Introduction à la reproductibilité en calcul scientifique

Céline Acary-Robert
LJK, GRICAD

ANF CNRS, 1-3 Juillet 2025

Reproductibilité : une définition générale

On dispose :

- d'un code,
- des données associés

On voudrait :

- refaire un calcul existant, changer un paramètre
- ajouter une contribution au code pour étendre ses fonctionnalités

Qui ?

- le chercheur lui-même, des membres de son équipe
- d'autres chercheurs

Quand ?

- maintenant ou plus tard

Bénéfices pour le chercheur

L'intérêt d'assurer la reproductibilité des résultats ?

- pour faciliter la réutilisation de votre travail, même localement
- pour avoir confiance dans le code qu'on reprend plus tard
- pour gagner du temps pour vous :
  - par exemple :
    - ➜ vous avez une suggestion d'un lecteur, d'un collègue pour améliorer un point
    - ➜ vous reprenez votre code 5 ans après ...
    - ➜ Un étudiant en thèse veut comparer ses résultats avec ceux d'un étudiant précédent

Bénéfices pour la recherche

- Pérenniser les résultats de recherche
- Donner confiance au lecteur/reviewer
- Augmenter la rapidité de diffusion des savoirs
- Augmenter l'efficacité de la recherche en réutilisant des choses déjà faites
- Augmenter la qualité des codes : un bug peut être signalé
- Favorise les collaborations : des contributions peuvent être apportées, ...

➜ Augmenter la confiance dans les produits de la recherche, pour soi, pour les collègues et pour le grand public en général

En pratique, pour le lecteur d'une publication

De quoi a-t-on besoin ?

- article, disponible sur une archive
- code, disponible grâce à un lien web, un identifiant pérenne
- jeu de données, disponible sur le web (une page ou un entrepôt)

➜ Les principes de la Science Ouverte offre un cadre idéal pour assurer la reproductibilité d'un travail.

Science Ouverte

Objectifs

- Diffusion sans entrave des publications, des données et des codes de la recherche
- Favoriser les avancées scientifiques (science cumulative) et constituer un levier pour l'intégrité scientifique

Cadre légal

- Science Ouverte : loi pour une république numérique 1ère loi 2016 : circulation des données et du savoir : ouverture des données publiques par défaut
- Loi pour l'intégrité scientifique : décembre 2020

Science Ouverte

Mise en œuvre : Plan National pour la Science Ouverte

- 1er plan : 2018 - 2021
- 2e plan : 2021 - 2024 : 4 axes pricipaux
  - Généraliser l'accès ouvert aux publications
  - Structurer, partager et ouvrir les données de la recherche publique (FAIR : Facile à trouver, Accessible, Interopérable, Réutilisable)
  - Ouvrir et promouvoir les codes sources produits par la recherche publique
  - Transformer les pratiques pour faire de la science ouverte le principe par défaut

Contexte général : science ouverte

Organismes français

- COSO : COmité pour la Science Ouverte (MESR) :
  - assure la mise en œuvre de la politique de Science Ouverte
  - édite des guides de bonnes pratiques : passeports pour la science ouverte
- Ateliers de la donnée : gestion des données de la recherche
  - apporte localement une expertise dans la gestion raisonnée des données.

Organismes à l'international

- Politiques et projets européens autour de l'Open Science (openAIRE, ...)
- COS: Center for Open Science US

Intégrité scientifique

L'intégrité scientifique est désormais inscrite dans le code de la recherche depuis 2020.

