr.ca/cycledata>

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

DEBUT

Budget

Récolte

Volumétrie

MILIEU

identification

Analyse Nettoyage,
vérification, sélection

FIN

Ethique

Sécurité

Publication

Transfert

Format

Stockage

Partage

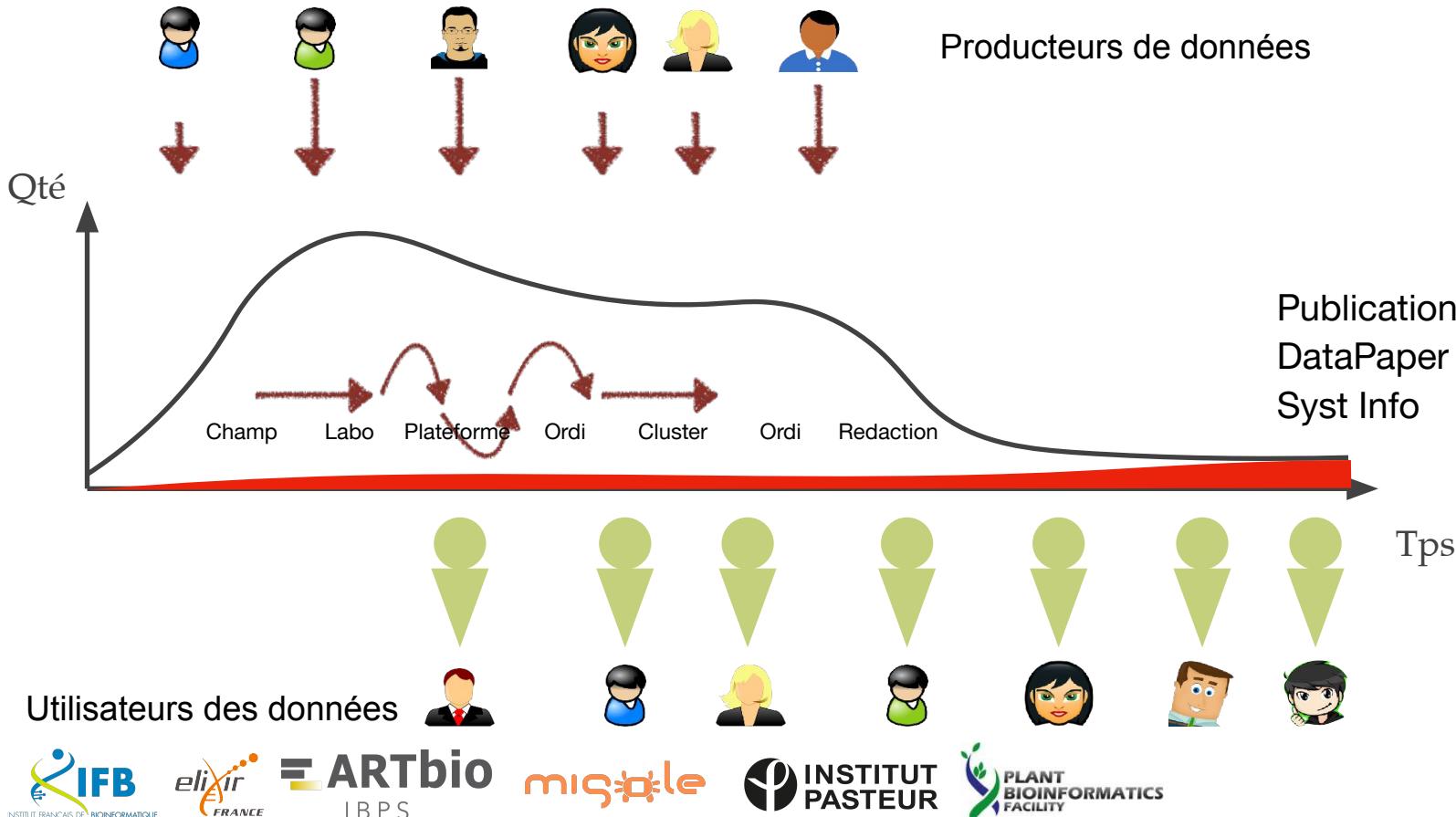
Classement

Versionning

Confidentialité

