Principes et enjeux de la science ouverte

Gaëlle Leroux

Université Claude-Bernard Lyon 1
Institut des Sciences et Techniques de Réadaptation
Master 1 Santé
UE Insertion professionnelle

Principes et enjeux de la science ouverte

Gaëlle Leroux, PhD
Research engineer
gaelle.leroux@cnrs.fr
https://orcid.org/0000-0003-2729-4945

Soutien Méthodologique aux projets d'imagerie

REMi
Réseau d'Entraide Multicentrique en IRM

Ouvrir la science !
Research data College

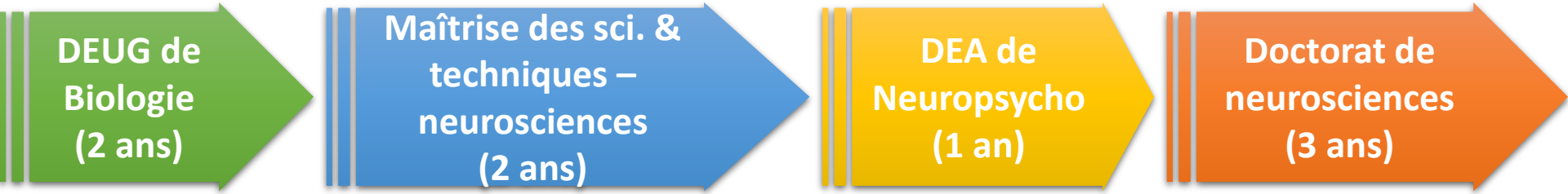
7 piliers de la science ouverte

- publications ouvertes (*open access*)
- données ouvertes (*open data*)
- sources et pratiques ouvertes (*open source & open materials*)
- éducation ouverte (*open educational resources*)
- science citoyenne (*citizen science*)
- évaluation responsable (*assessment*)
- intégrité scientifique (*scientific integrity*)

Source

Mon parcours professionnel

Formation initiale à Caen



Traitement de données en SPECT & MEG de patients épileptiques

Collecte & traitement de données comportementales, en EEG, MEG et IRMf chez des adultes sains

Expériences professionnelles

- **Post-doc** (CDD) **en neurosciences cognitives développementales** (2 ans) à Stockholm
- **Ingénieure de recherche biologiste en analyse de données**
  Univ. Paris Descartes (4 ans), Univ. Bordeaux (3 ans), CNRS (9 ans) Bordeaux puis Lyon

Spécialité

- **Conception et implémentation de protocoles de recherche** (rédaction de protocoles, demandes éthiques, choix logiciels, prog. logicielle de chronométrie, design des stimuli)
- **Acquisition de données** (comportementales, en neuroimagerie non invasive)
- **Analyse des données et valorisation** (gestion de bases de données, pré- & post-traitements, statistiques, publications, diffusion de données)

Ce qui m'a sensibilisé à la science ouverte

Évolution des jeux de données étudiés

2001 6 participants avec 2 modalités d'imagerie + qqes données cliniques
(labo 1) données stockées sur mon ordinateur du labo
1 pers traitant données

2002-07 < 25 participants avec 1-3 modalités d'imagerie + quelques tests comportementaux
(labo 2) données stockées sur CD + 1 serveur du labo
(labo 3) 2 pers traitant données

2008-11 60 participants avec 1 modalité d'imagerie + batterie tests comportementaux
(labo 4) données stockées sur 2 serveurs du labo
2 pers traitant données

2011-18 450 participants avec 4 modalités d'imagerie + batterie de tests comportementaux
(labo 5) + données génétiques
données stockées sur 3 serveurs du labo
6 pers traitant données

Si je devais répliquer les résultats & figures dans les 15 articles publiés…aïe aïe aïe

Pourquoi on parle d'« ouvrir » la science ?

Elle est ***issue d'un mouvement de contestation de la situation de monopoles des grands éditeurs par la promotion de l'open access****.*

La notion de science ouverte s'est développée afin de caractériser
un **double mouvement d'ouverture des résultats de la science**
(publications, données de la recherche lorsque c'est possible),
mais également une nouvelle façon de faire de la recherche,
passant par l'ouverture des processus, des codes, des méthodes ou des protocoles.

Source : guide Couperin

« La science ouverte est un mouvement qui consiste à **ouvrir le processus de recherche à tous types d'acteurs** (partenaires, citoyens) et à **rendre accessibles et réutilisables les produits de la recherche** (publications scientifiques, données, logiciels, etc.) **par la communauté scientifique et la société.**

Partager les résultats de recherche, non seulement entre scientifiques, mais aussi avec la société, est en effet le meilleur moyen de **faire progresser la connaissance** et de **développer des relations de confiance avec les différents acteurs**. » INRAE. https://www.inrae.fr/inrae-engage-science-ouverte

Contours de la science ouverte

La science ouverte n'est pas différente de la science traditionnelle.

Elle signifie simplement que l'on effectue ses recherches de manière plus transparente et plus collaborative.

La science ouverte s'applique à toutes les disciplines de recherche.

Source

Un historique rapide de la science ouverte

1ères initiatives : 1990's

1991 Création de l'archive ouverte **ArXiv**

1992 Création des licences ***Creative Commons***

3 textes fondateurs de l'Open Access au niveau international :

2002 **Initiative de Budapest** sur l'Open Access

2003 **Déclaration de Berlin** sur le libre accès à la connaissance

2003 **Bethesda statement** on Open Access publishing

Source

Un historique rapide de la science ouverte

1ères initiatives : 1990's

Dates clefs en France

2002 **HAL : archives ouvertes des publications scientifiques**

Un historique rapide de la science ouverte

Source

1ères initiatives : 1990's

Dates clefs en France

2002 HAL : archives ouvertes des publications scientifiques

2016 **Loi pour une République numérique** (ou "Loi Lemaire")

Publications

L'article 30 **garantit le droit aux chercheurs de diffuser la version acceptée pour publication** (postprint ou AAM) **de leur article**
au plus tard 6 mois après la première date de publication par l'éditeur **en sciences, techniques et médecine (STM)**
et 12 mois pour les chercheurs **en sciences humaines et sociales (SHS).**

Cette loi s'applique, quel que soit le contrat signé avec l'éditeur de la revue, que la revue soit française ou étrangère.

Un historique rapide de la science ouverte

Source

1ères initiatives : 1990's

Dates clefs en France

2002 HAL : archives ouvertes des publications scientifiques

2016 Loi pour une République numérique (ou "Loi Lemaire")

Données

« II. Dès lors que les données issues d'une activité de recherche financée au moins pour moitié par des dotations de l'Etat, des collectivités territoriales, des établissements publics, des subventions d'agences de financement nationales ou par des fonds de l'Union européenne ne sont pas protégées par un droit spécifique ou une réglementation particulière et qu'elles ont été rendues publiques par le chercheur, l'établissement ou l'organisme de recherche, leur réutilisation est libre. »

« III. L'éditeur d'un écrit scientifique mentionné au I ne peut limiter la réutilisation des données de la recherche rendues publiques dans le cadre de sa publication. »

« IV. Les dispositions du présent article sont d'ordre public et toute clause contraire à celles-ci est réputée non écrite. »

Mise en place du principe d'Open Data par défaut.

