23 mars 2022

Module 1 Séquence 2

Kirsley Chennen

Formation “Science Ouverte & PGD”
Module 1: vers le FAIR

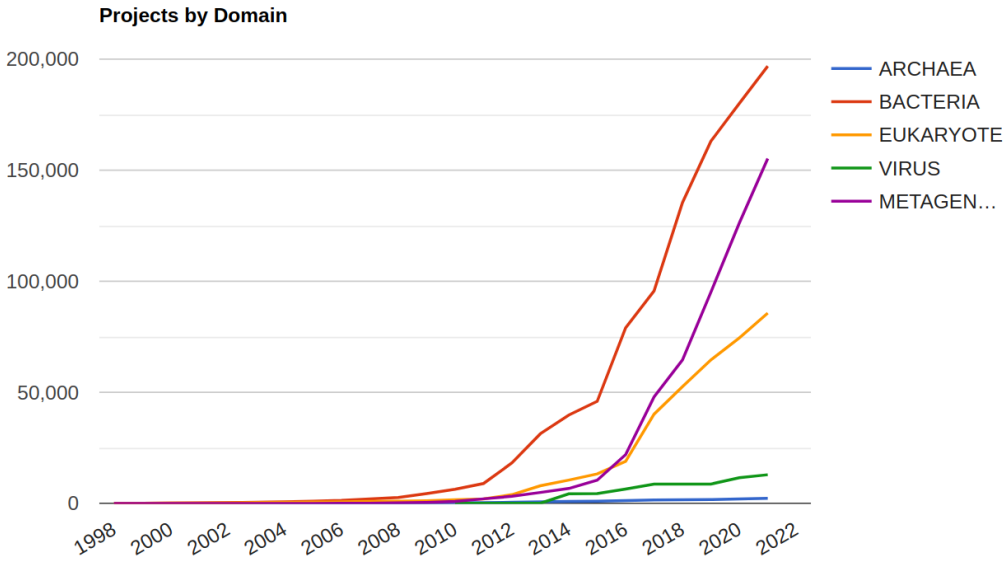
23 mars 2022

Crise de reproductibilité

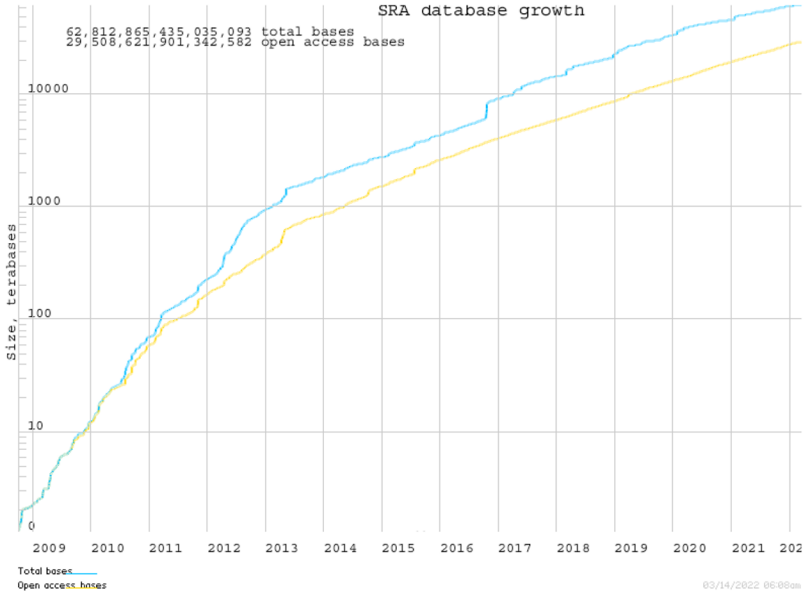
Kirsley Chennen, ICube, Strasbourg
https://orcid.org/0000-0001-9268-6748

Adapté de
Frédéric de Lamotte, INRAe, INRAE Montpellier
https://orcid.org/0000-0001-5102-0632
Thomas Denecker, IFB, Paris
https://orcid.org/0000-0001-5102-0632

Explosion des données en biologie



Total projects in GOLD
(https://gold.jgi.doe.gov/statistics)



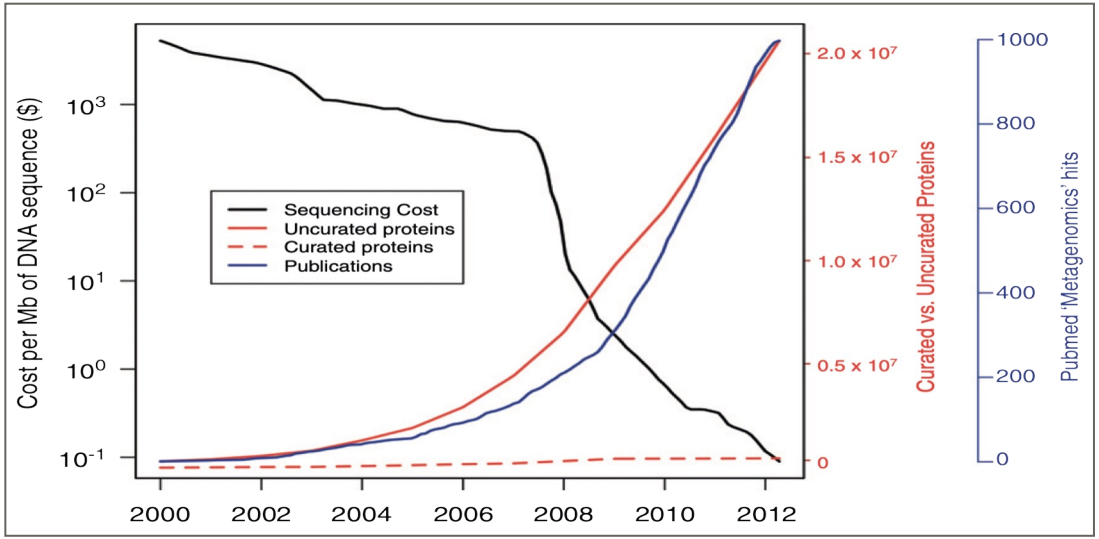
Sequences in SRA database
https://www.ncbi.nlm.nih.gov/sra/docs/sragrowth/