Suppression

Intégrité

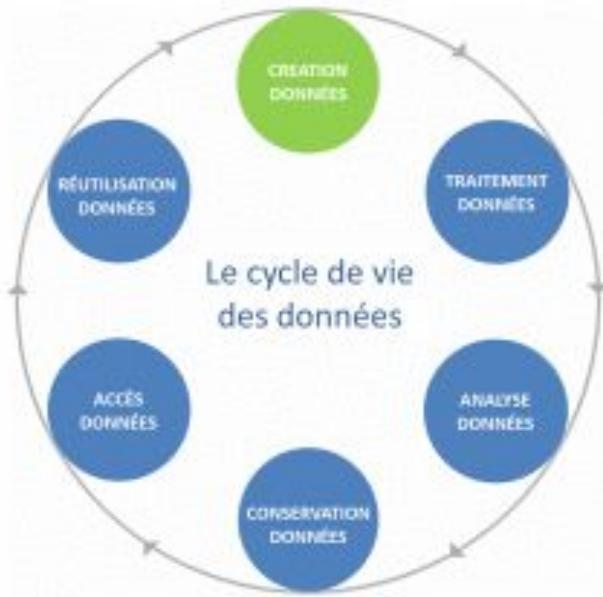




Le modèle de UK Data Archive définit les six étapes suivantes :

- **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).

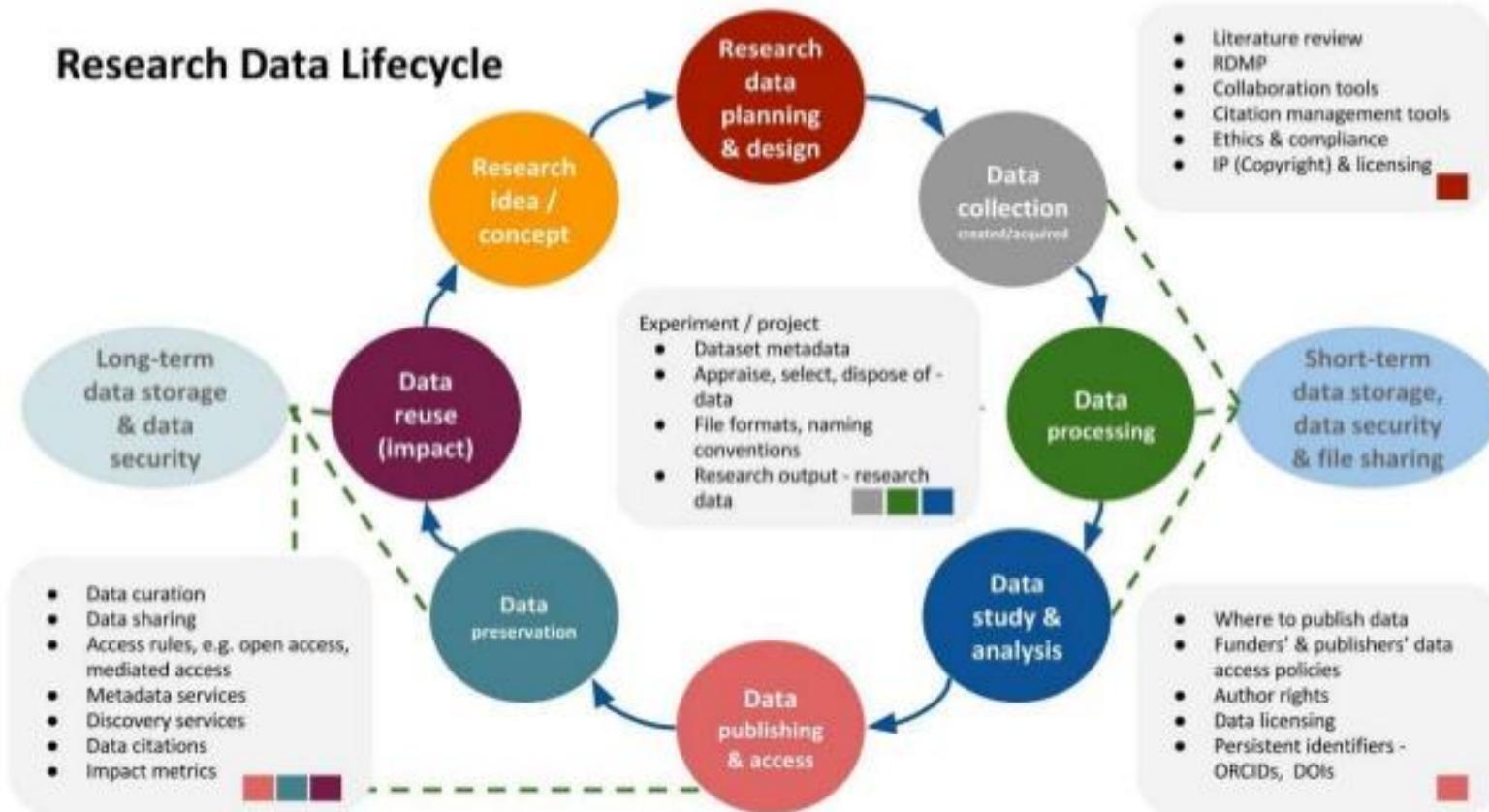
Une introduction à la gestion et au partage des données de la recherche - Le cycle de vie des données