Un historique rapide de la science ouverte

Source

1ères initiatives : 1990's

Dates clefs en France

2002 HAL : archives ouvertes des publications scientifiques
2016 Loi pour une République numérique (ou "Loi Lemaire")
2018-21 1er plan national pour la science ouverte ("Plan S")
2021-24 2nd plan national pour la science ouverte

Codes & logiciels

3ème axe :
« ouvrir et promouvoir les codes sources produits par la recherche »


Reproductibilité des résultats
Intégrité scientifique
Stimulation de l'innovation
Enjeux de la Science Ouverte
Science citoyenne
Facilité d'accès au savoir pour tous
Promotion du travail collaboratif
Et en + :
meilleur taux de citation pour les articles et données de la recherche librement accessibles

Source

Crise de la reproductibilité des résultats

RESEARCH ARTICLE

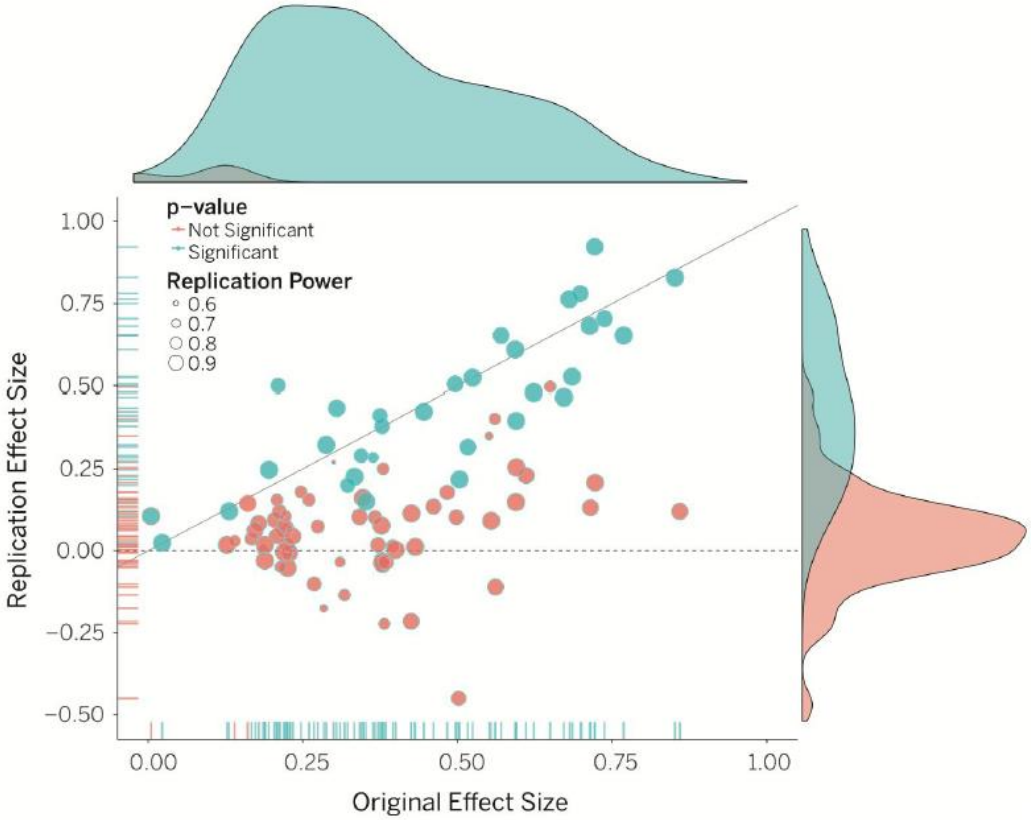
Estimating the reproducibility of psychological science

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

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

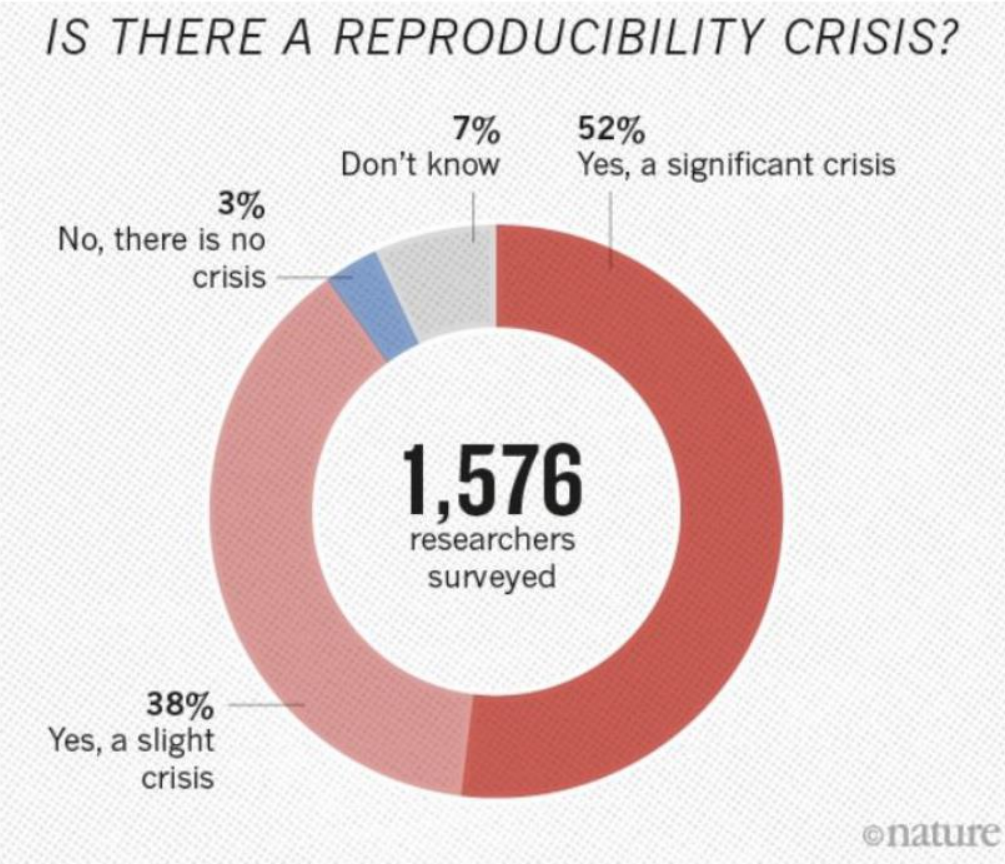
The *Reproducibility project* set out to replicate 100 experiments published in high-impact psychology journals.