Le défi de la valorisation

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

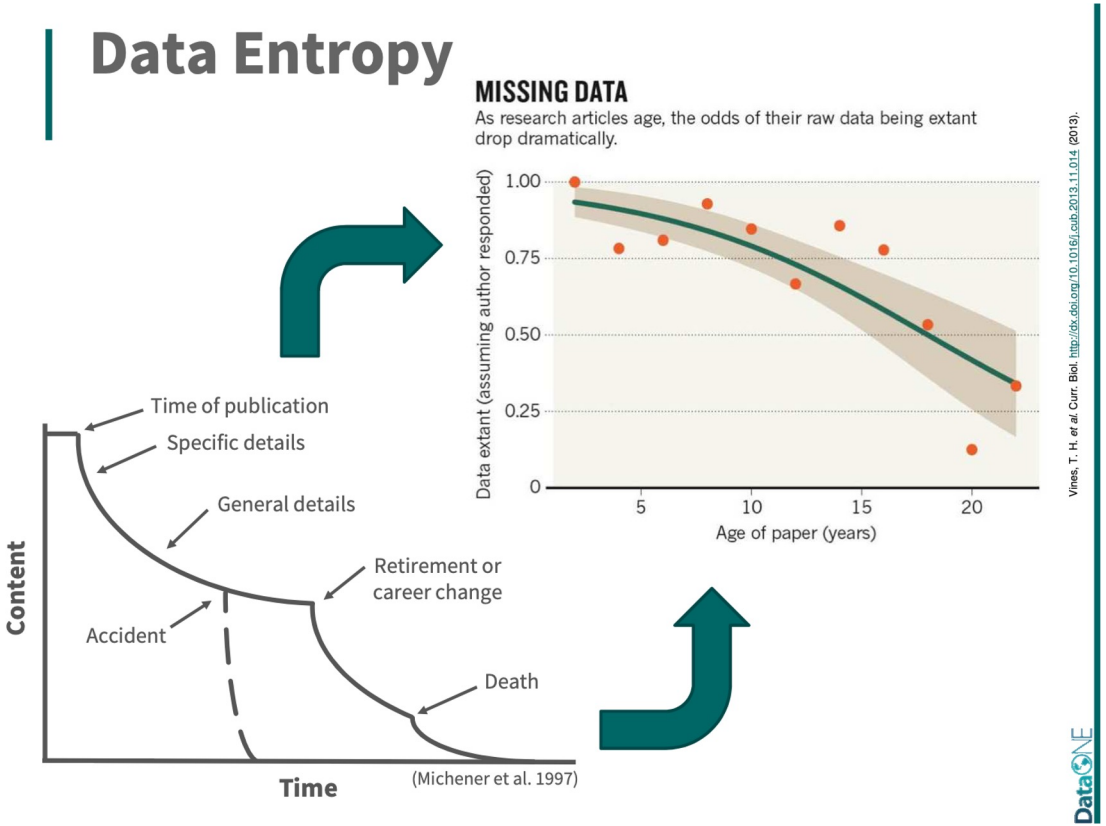
Génome humain :

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

Les ravages du temps



Les défis de la reproductibilité

RESEARCH ARTICLE

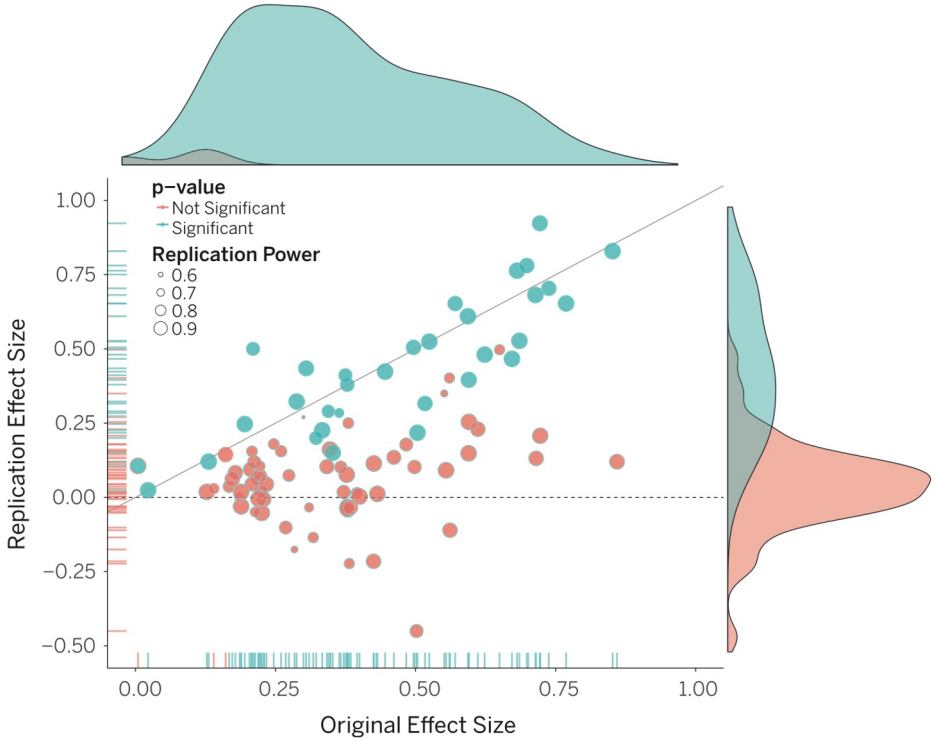
Estimating the reproducibility of psychological science