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



Research Data Lifecycle





- **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**

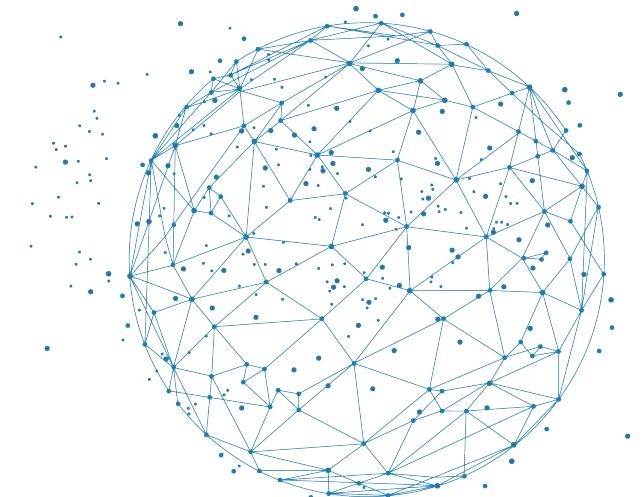
- des petits, des gros
- un peu partout (PC, cloud, cluster)
- des données brutes, du code, des résultats

- **De la connaissance**

- des méthodes, du code
- des systèmes d'information
- des publications



Plan de Gestion de Données





PGD de projet

Document qui définit comment seront gérées les données d'un projet **pendant et après le projet**

- La difficulté : penser à traiter toutes les étapes du cycle de vie des données
- L'avantage : les modèles vous aident à penser à tout, en vous posant une série de questions

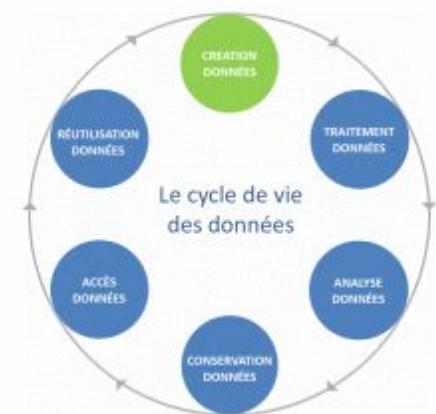
QUAND

Document évolutif = au moins 3 versions :

- Première version au début du projet de recherche
- Mises à jour régulières au cours du projet : versions intermédiaires
- Version finale à la fin du projet

QUI

Rédaction par l'équipe de recherche coordinatrice du projet





Modèle de PGD = une liste de questions à remplir pour rédiger un PGD

Modèles existants :

- Modèles des financeurs : ANR, Commission Européenne...
- Modèle Science Europe - <https://doi.org/10.5281/zenodo.4915862>
- Modèles institutionnels, des centres de calcul/stockage... (INRAe, CEA, Institut Pasteur, IN2P3, Universités, Grandes écoles...)