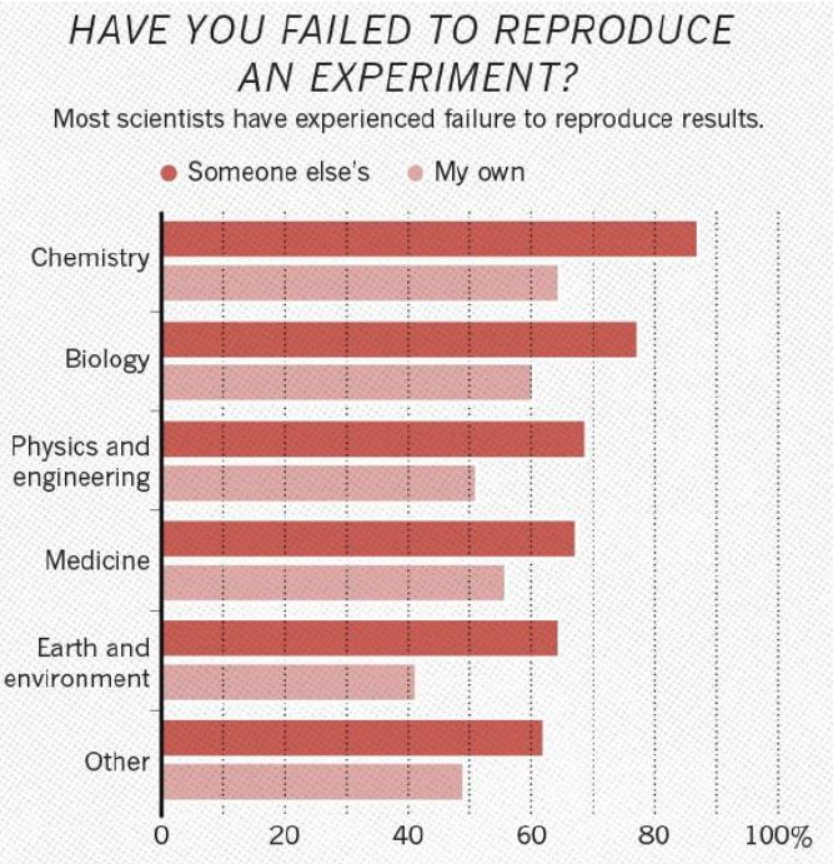
About one-half to two-thirds of the original findings could not be observed in the replication study.



Extrait de la formation IFB “Science ouverte et PGD”. Nov. 2021

Crise de la reproductibilité des résultats





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

Source

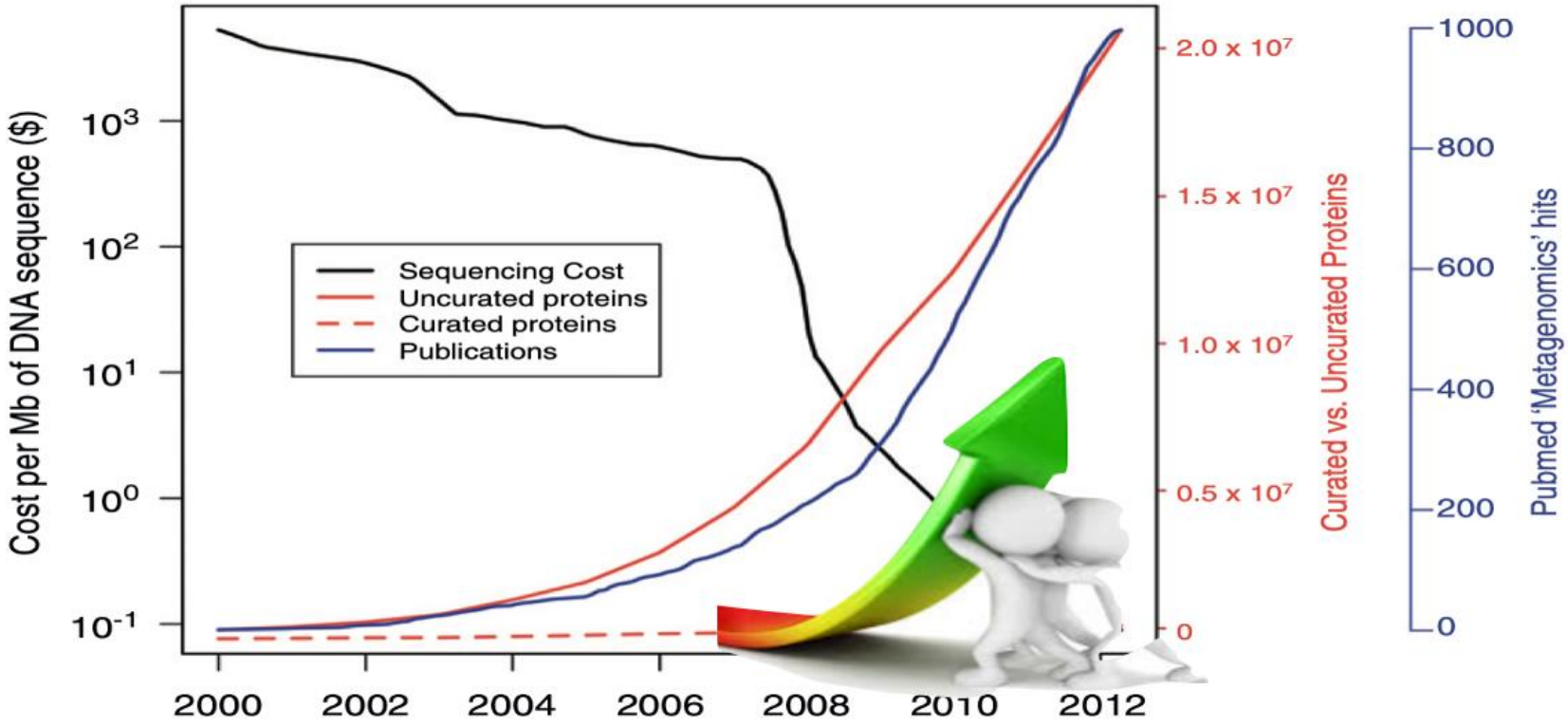
Le déluge de données en science

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** $



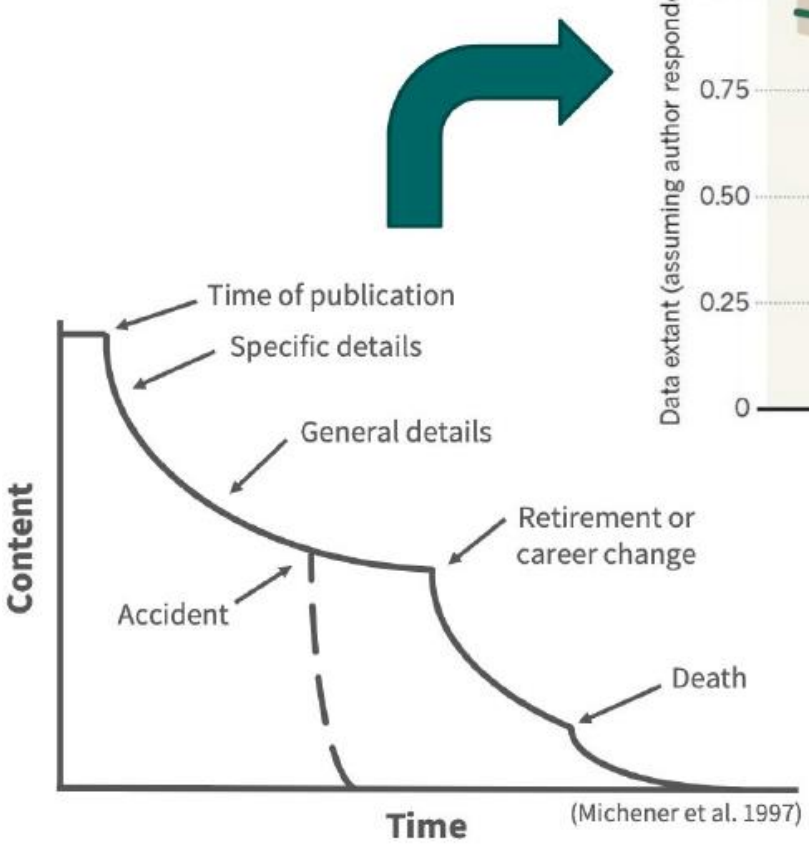
Par conséquent :