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

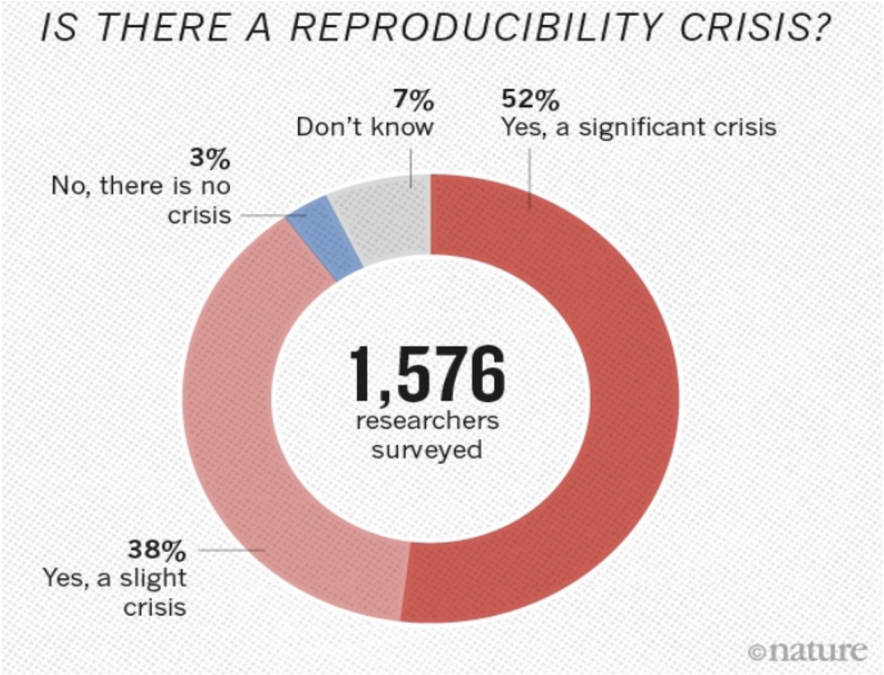
Science 28 Aug 2015:
Vol. 349, Issue 6251, aac4716
DOI: 10.1126/science.aac4716

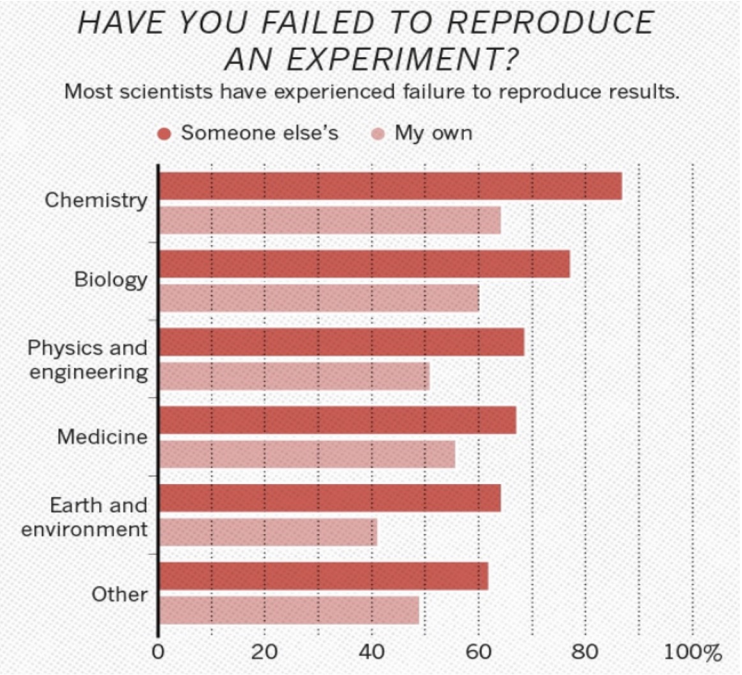
The *Reproducibility project* set out to replicate 100 experiments published in high-impact psychology journals.