Peu / pas de modèles spécifiques à des types de données



- **Plan** : on planifie (donc on anticipe)
 - Se poser les bonnes questions, le plus tôt possible
- **Gestion** : on gère, on fait fructifier (on commence déjà par ne plus perdre)
 - Penser à toutes les démarches à effectuer à chaque étape du cycle de vie des données
- **Données** : à bien définir au préalable



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

Pour l'équipe de recherche

Se référer au PGD pour :

- Retrouver les données
- Comprendre les données
- Savoir où sont conservées les données ...

Pour la communauté scientifique

Publier le PGD pour indiquer :

- Quelles données existent
- Où elles sont conservées
- Qui peut y accéder, sous quelles conditions...



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

tout développer | tout réduire

1.1 Description générale du produit de recherche >

1.2 Est-ce que des données existantes seront réutilisées ? >

1.3 Comment seront produites/collectées les nouvelles données ? >

→ Objectif : Assurer la reproductibilité des expériences

- Décrire les données



2. Documentation et qualité des données

tout développer | tout réduire

2.1 Quelles métadonnées et quelle documentation (par exemple mode d'organisation des données) accompagneront les données ?



2.2 Quelles seront les méthodes utilisées pour assurer la qualité scientifique des données ?



4. Traitement et analyse des données

tout développer | tout réduire

4.1 Comment et avec quels moyens seront traitées les données ?

- Objectif : Faciliter la réutilisation des données
- Garantir la compréhension des données



3. Exigences légales et éthiques, code de conduite

tout développer | tout réduire

3.1 Quelles seront les mesures appliquées pour assurer la protection des données à caractère personnel ? >

3.2 Quelles sont les contraintes juridiques (sensibilité des données autres qu'à caractère personnel, confidentialité, ...) à prendre en compte pour le partage et le stockage des données ? >