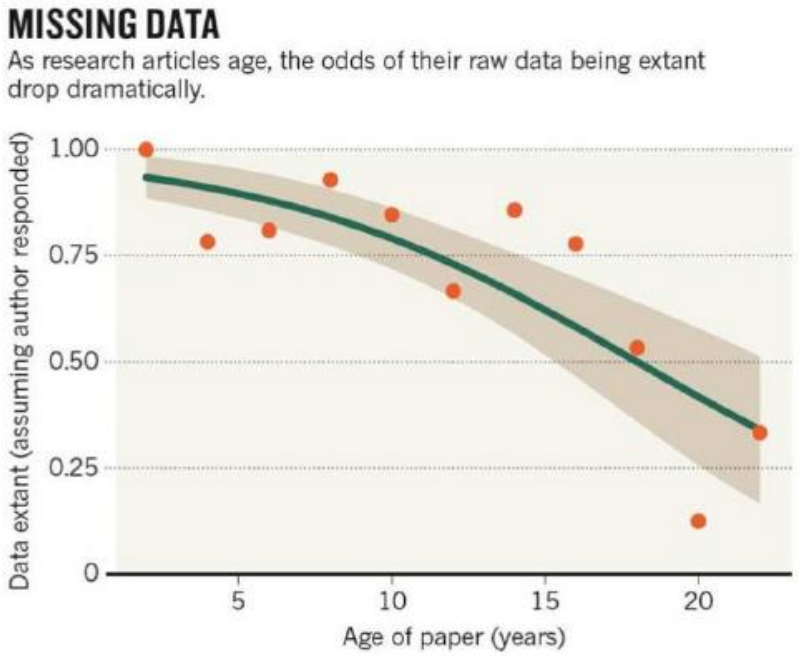
- La quantité de données à stocker et analyser explose
- Le rendement d'analyse chute

Source

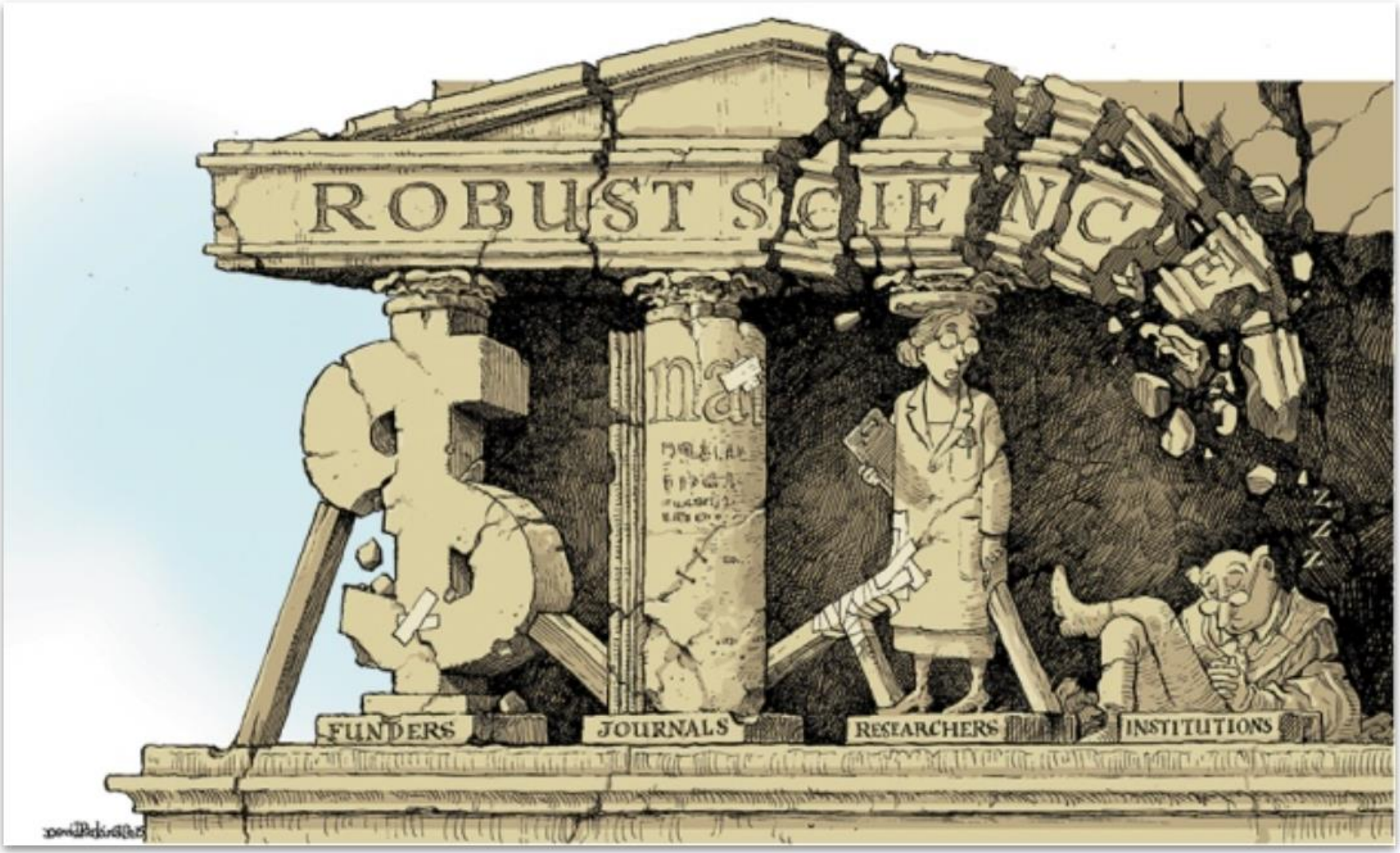
Les ravages du temps

Data Entropy





Source



Nature, September 1, 2015.
Reproduced by courtesy of Nature Publishing Group.