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



A reproducibility problem, Biology

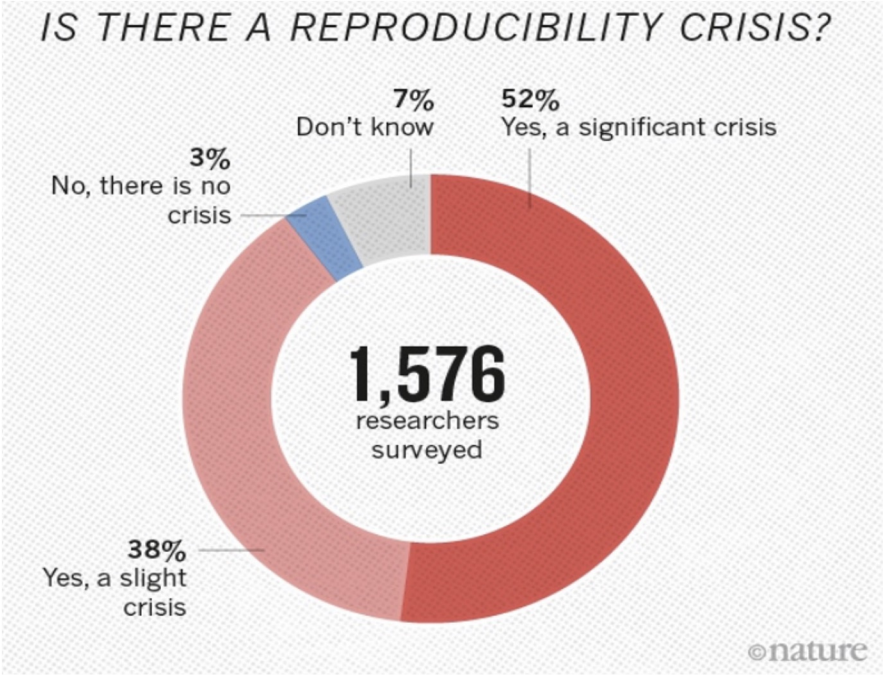


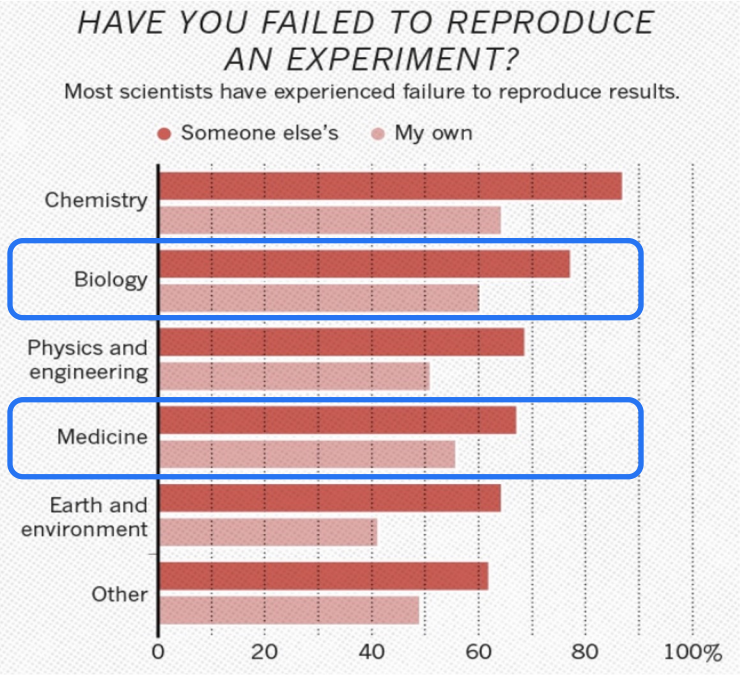


>70% of the analyses in Experimental Biology are **not** reproducible

"1,500 scientists lift the lid on reproducibility". Monya Baker. Nature. 533: 452–454 - 2016

A reproducibility problem, Biology

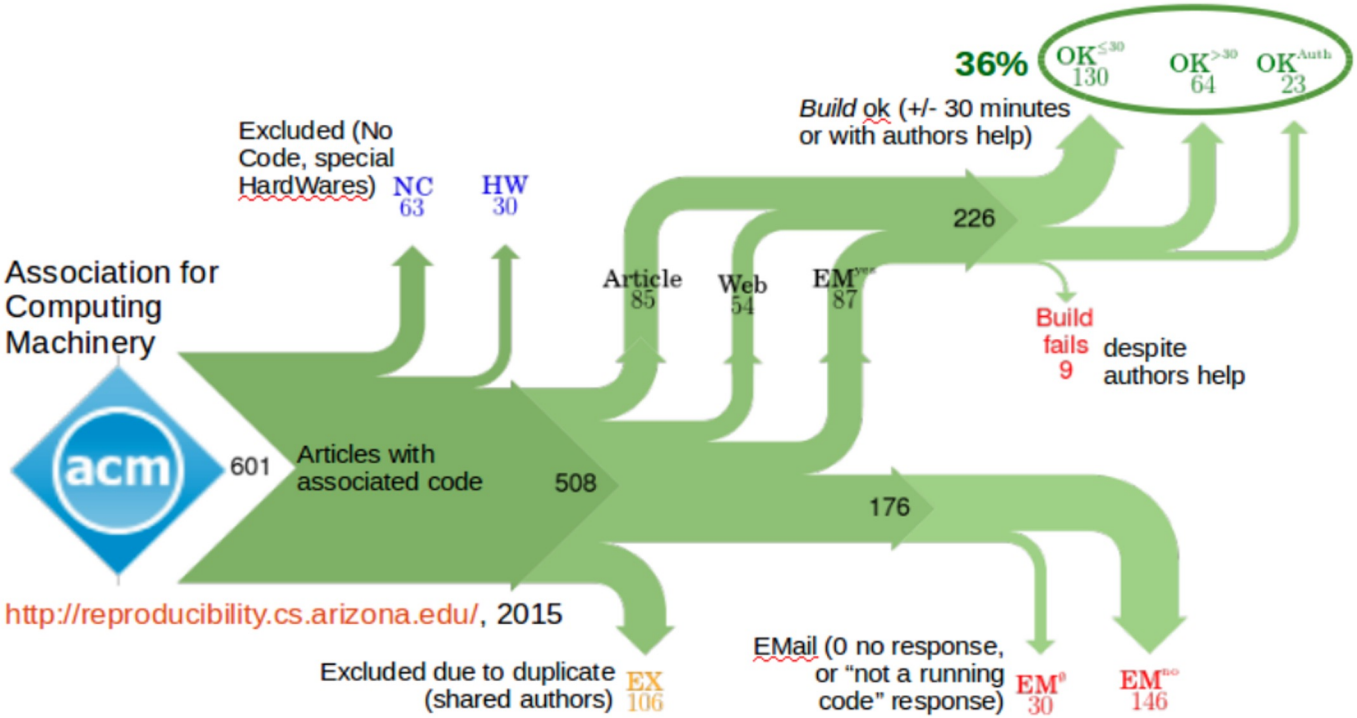




>70% of the analyses in Experimental Biology are **not** reproducible

"1,500 scientists lift the lid on reproducibility". Monya Baker. Nature. 533: 452–454 - 2016

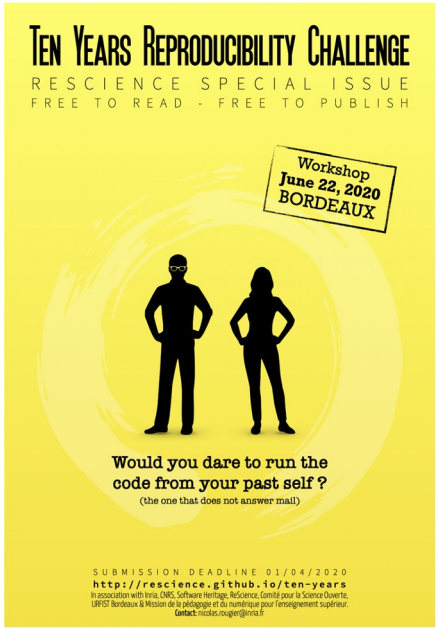
A reproducibility problem, Computer Sciences