Office Francais de l'Intégrité Scientifique

- sous l'égide de l'HCERES
- intégrité scientifique devient une notion juridique

Principes de l'intégrité scientifique :

- éthique de la recherche
- caractère honnête et scientifiquement rigoureux
- impartialité et objectivité des recherches

  ➜ notion de fiabilité

Problématique de la reproductibilité

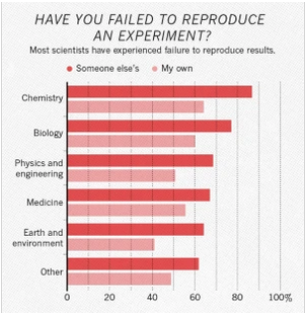
Diverses études sur la reproductibilité des résultats

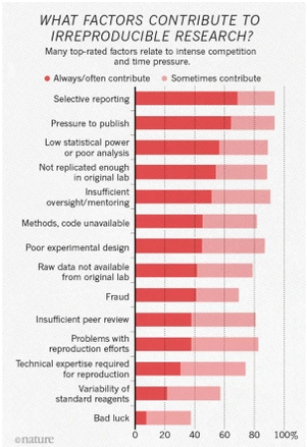
- notion qui émergée dans les années 2000, prise de conscience importante autour des années 2010
- “crise de la reproductibilité”
- cadre des recherche sur les méthodes de recherche : méta-science, méta-recherche

Reproductibilité comme objet d'études

- 2016 : **Questionnaire** en ligne envoyé à 1576 chercheurs de disciplines très différentes
  - 70 % ont échoué à reproduire un travail effectué par un autre chercheur
  - + de 50 % ont échoué à reproduire leur propre expérience

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





Causes du manque de reproductibilité

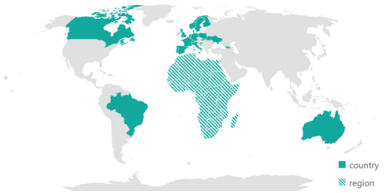
- reporting sélectif
- course à la publication : manque de temps
- manque de sensibilisation / formation / support sur ces enjeux
- fraude, erreur humaine
- ...

Reproductibilité comme objet d'études

- Economie
  - Reproductibilité de 60 articles (données + code + préanalyse disponibles)
  - 1/3 papiers reproduits, +10% après contact avec les auteurs
    Chang, Andrew C., and Phillip Li. 2017. "A Preanalysis Plan to Replicate Sixty Economics Research Papers That Worked Half of the Time." American Economic Review, 107 (5): 60–64. DOI: 10.1257/aer.p20171034
- Psychologie appliquée
  - Reproductibilité de 35 articles publiés entre 2015 et 2018 dans la revue *Cognition* (données disponibles)
  - 11 articles reproduits, 11 après contacts avec les auteurs : 63% max
    Estimating the Prevalence of Transparency and Reproducibility-Related Research Practices in Psychology (2014–2017), Tom E. Hardwicke, Robert T. Thibault, [...], and John P. A. Ioannidis, Volume 17, Issue 1, https://doi.org/10.1177/1745691620979806
- Ten Years Reproducibility Challenge Oct 11, 2019
  - Objectif : refaire tourner son code publié 10 ans après
  - Difficultés : retrouver le code + env. logiciel
    Challenge to scientists: does your ten-year-old code still run? Nature vol. 584 iss. 7822, 2020

Contexte international

- Beaucoup de communautés différentes se sont emparées du sujet
- Création de plusieurs réseaux en Europe et dans le monde : UKRN en 2019,...
- Projets européens : horizon europe (programme européen pour la recherche et l'innovation) –> projets autour de la metascience et de la reproductibilité (OSIRIS, TRUSTparency, iRISE, ...)



Les revues de publications

Publications

- Revues qui ont des pré-requis (variable en fonction des disciplines)
  - IPOL (machine learning), Computo (statistique), ReScience C (numérique)
- Badges de l'ACM (Association for Computing Machinery) : https://www.acm.org/publications/policies/artifact-review-and-badging-current
- The Graphics Replicability Stamp Initiative (GRSI) : liste de conférences (SIGGRAPH, ...)
- CASCAD en économie (https://www.cascad.tech/)

Contexte français

- 2022 : création du réseau français
- en complément de la science ouverte
- soutien du ministère
- adhésion : abonnement à la liste de diffusion

https://www.recherche-reproductible.fr

Contexte français

Vous pouvez contribuer !