Source

Pas si simple pour autant

Reproducibility is not black an white, but rather a continuum

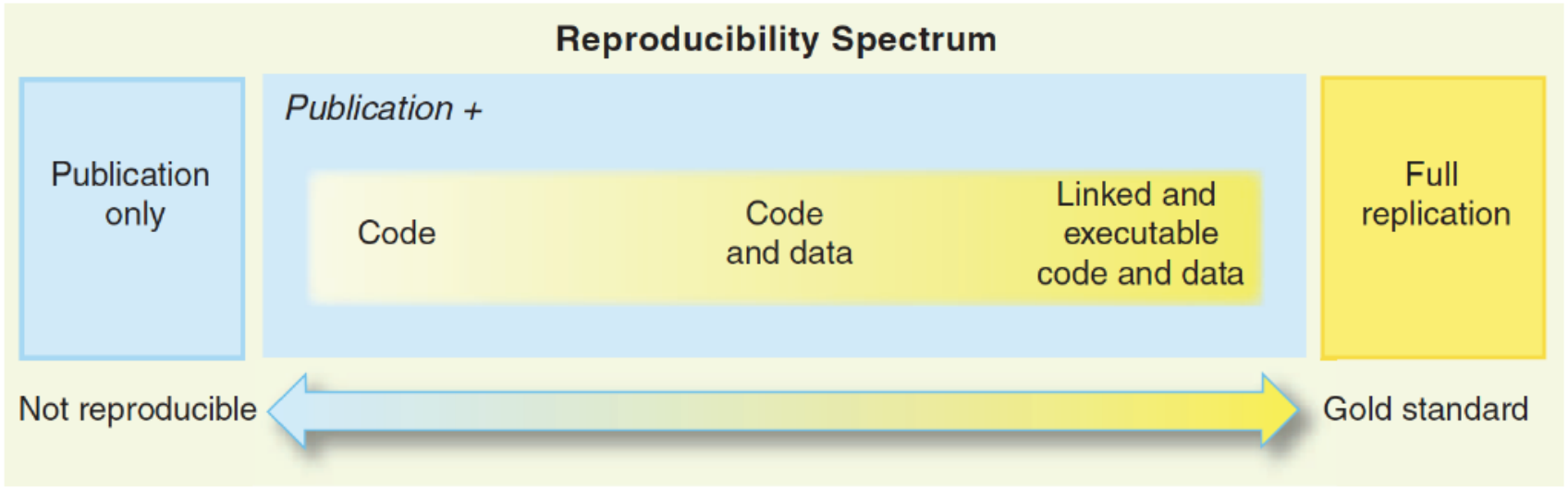


Fig. 1. The spectrum of reproducibility.

Peng, Science, 2011

Source

Pourquoi une si faible reproductibilité des résultats ?

- **Absence d'une "culture de la reproductibilité" profondément enracinée**
- **Manque de formation aux méthodes de recherche robustes**
- **Pratiques de recherche**
  - Mauvaise organisation
  - Mauvaise documentation
  - Mauvaise automatisation
  - Mauvaise diffusion
  - Données ne sont pas disponibles

Source

Comment l’organisation aide à la reproductibilité

Principes d'organisation

1. Mettre chaque projet dans son propre répertoire de projet
2. Garder ensemble ce qui doit l'être
3. Gardez les données source
4. Rendez les noms de fichiers lisibles par des humains
5. Rendez les noms de fichiers lisibles par des machines
6. Rendre les noms de fichiers faciles à trier
7. Sauvegardez vos données

Source

Comment l'organisation aide à la reproductibilité

Ziemann *et al. Genome Biology* (2016) 17:177
DOI 10.1186/s13059-016-1044-7

Genome Biology

COMMENT Open Access

Gene name errors are widespread in the scientific literature

Mark Ziemann[1], Yotam Eren[1,2] and Assam El-Osta[1,3*]

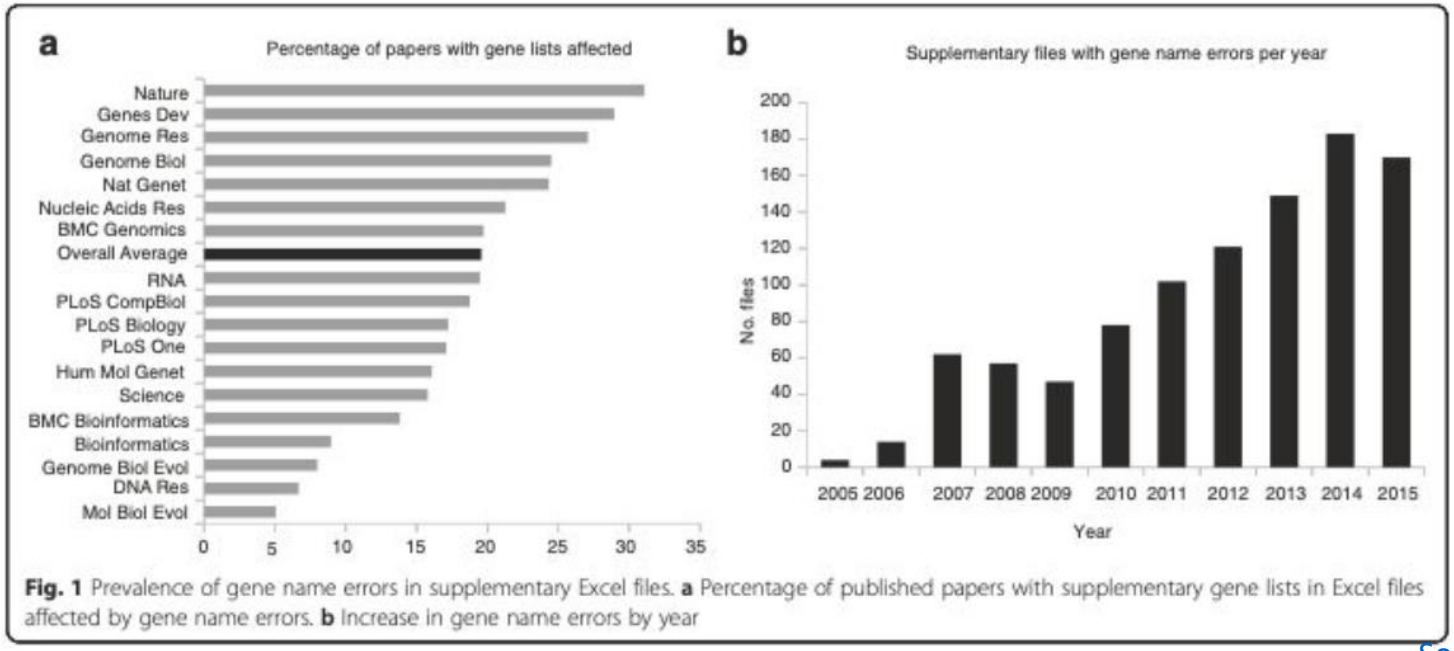


Fig. 1 Prevalence of gene name errors in supplementary Excel files. **a** Percentage of published papers with supplementary gene lists in Excel files affected by gene name errors. **b** Increase in gene name errors by year

Source

Comment l'organisation aide à la reproductibilité

DONNER UN NOM BREF ET EXPLICITE et …

Pas d'espace Ni de caractères spéciaux (& / + > : ? % …)

Règles dénomination fichiers ✗

ReglesDenominationFichiers ✓

Dates au **format AAAAMMJJ** (année, mois, jour)

20150405_CR 20160310_CR 20160515_CR

Versionnez

Convention_V01 Convention_V02 Convention_VF

Rangez

Reunion

20150407_CR

20150407_Minutes

20150407_OJ

Et documentez vos règles !