64% failure to reuse the software

Collberg et al, University of Arizona TR 14-04, 2015

A reproducibility problem, Bioinformatics



https://rescience.github.io/

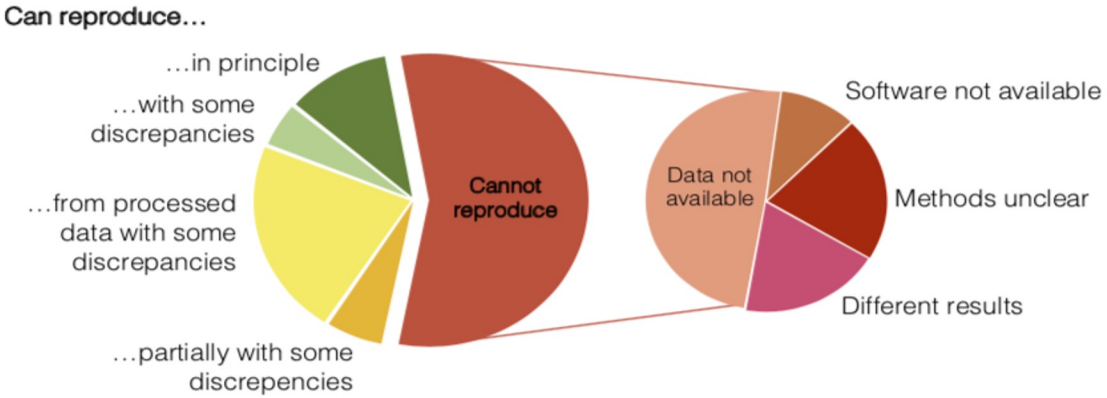
Ten-Year Reproducibility Challenge, Konrad Hinsen Can your 2009 code still run? special issue of ReScience and result comments in Nature

Who's never wanted to take over a protocol, a pipeline, a tool or a dataset without running into it?

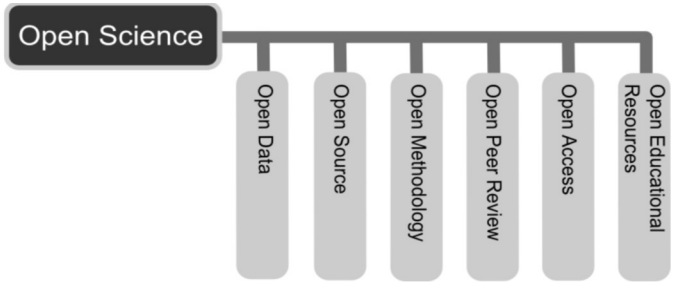
- Obsolete documentation or URL (E404!!!)
- tools: not compatible OS, not availability of dependencies tool update ⇒ codes unusable: python 2 vs. 3, change of function arguments (R)
- inability to reproduce the results of computational analysis: package versions, IDE: stable version of the language different according to the OS (Rstudio)

A reproducibility problem, Bioinformatics

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

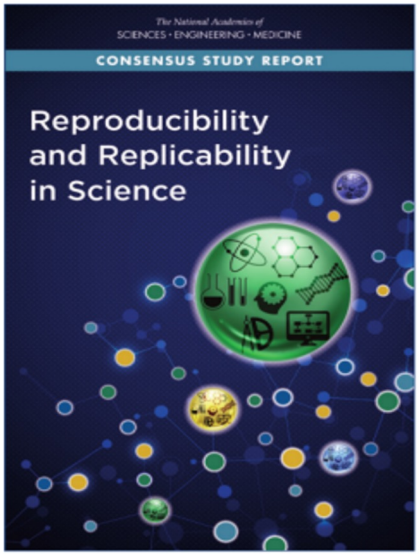


Summary of the efforts to replicate the published analyses.
Adopted from: Ioannidis et al. Repeatability of published microarray gene expression analyses. *Nature Genetics* **41** (2009) doi:10.1038/ng.295



Reproducibility in science

Reproducible research, Repeatability, Replicability, Reproducibility, Replication: overlapping semantics
⇒ a plethora of definitions!



National Academies of Sciences, Engineering, and Medicine (2019).

ACM definition (2016):
Repeatability - Same team, same exp. setup
Replicability - Different team, same exp. setup
Reproducibility - Different team, different exp. Setup

Whitaker's matrix of reproducibility (2017):

		Data: Same	Data: Different
Analysis	Same	Reproducible	Replicable
Analysis	Different	Robust	Generalisable

Exo 1.1: Réutiliser les données? Où est le problème?

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

https://youtu.be/66oNv_DJuPc

Mentimeter code: **9690 6082**
https://www.menti.com/z2mggjobu1

Thérapie de groupe

Quelles expériences avez-vous concernant l’accès et la réutilisation de données de recherche ?

Exo 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: code: 15 63 57 3 / https://www.menti.com/x9zth8t1yt)
   Attention : notez un seul mot à la fois et en français, non composé et sans majuscule.

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

Un autre exemple...

23 mars 2022

Module 1 Séquence 3

Kirsley Chennen Arnaud Kress

23 mars 2022

Vers le FAIR

Kirsley Chennen, ICube, Strasbourg
https://orcid.org/0000-0001-9268-6748

Arnaud Kress, ICube, Strasbourg
https://orcid.org/0000-0002-7616-8876

Frédéric de Lamotte, INRAe, INRAE Montpellier
https://orcid.org/0000-0001-5102-0632

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 sont accessibles à travers un **protocole de communication standard**
- 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**

https://doranum.fr/enjeux-benefices/principes-fair/

Vos données sont-elles FAIR ?

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

- Les données ont un **PID** (Persistent IDentifier ou identifiant pérenne en français)
- 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

https://doranum.fr/enjeux-benefices/principes-fair/

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**

https://doranum.fr/enjeux-benefices/principes-fair/

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**

https://doranum.fr/enjeux-benefices/principes-fair/

Vos données sont-elles FAIR ?

Reusable – Quelle licence?