3.3 Quels sont les aspects éthiques à prendre en compte lors de la collecte des données ? >

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



5. Stockage et sauvegarde des données pendant le processus de recherche

tout développer | tout réduire

5.1 Comment les données seront-elles stockées et sauvegardées tout au long du projet ?



→ Objectif : Éviter les pertes de données

- Assurer un stockage adapté



6. Partage des données et conservation à long terme

tout développer | tout réduire

6.1 Comment les données seront-elles partagées ? >

6.2 Comment les données seront-elles conservées à long terme ? >

→ Objectif : Clarifier les droits de réutilisation

- Spécifier les modalités de partage

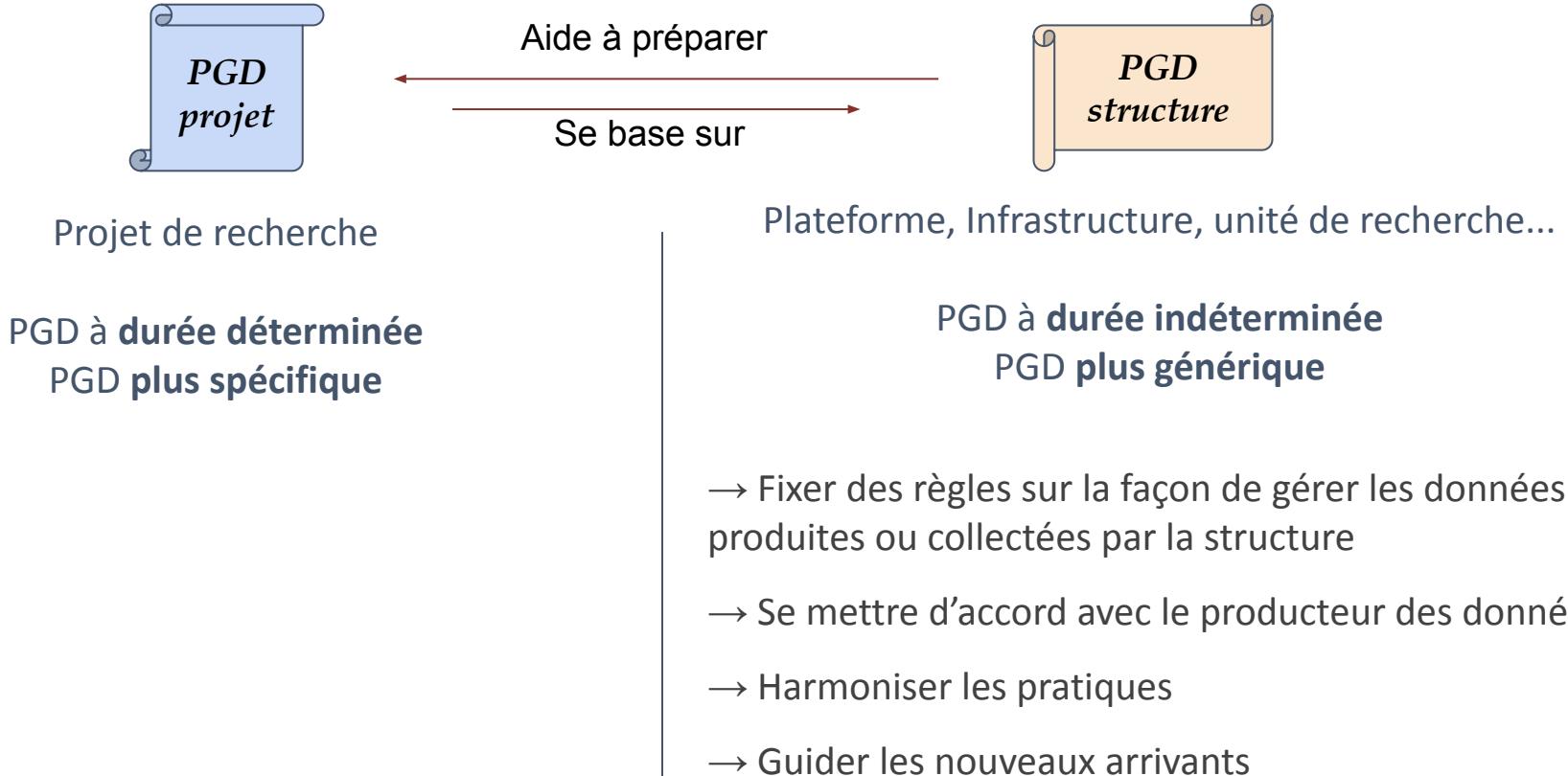


Liste des personnes contribuant à la gestion des produits de recherche au cours d'un projet et leurs rôles. L'attribution d'un rôle à une personne s'effectue dans l'onglet "Rédiger".

Nom	Affiliation	Rôles attribués (Produits de recherche associés)	
Anne-Caroline Delétoille	Institut Pasteur	<ul style="list-style-type: none">Personne contact pour les données (JD2 - PCR)	<input checked="" type="checkbox"/> <input checked="" type="checkbox"/>
Fanny SEBIRE	Institut Pasteur	<ul style="list-style-type: none">Coordinateur du projetPersonne contact pour les données (JD1 - Images)Responsable du plan	<input checked="" type="checkbox"/> <input checked="" type="checkbox"/>