REGLES DE NOMMAGE DES FICHIERS

EGE-10-Sec7.2.2a-v0.7 | **Domaine:** Systèmes Information

Page: 1/13

REPUBLIQUE ET CANTON DE GENEVE
Collège spécialisé des systèmes d'information

DIRECTIVE TRANSVERSALE

REGLES DE NOMMAGE DES FICHIERS	
EGE-10-Sec7.2.2a-v0.7	**Domaine** : Systèmes Information
Date : 26.11.2012	**Entrée en vigueur** : Immédiate
Rédacteur(s): *Groupe Records management-archives définitives (RM-Archdéf)*	**Direction/Service transversal(e):** *CSSI*
Responsable(s) de la mise en œuvre: Archivistes de département et d'institution **Date: 21.11.2012**	**Approbateur :** Collège spécialisé Systèmes d'Information **Date: 21.11.2012 /mise à jour de l'annexe : décembre 2015**

Source

Comment l'organisation aide à la reproductibilité

Type	Format conseillé	Format non conseillé
Document texte	PDF, TXT, ODT	MS Word, RTF
Feuille de calcul	ODS, CSV	MS Excel, PDF, OOXML
Base de données	SQL, SIARD, DB tables (.CSV)	MS Access, dBase (.dbf), HDF5
Données statistiques	SPSS Portable, STATA, XML, CSV, TXT	SAS et R
Images	JPEG, TIFF, PNG	DICOM
Audio	BWF, MXF, Matroska (.mka), FLAC, OPUS	WAVE, MP3, AAC, AIFF, OGG
Video	MXF, MKV	MPEG-4, MPEG-2, AVI, QuickTime (.mov, .qt)
Information géographique	GML, MIF/MID	ESRI Shapefiles, MapInfo, KML
Images géoréférencées	GeoTIFF (.tif, .tiff)	TIFF World File
Raster	ASCII GRID (.asc, .txt)	ESRI GRID

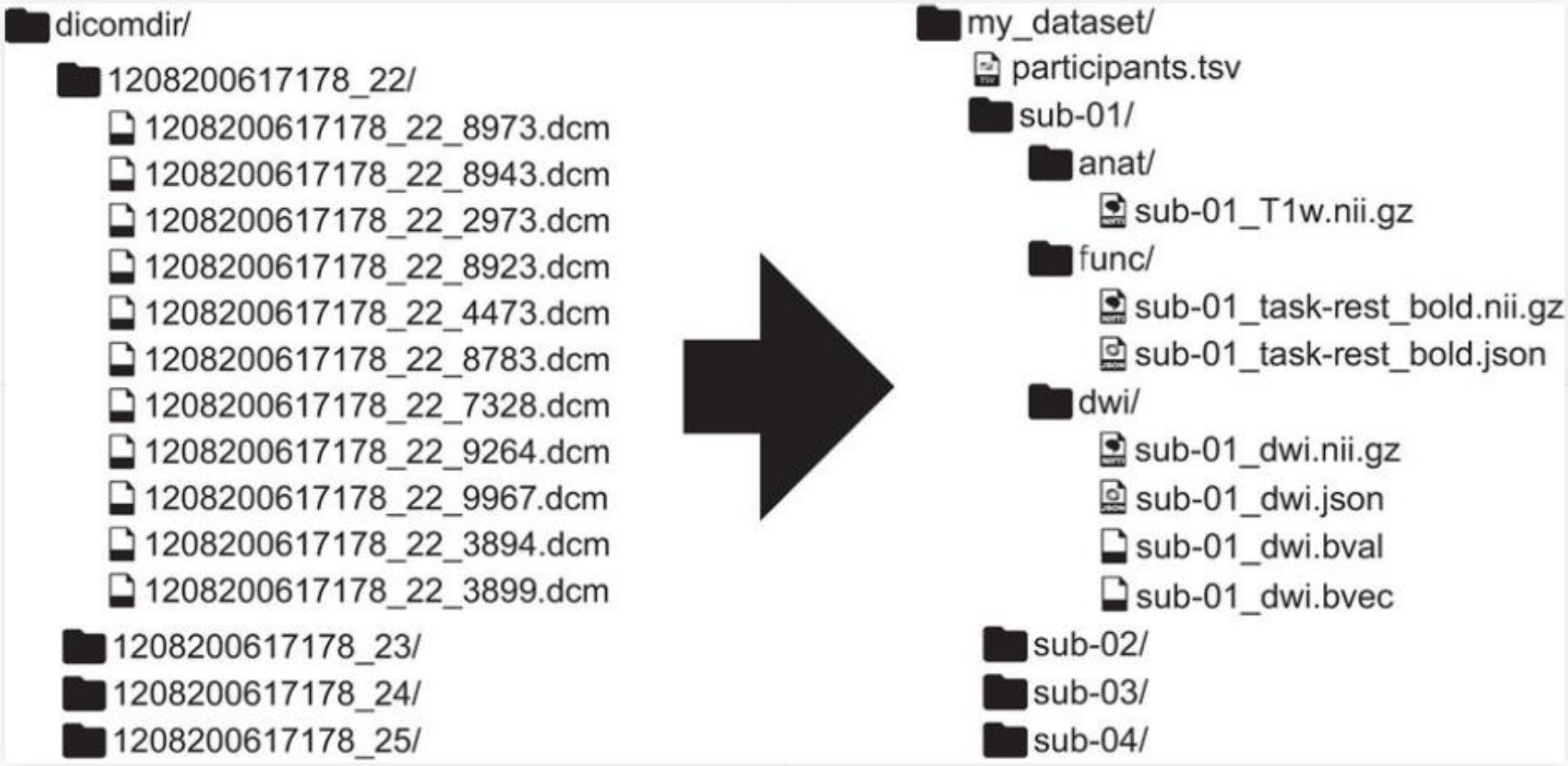
https://facile.cines.fr/ service de validation des formats

En pratique, on peut souvent travailler avec un format fermé populaire et le **convertir** en format ouvert. **Mais il faut vérifier si la conversion altère les informations, et prendre des mesures de compensation si nécessaire.**
Ex : la conversion XLSX -> CSV perd les mises en forme.

Source

Comment l'organisation aide à la reproductibilité

The Brain Imaging Data Structure (BIDS)



Source & further reading: Gorgolewski et al. Sci Data, 2016, http://bids.neuroimaging.io/

Source

Comment l'organisation aide à la reproductibilité

Stocker et sécuriser : quels compromis ?