- Exemple: Creative Commons (https://creativecommons.org/choose/) qui autorise le partage, la réutilisation en échange de la citation de l'auteur

- Il est possible d'ajouter des restrictions sur l'utilisation commerciale ou les conditions de partage:

Attention ! Il y a d'autres licences plus adaptées pour les logiciels

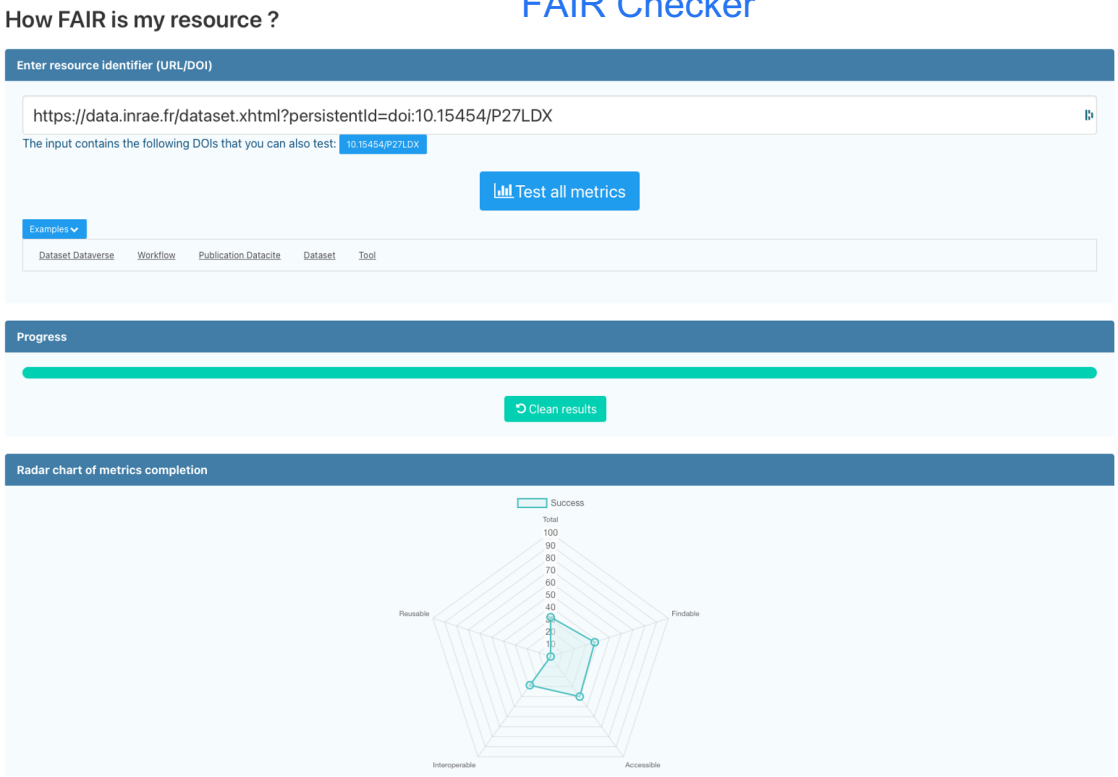
https://choosealicense.com

Exo 1.3: Vos données sont-elles FAIR ?

Findable?

Interoperable?

FAIR Checker



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

"ALL RESEARCH SHOULD AIM TO BE F.A.I.R."

#FIGSHAREFEST

	GOOD	BAD
FINDABLE	ONLINE DATABASE	FILING CABINET IN A BATH IN THE BASEMENT UNDER A LEAKING PIPE
ACCESSABLE	OPEN ACCESS FOR EVERYONE (NO LOGIN)	THE FILING CABINET ALSO IS HOME TO A NEST OF WILD BADGERS
INTEROPERABLE	ALL DATA IS IN OPEN FORMATS	ALL DOCUMENTS ARE PRINTED IN COMIC SANS AND WRITTEN IN ESPERANTO
REUSEABLE	GOOD META DATA AND SECURELY STORED FOR 10 YEARS	THE PAPER EXPLODES IF IT'S READ

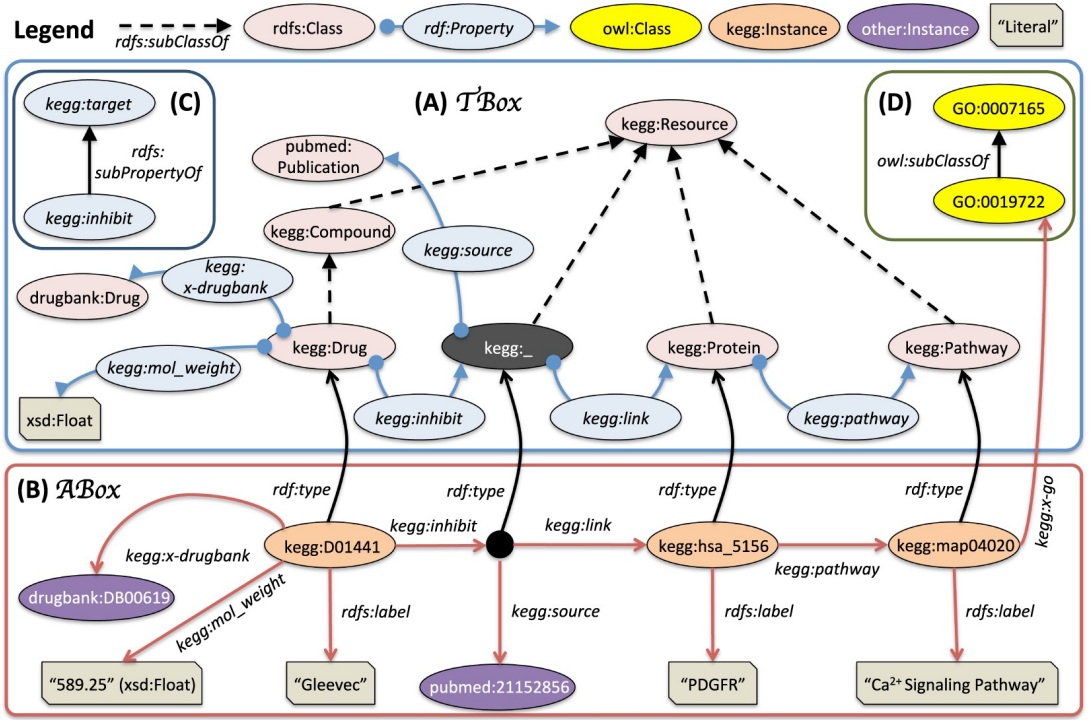
ERRANTSCIENCE.COM

Où voulons-nous aller ?