- Proposer des wébinaires
- Contribuer au GT “Logiciels” (rédiger des fiches de bonnes pratique)
- Alimenter la base de données bibliograhique
- Proposer la synthèse d’un article en lien avec la reproductibilité
- Proposer une sensibilisation dans votre laboratoire

https://www.recherche-reproductible.fr

Reproductibilité : une grande diversité

Les bonnes pratiques vont dépendre de

- de la discipline
- mais aussi des techniques employées
- de la gestion des données
- etc ...

➜ Nécessité de préciser la terminologie

Définitions : catégorisation en fonction des méthodologies

Différents type de reproductibilité en fonction des contextes de recherche

- empirique : trace des conditions dans lesquelles l'expérience a eu lieu
  - paramètres : le PH, la température...,
  - ingrédients qui ont composé cette expérience (les espèces chimiques, ...)
  - protocole suivi pour la réaliser (ordre des étapes, matériel utilisé...).
- observationnelle : trace des conditions dans lesquelles les observations ont été réalisées
  - protocole d'observation
  - contexte, la durée, et les conditions de l'observation.
  - liste des éléments à observer, critères d'évaluation, etc.
  - détails de toutes les étapes de l'observation (lieux, horaires, outils utilisés).

➜ La reproductibilité empirique ou observationnelle sont souvent difficiles à atteindre parce que certaines expériences requièrent une précision importante dans la reproduction des conditions expérimentales ou observationnelles.

Définitions / thématiques

Différents type de reproductibilité en fonction des contextes de recherche

- statistique : trace de la raison qui fait qu'un résultat est significatif au sens statistique
  - justification du choix du test statistique
  - choix des paramètres du modèle
  - des valeurs de seuil, de la taille des échantillons
- computationnelle : trace de la production d'un résultat sur ordinateur
  - codes informatiques utilisés + leur version + enchainement
  - environnement logiciel, type de machine, logiciels installés utilisés
  - les jeux de données

➜ Pour la reproductibilité computationnelle : il existe plein d'outils informatiques qui facilitent grandement le travail !

Définitions qui font consensus pour le calcul numérique

- Répétabilité : la même équipe de recherche, en utilisant le même code (protocole, ..) et les mêmes données, on aboutit à des *conclusions/résultats/..* identiques
- Reproductibilité : une équipe différente, en utilisant le même code (protocole, ..) et les mêmes données, on aboutit à des *conclusions/résultats/..* équivalents (avec une tolérance)
- Réplicabilité : en prenant un jeu de données différents et/ou une implémentation différente de la méthode, on aboutit à des *conclusions/résultats/..* équivalents

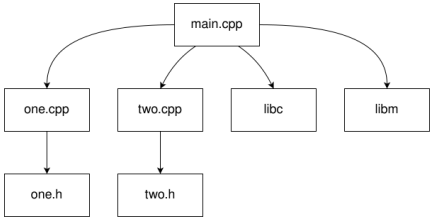
Sources de non-reproductibilité en calcul scientifique

- Différence de version de l'interpéteur et des dépendances
- Choix d'optimisation à la compilation
- Arrondis sur les nombres à virgule flottante
- Parallélisme qui change l'ordre des opérations
- CPU / GPU

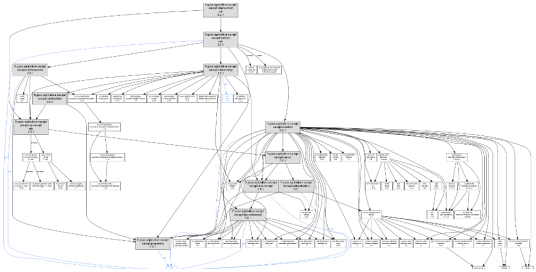
➜ Choix de la précision de la reproductibilité souhaitée

Exemples de problèmes de reproductibilité en calcul scientifique

Le code



Les dépendances



Exemples de problèmes de reproductibilité en calcul scientifique

Version de python3 :