→ Objectif : Établir le rôle de chacun

- Définir les responsabilités





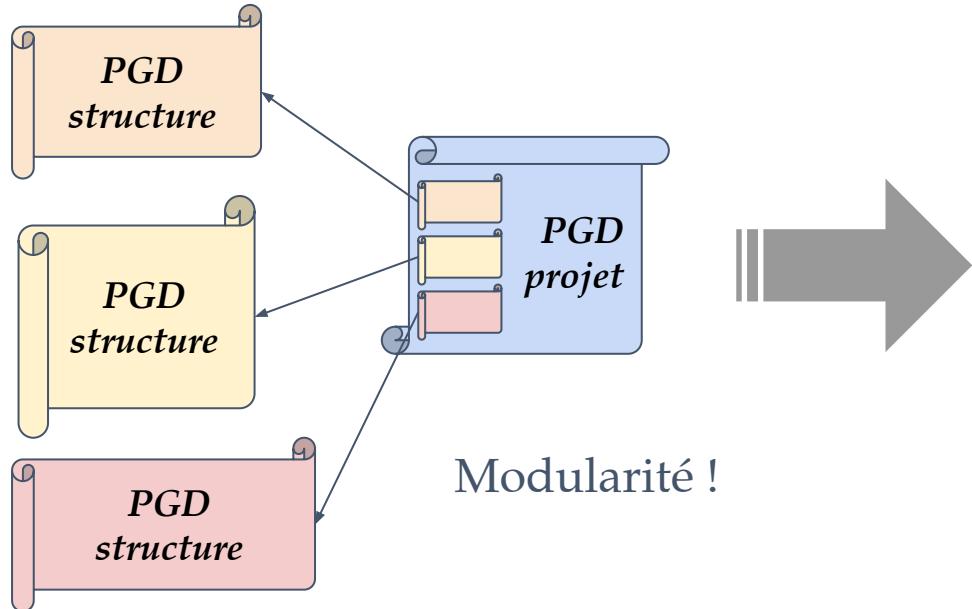
PGD publics sur DMP OPIDoR (projet et structure) : https://dmp.opidor.fr/public_plans

Exemples de PGD de projet :

- [PGD du projet INFRAVEC2](#)
- [Collection de 841 PGD H2020](#)
- [11 PGD primés](#)

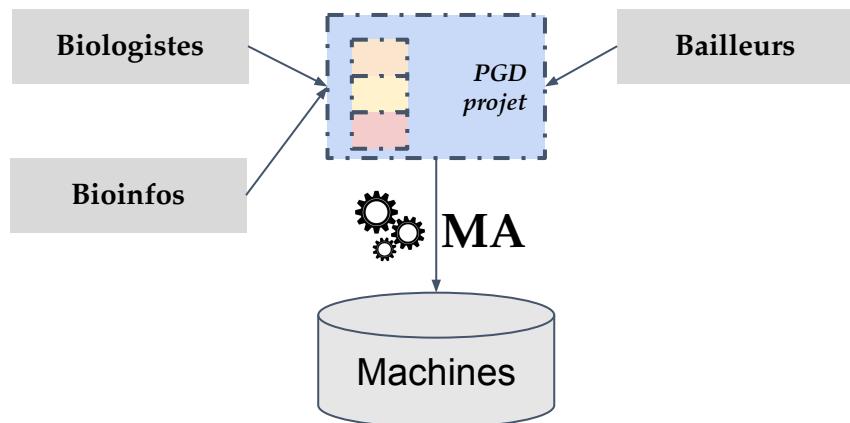
Exemples de PGD de structure :

- [Plan de Gestion des Données du CIRM-BIA](#) :
- [Data management plan of the Plant Bioinformatics Facility](#)



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



Différents outils pour rédiger des PGD



- DMP OPIDoR - solution nationale
- DSW - Data Stewardship Wizard - solution européenne (ELIXIR)
- ARGOS - solution de la Commission Européenne



DS Wizard

20210224_exemple_canevas_IFB_bioimage

Chapters

- I. Préface
- II. Introduction
- III. Informations générales
- IV. Données de la recherche

1.a.2.a.3 Quelles mesures de contrôle de la qualité sont prises pour ce jeu de données ?

Ceci concerne l'acquisition des images. Voir le premier white paper de <https://quarep.org/>, <https://arxiv.org/pdf/2101.09153.pdf>. La liste d'options présentée ici est en anglais : ne devrait-on pas traduire cela pour être consistant/coherente avec le reste de ce document ? (quoique les modes d'acquisition sont bien donnés en anglais...)

- a. Illumination power
- b. Detection system performance
- c. Field of view uniformity, flatness
- d. Chromatic aberrations
- e. Lateral and axial resolution
- f. Image quality
- g. Commentaires

1.a.2.a.2.a.4 Des versions différentes du jeu de données sont-elles créées ?

- Get familiar with the **DMP Lifecycle** and excel in **Open** and **FAIR** RDM planning
- Co-create DMPs and manage workload
- Publish and cite DMPs as living documents
- Configure DMPs to tailored community needs
- Link DMPs to research outputs, EOSC services and the **OpenAIRE Research Graph**