Comparatif de systèmes de stockage des données

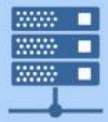
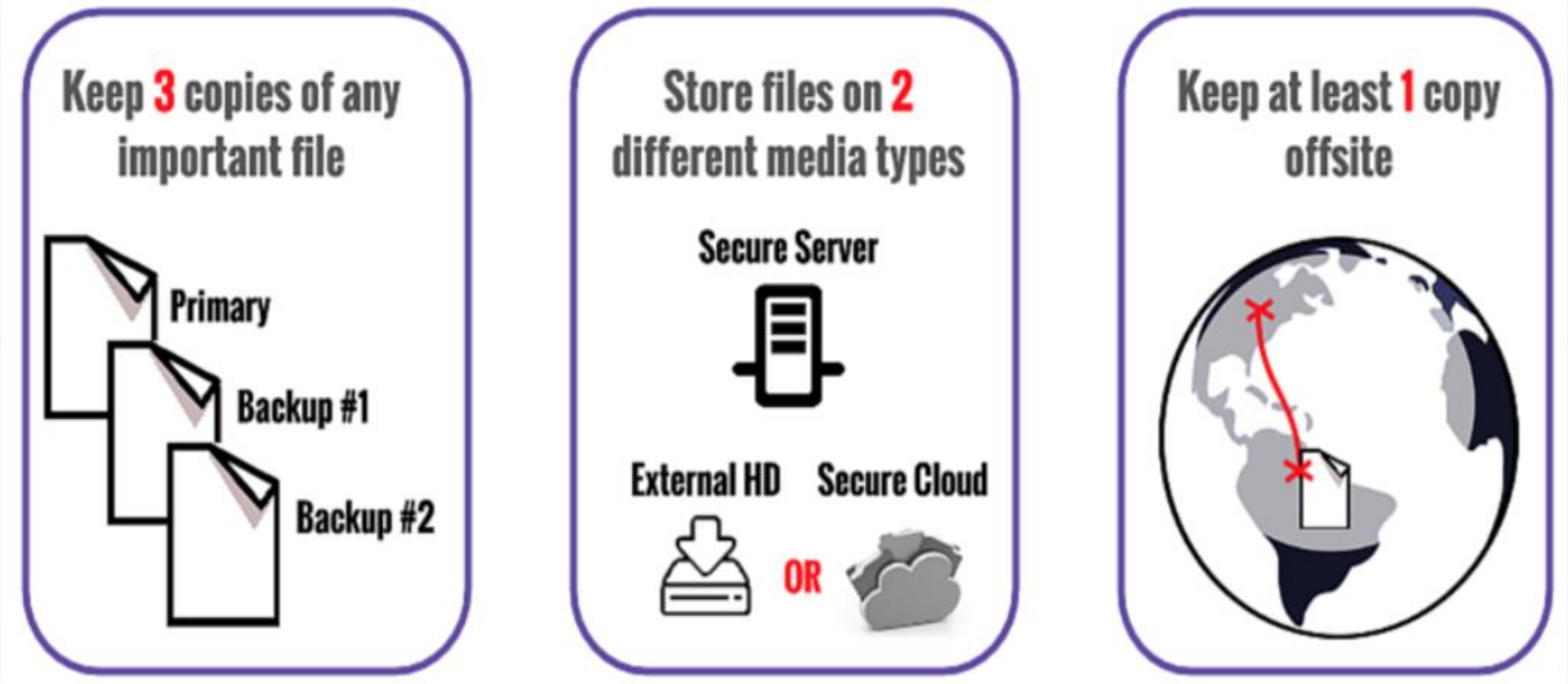
Support de stockage	Sécurité	Accès	Coût	Remarque d'utilisation
 Ordinateur professionnel	★★☆☆ Sujet au piratage informatique, aux détériorations et pannes	★☆☆☆ Pas adapté au partage, nécessite l'utilisation d'un support externe ou d'Internet (mail, cloud…)	★★★★ Pas de coût supplémentaire ou coût peu important	- Pour un stockage temporaire - Nécessité de crypter les données confidentielles et sensibles
 Support externe	★☆☆☆ - Sujet au vol, à la perte du support - Durée de vie limitée (dégradation du matériel)	★★★☆ Facilement transportable, il permet de transférer les données vers un autre ordinateur	★★★★ Pas de coût supplémentaire ou coût peu important	- Pour un stockage temporaire - Nécessité de crypter ou de sécuriser physiquement les données confidentielles et sensibles
 Serveur institutionnel	★★★★ Stockage fiable, durable et sécurisé (contre le vol, le piratage, les incendies…)	★★☆☆ La connexion au serveur institutionnel ne facilite pas le travail avec des personnes extérieures	★★☆☆ Coût assez important mais pas forcément répercuté sur l'usager	- Pour un stockage plus pérenne - Adapté pour le stockage de données sensibles et des versions « stables » de vos données - Toutes les institutions ne proposent pas ce service
 Serveur Cloud	★★☆☆ On ne sait pas vraiment où sont stockées les données, ni ce qu'elles deviennent	★★★★ Permet un travail synchronisé avec toutes les personnes ayant été autorisées au partage	★★☆☆ Payant à partir d'une certaine limite de stockage	- Pour un partage avec des personnes externes à l'institution - Ne pas y mettre de données sensibles ou confidentielles - Pas de contrôle sur la procédure de sauvegarde des données

Tableau tiré de http://doranum.fr/le-stockage-des-donnees/

Source

Comment l'organisation aide à la reproductibilité



Source: NYU libraries

Source

Comment la documentation aide à la reproductibilité

Principes de documentation :

8. Créez une vue d'ensemble de votre projet
9. Rendre les dépendances logicielles explicites
10. Décrire les variables dans un livre de codes
11. Documentez votre code et son utilisation
12. Donner des noms significatifs aux fonctions et aux variables
13. Gardez la trace de la façon dont chaque résultat a été produit

Source

Comment la documentation aide à la reproductibilité

Variable Name	Description	Type	Values or Characteristics
patientid	Patient ID	Numeric	Integers
sex	Patient Gender	Numeric	1=Female, 2=Male
proceduredate	Date of Procedure	Date, M/D/YYYY	
dob	Date of Birth	Date, M/D/YYYY	
patientrace	Patient Race	Numeric	0=Not Latino, 1=Latino,
tx	Treatment Group	Numeric	1-Tx 1, 2=Tx 2, 3=Tx 3
dischargeppt	PPt at Discharge	Numeric	2 decimal places
notes	Notes	Character	
location	Location	Numeric	1=local, 2=regional, 3=distant
patientdied	Patient is Dead	Numeric	0=No, 1=Yes
deathdate	Date of Death	Date, M/D/YYYY	
clinicaloutcome	Clinical Outcome	Character	
outcomecomments	Outcome Comments	Character	

Source: http://domstat.med.ucla.edu/pages/codebook2

Further reading: Codebook cookbook - A guide to writing a good codebook for data analysis projects in medicine, Creating a codebook - Document, Discover, and Interoperate alliance

Tools: codebook generator for R: memisc

Source

Comment la documentation aide à la reproductibilité

Comment l'automatisation aide à la reproductibilité

Principes d'automatisation :

14. Éviter les étapes de manipulation manuelle des données
15. Modulariser le code plutôt que de copier et coller.
16. Réutiliser le code plutôt que de le réécrire.
17. Ne commentez pas et ne décommentez pas les sections de code pour contrôler le comportement du programme
18. Ajouter des assertions aux programmes pour vérifier leur fonctionnement
19. Enregistrer la graine du générateur de nombres aléatoires
20. Enregistrer tous les résultats intermédiaires

Comment la diffusion aide à la reproductibilité

Principes de diffusion :

21. Stocker les données dans des **formats ouverts**
22. Traiter les données avec des **logiciels ouverts**
23. Éliminer les chemins codés en dur dans votre code
24. Relier les déclarations textuelles aux résultats sous-jacents
25. Soumettez le code et les données à un dépôt émettant des DOI et faites-y référence dans votre article.
26. Rendre explicite la **licence du code et des données**

Les principes FAIR

Source

Des données/logiciels peuvent être « FAIR » sans être librement accessibles.

Suivre les principes FAIR permet de s'assurer que ses données sont réutilisables, qu'elles soient partagées ou non.