En Python 3.6, les valeurs par défaut de rel_tol et abs_tol ont été ajustées :

python3.5	python3.6
import math math.isclose(1.0, 1.0000001) :: False	import math math.isclose(1.0, 1.0000001) :: True

Exemples de problèmes de reproductibilité en calcul scientifique

Version de dépendances : changement de syntaxe

NumPy < 2.0 **removed member**	NumPy >= 2.0 **migration guideline**
Inf	Use np.inf instead
Infinity	Use np.inf instead
infty	Use np.inf instead
mat	Use np.asmatrix instead
NaN	Use np.nan instead
NINF	Use -np.inf instead
NZERO	Use -0.0 instead
longfloat	Use np.longdouble instead

https://numpy.org/doc/stable/numpy_2_0_migration_guide.html

Pandas < 0.20	Pandas >= 0.20
df.sort('column_name')	df.sort_values('column_name')

Exemples de problèmes de reproductibilité en calcul scientifique

Version des dépendances : résultat d'une fonction

```
import random
random.seed(42)
data = [1, 2, 3, 4, 5]
random.shuffle(data)
print(data)
```

random 3.2	random 3.3+
Résultat : [4, 2, 3, 5, 1]	Résultat : [1, 5, 2, 4, 3]

```
import numpy
np.float32(3) + 3.
```

NumPy < 2.0	NumPy $>= 2.0$
float64	float32

Exemples de problèmes de reproductibilité en calcul scientifique

Compilation, parallélisation

$$H_n = \sum_{k=1}^{n} 1/k = ln(n) + \gamma + o(1)$$

- Intel(R) Xeon(R) Gold 5218R CPU @ 2.10GHz
- gfortran 14 -fopenmp

Algorithme naif en double précision, parallélisation OpenMP :

```
Gamma = 0.5772156649015328606
!$OMP parallel do private(i) reduction(+ : sum)
do i=1,n
   sum = sum + 1.0/i
enddo
!$OMP end parallel do
p = abs(sum - log(float(n)) - Gamma)
```

Problèmes de reproductibilité : parallélisation avec OpenMP

n $= 1\,000\,000$

Calcul	Séquentiel	OpenMP 1 thread
Somme	14.392726788474306	14.392726788474306
	OpenMP 4 threads	OpenMP 4 threads -Ofast
	14.392726788474306	14.392726222394913

n$= 1\,000\,000\,000$

Calcul	Séquentiel	OpenMP 1 thread
Somme	21.300481573265063	21.300481573265063
	OpenMP 4 threads	OpenMP 4 threads -Ofast
	21.3004815726470 87	21.300480911325508

Problèmes de reproductibilité en calcul scientifique : calcul sur GPU

- Intel(R) Xeon(R) Gold 5218R CPU @ 2.10GHz
- 1 GPU nvidia V100
- nvfortran 22.3-0 64-bit target on x86-64 Linux

```
Gamma = 0.5772156649015328606
!$acc kernels copy(sum)
devicetype = acc_get_device_type()
ngpus = acc_get_num_devices(devicetype)
do i=1,n
   sum = sum + 1.0/i
enddo
!$acc end kernels
p = abs(sum - log(float(n)) - Gamma)
```

Exemples de problèmes de reproductibilité en calcul scientifique

nvfortran

n= 1 000 000

Calcul	CPU	GPU
sum	14.39272678847431	14.39272678847431

n= 1 000 000 000

Calcul	CPU	GPU
sum	21.30048157326506	21.30048156797372

Calcul scientifique : bilan

- Version des interpréteurs et leurs dépendances
  - ➜ Partager l'environnement logiciel complet
  - ➜ Entretenir le code dans le temps
- Précision obtenue
  - ➜ Choisir la précision nécessaire
  - ➜ Partager les détails de compilation et d'éxécution

Les bonnes pratiques à mettre en place

- Gestion des données : utilisation et production
- Gestion des codes : écriture, exécution et diffusion

➜ Pour chaque étape il y a des outils pour vous aider !