Vers la base de connaissance

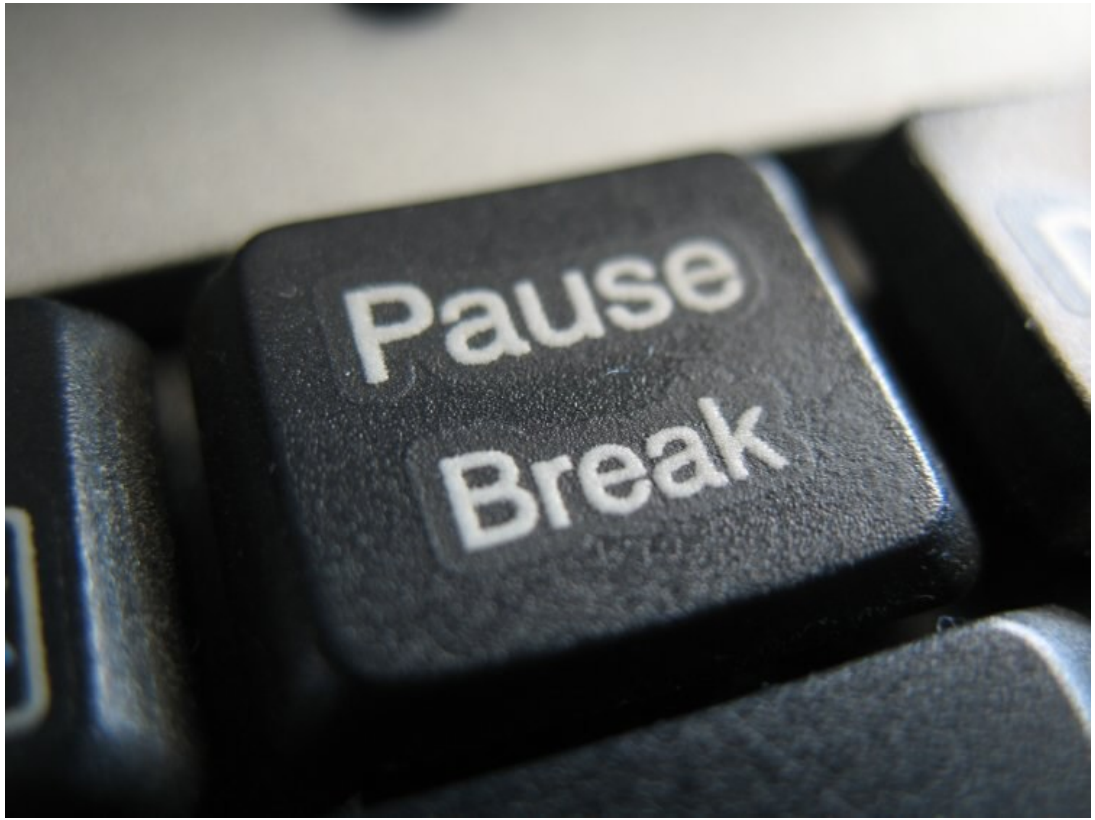
Dans l'idéal, les données sont toutes reliées à des ontologies et des bases documentées

https://www.nextprot.org/proteins/search



Kamdar, M.R., Musen, M.A. An empirical meta-analysis of the life sciences linked open data on the web. *Sci Data* **8,** 24 (2021). https://doi.org/10.1038/s41597-021-00797-y

Pause !



23 mars 2022

Module 1 Séquence 4

Arnaud Kress

Formation “Science Ouverte & PGD”
Module 1: vers le FAIR

23 mars 2022

Cycle de vie des données

Arnaud Kress, ICube, Strasbourg
https://orcid.org/0000-0002-7616-8876

Frédéric de Lamotte, INRAe, INRAE Montpellier
https://orcid.org/0000-0001-5102-0632

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 …

Session post-it!

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

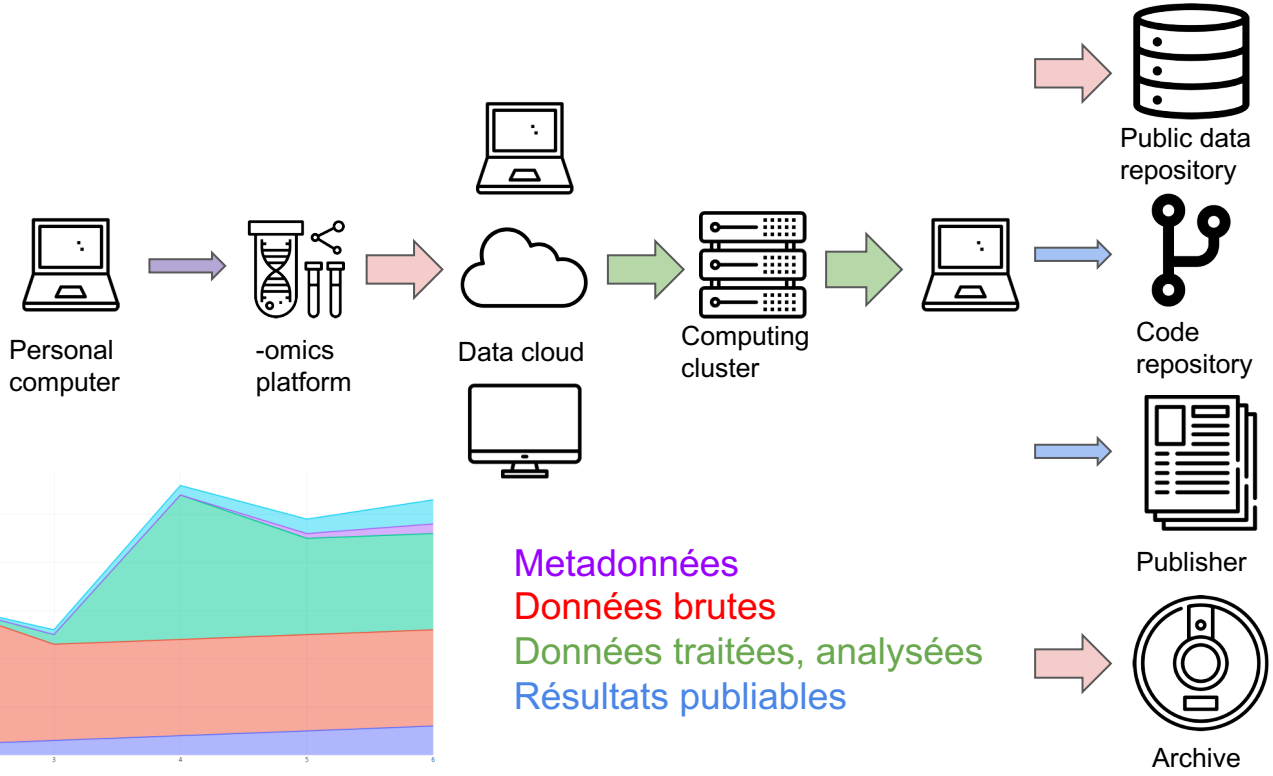
Les étapes

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).