Facile à trouver :

- Dans un **entrepôt**
- **Identifiant** unique et pérenne (PID)
- **Métadonnées** riches

Accessible :

- Définir les conditions d'accès aux données
  - rendre les données accessibles librement
  - sinon : rendre accessibles les métadonnées pour signaler l'existence des données

Interopérable :

- **Formats ouverts**
- **Mettre à disposition le code source** du logiciel nécessaire pour lire, traiter, analyser les données
- **Privilégier les standards de métadonnées**
- **Indiquer des liens vers d'autres ressources** (autres données, publication…)

Réutilisable :

- **Licence** de diffusion aux données
- **Documentation** pour décrire les données de façon détaillée, les contextualiser, les rendre compréhensibles…


Reproductibilité des résultats
Intégrité scientifique
Stimulation de l'innovation
Enjeux de la Science Ouverte
Science citoyenne
Facilité d'accès au savoir pour tous
Promotion du travail collaboratif
Et en + :
meilleur taux de citation pour les articles et données de la recherche librement accessibles

Source

Science ouverte et intégrité scientifique

Principes fondamentaux de l'intégrité scientifique :

- **Fiabilité**, autrement dit garantir la qualité de la recherche, qui transparaît dans la conception, la méthodologie, l'analyse et l'utilisation des ressources
- **Honnêteté**, autrement dit élaborer, entreprendre, évaluer, déclarer et faire connaître la recherche d'une manière transparente, juste, complète et objective
- **Respect** envers les collègues, les participants à la recherche, la société, les écosystèmes, l'héritage culturel et l'environnement
- **Responsabilité** [partagée entre les acteurs de la recherche] **assumée pour les activités de recherche**, de l'idée à la publication, leur gestion et leur organisation, pour la formation, la supervision et le mentorat, et pour les implications plus générales de la recherche.

Code de conduite européen pour l'intégrité en recherche, 2018, page 4

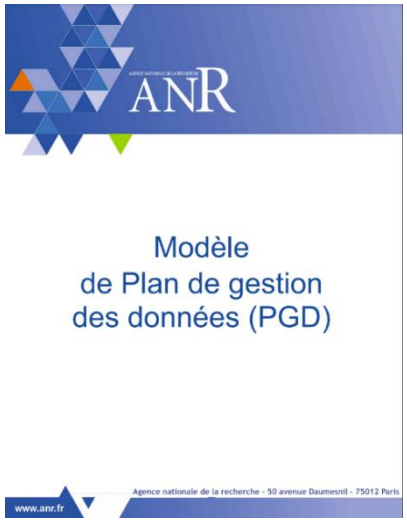
Source


Reproductibilité des résultats
Intégrité scientifique
Stimulation de l'innovation
Enjeux de la Science Ouverte
Science citoyenne
Facilité d'accès au savoir pour tous
Promotion du travail collaboratif
Et en + :
meilleur taux de citation pour les articles et données de la recherche librement accessibles

Source

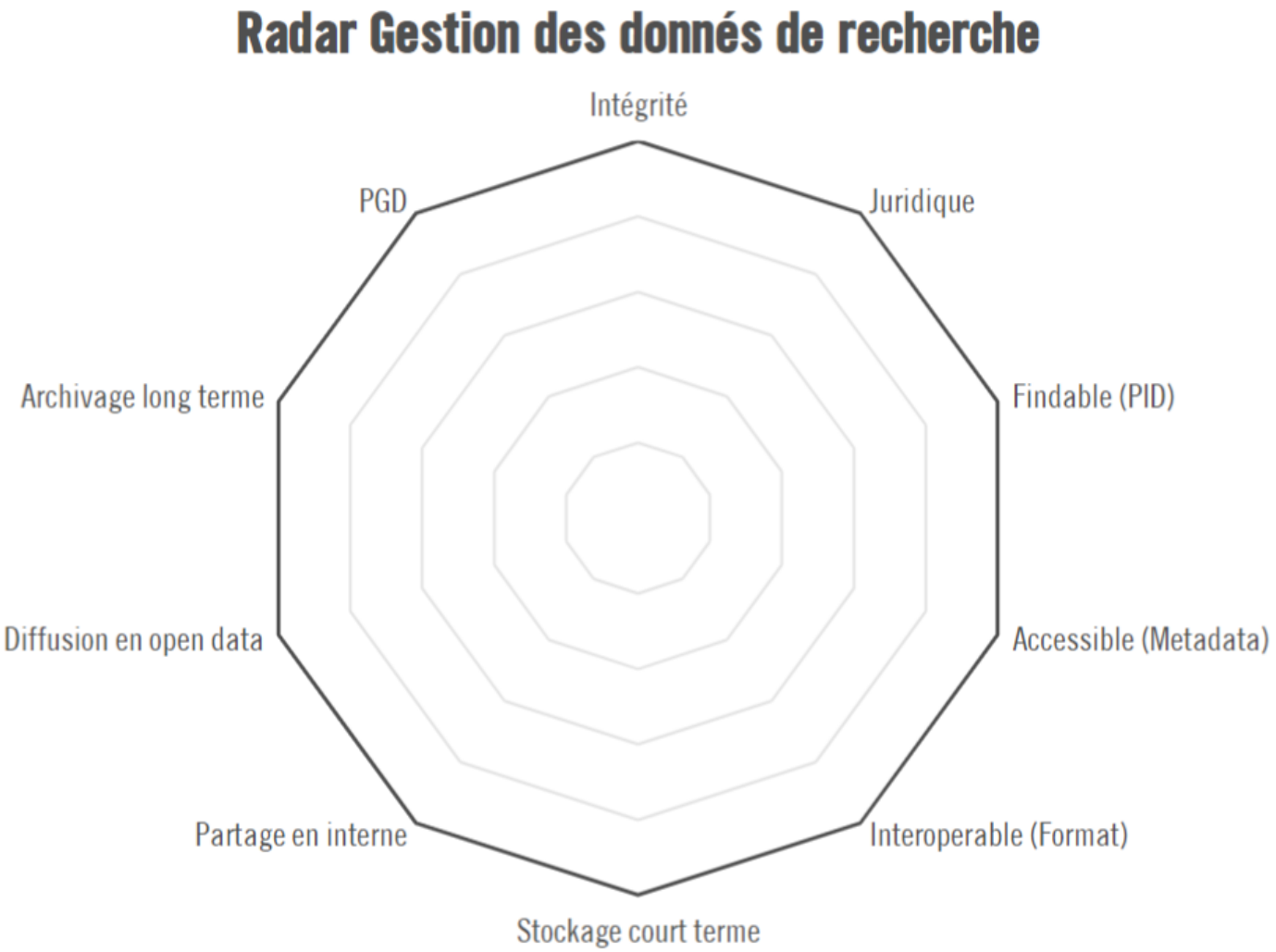
Formalisation des efforts dans un document

Plan de Gestion des Données (PGD) / *Data Management Plan (DMP)*



Source

Formalisation des efforts dans un document



Source

Ressources

Guides pratiques :
https://scienceouverte.couperin.org/
https://www.ouvrirlascience.fr/accueil/

Publications ouvertes : https://hal.science/

Données ouvertes : https://recherche.data.gouv.fr/fr

La science ouverte dans l'ED NSCo https://nsco.universite-lyon.fr/science-ouverte/

Quelques outils :
Cahier de laboratoire électronique https://datacc.elab.one/login.php
Plan de gestion de données https://dmp.opidor.fr/
Aide à la gestion & partage des données https://doranum.fr/

Exemple de feuille de route dans un labo : https://www.crnl.fr/fr/page-base/science-ouverte