Calcul scientifique : les bonnes pratiques

Gestion des données d'entrée

Dans une publication, vous indiquerez :

- auteurs, date de publication
- identifiant : DOI + lien téléchargement
- pré-taitement : conversion de format, calculs
- façon dont elles sont utilisées dans votre simulation

Calcul scientifique : les bonnes pratiques

Gestion des données de sortie : respect des principes FAIR

- format standard
- licence
- DOI + entrepôt
- description
- métadonnées

Calcul scientifique : les bonnes pratiques

Ecriture du code

- Mettre en œuvre les bonnes pratiques de développement
  ➜ pour faciliter la réutilisation du code
- Si possible, utiliser un gestionnaire d'environnement logiciel reproductible (Guix, Nix)
- Réfléchir à une licence : va réglementer la vie du logiciel
- Utiliser un outil de versionnement (git)
- Utiliser une forge logicielle pour automatiser les tests
- Rédiger une documentation + automatisation

Calcul scientifique : les bonnes pratiques

Exécution du code

Penser à sauvegarder dans des fichiers adaptés :

- Description des moyens de calcul (portable, cluster, ...)
- Environnement logiciel
  - description précise de l'environnement logiciel
  - les moyens de le reproduire
- Jeux de paramètres
- Workflow d'exécution

Calcul scientifique : les bonnes pratiques

Diffusion du code

- Mettre une license (licence, contributeurs, etc ..)
- Identifiant pérenne : HAL + SWH

  - référence à une version à l'instant t + bout de code
- Description de l'environnement logiciel
- Fichier de workflow

Personne n'est parfait !

Idée : changer ses pratiques en douceur 😉

- Mener une réflexion sur vos pratiques pour identifier les freins
  - facteurs techniques
  - facteurs structurels : incitation à la publication
- Identifier les points que vous pouvez améliorer et les outils qui peuvent vous aider

Conclusions

En pratique

- On perd un peu de temps à la mise en place du protocole
- On en gagne vraiment beaucoup *a posteriori*

Qui est gagnant ?

- Vous : rigueur + gain de temps
- Les reviewers : confiance
- Vos collègues : facilite la transmission des codes, les collaborations
- Le grand public : confiance

Exercice

Mise en pratique

- par groupe de 4 ou 5
- présentations
- identifier une bonne pratique que vous avez
- réfléchissez à un élément que vous pourriez facilement améliorer
- identifier un frein (formation, temps, autre ..)
- discussion
- un membre du groupe fait une restitution

Ressources et bibliographie

- MOOC recherche reproductible pour une science transparente I et II
- https://www.inshs.cnrs.fr/fr/cnrsinfo/replication-reproductibilite-reutilisation-concepts-et-enjeux
- https://doc.hal.science/ressources/essentiels/CCSD_essentiels_logiciels.fr.pdf
- https://inria.hal.science/hal-01872189v2/
- https://indico.math.cnrs.fr/event/10998/contributions/10948/attachments/4774/7400/SWH-HAL_RNBM-Mathrice-v7.pdf
- https://www.hceres.fr/sites/default/files/media/downloads/lofisfaitlepointjanvier2024.pdf
- https://www.allea.org/wp-content/uploads/2017/05/ALLEA-European-Code-of-Conduct-for-Research-Integrity-2017.pdf
- https://www.ouvrirlascience.fr/le-comite-pour-la-science-ouverte/
- https://inria.hal.science/hal-04930405v1/document
- NicolasBonneel,DavidCoeurjolly,JulieDigne,NicolasMellado,CodeReplicabilityinComputerGraphics, ACMTrans.onGraphics(ProceedingsofSIGGRAPH2020),39:4
- https://www.acm.org/publications/artifacts
- https://www.sciencedirect.com/science/article/abs/pii/S157401372400039X
- https://indico.engineering.univ-lyon1.fr/event/12/contributions/68/attachments/80/113/2024_04_04_JRS-Michel-KERN.pdf