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

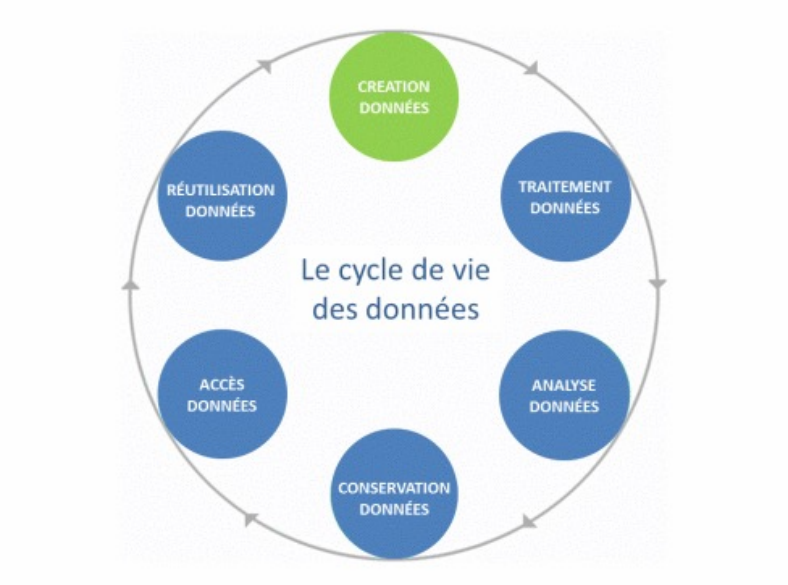
Le parcours des données



Metadonnées
Données brutes
Données traitées, analysées
Résultats publiables

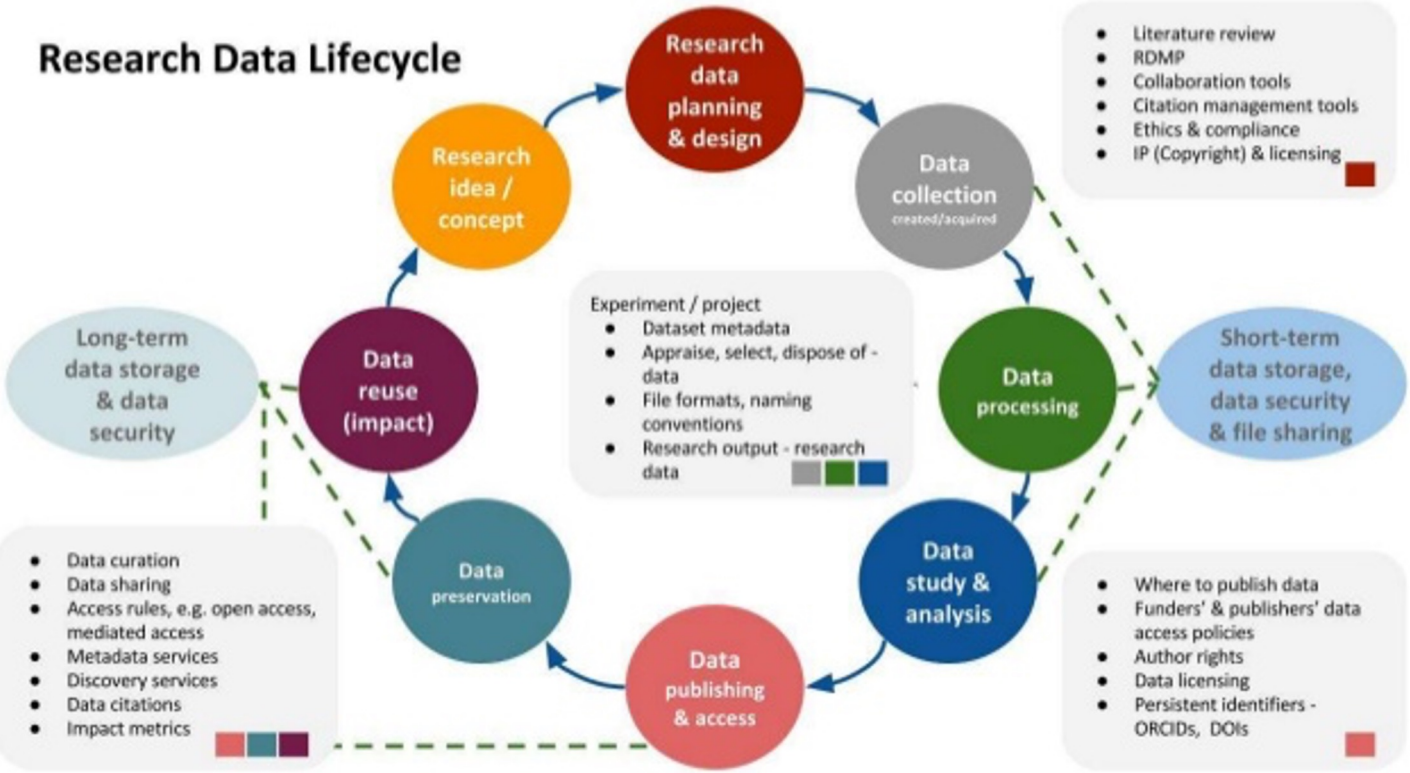
flaticon.com

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