Reproductibilité des données de recherche : définitions, enjeux, risques, bénéfices, recommandations

Eric Cahuzac, INRAE-DipSO

GTSO Données Couperin, 4 décembre 2023

Thanks to Odile Hologne et Esther Dzale

Contexte

- De quoi n’allons nous pas parler
  - Crise de la reproductibilité de la recherche et de ses causes...
  - Les critères d’évaluation de la recherche, la course à la publication, aux financements de la recherche...
  - De l’éthique de la recherche...
  - tout cela bouge...
- De quoi allons nous parler: de science ouverte
  - une science disponible, qui s’ouvre aux autres
  - une science plus précise,
  - une science qui capitalise le savoir
  - une science qui innove plus rapidement
  - ...

La DipSO à INRAE

- C’est quoi
  - créée pour la mise en œuvre et l’animation de la politique SO INRAE
  - une direction qui promeut les valeurs de la SO au sein de l’établissement, par des actions autour
    - de l’édition scientifique
    - des sciences participatives
    - d’une offre numérique pour permettre aux chercheurs de produire une science plus ouverte
  - C’est donner les moyens de faire de la science au plus proche des principes FAIR
  - https://science-ouverte.inrae.fr/

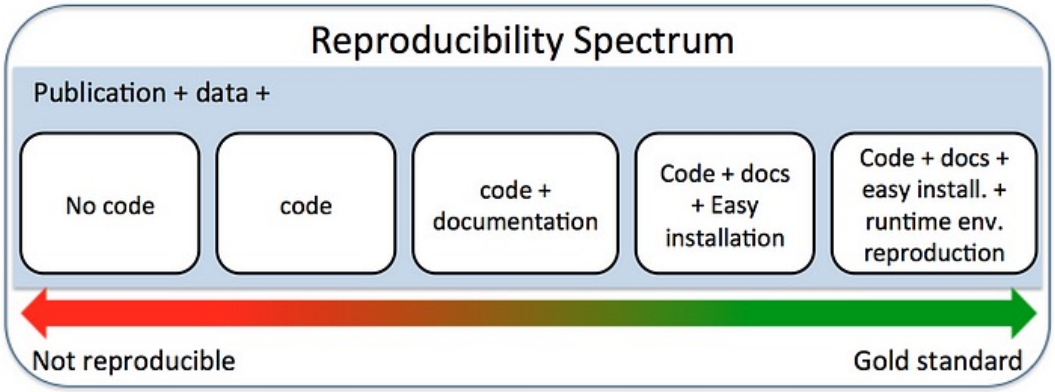
→ En faisant de la science ouverte on fait un grand pas vers la reproductibilité

De l'ouverture à la reproductibilité, via le FAIR

- Ouvrir c'est un bon début
  - c'est se mettre dans les bonnes conditions pour être reproductible
  - FAIR ce sont des principes très utiles pour s'en rapprocher

- FAIR ne garanti pas la reproductibilité
  - R c'est définir le type de licence pour cadrer la réutilisation
  - R ne veut pas dire ouvert, ni Reproductible

- Être reproductible ce n'est pas un état, c'est une cible que l'on se fixe
  - tout comme pour FAIR, on n'est pas reproductible/non reproductible
  - il faut améliorer nos pratiques pour le devenir
  - partager les résultats les données et les codes ce n'est pas suffisant



https://medium.com/@aakalin/scientific-data-analysis-pipelines-and-reproducibility-75ff9df5b4c5

La reproductibilité de la recherche: définition?

- Qu'entend-on par reproductibilité
  - Dans les premières travaux, c'est la notion de « ré-exécutable »
  - On trouve ensuite des notions plus précises, selon si :
    - c'est vous/votre équipe qui rejoue l'expérimentation, ou une autre équipe indépendante
    - l'expérimentation est rejouée en utilisant le même dispositif expérimental (y compris données, logiciels et codes) ou pas
  - On peut s'appuyer sur des standards (BIPM, NISO):
    - repeatability : même équipe, même dispositif expé.
    - reproductibility : équipe différente, même dispositif expé.
    - replicability : équipe différente, dispositif expé. différent
  - Une approche pragmatique de la reproductibilité :
    - recherche qui met à disposition toutes les informations qui permettent à n'importe quel chercheur indépendant de reproduire le résultat

R Di Cosmo,M Gruenpeter,S Zacchiroli(2020) Referencing Source Code Artifacts : a Separate Concern in Software Citation, Computing in Science & Engineering, DOI: 10.1109/MCSE.2019.2963148

Les enjeux

- Pourquoi faut-il être reproductible dans ses recherches?
  - Confiance: « seuls les résultats reproductibles sont des résultats scientifiques » Huschka (2013).
  - Performance: « se donner du temps pour inventer de nouvelles choses plutôt que de recréer celles qui existent déjà » Vandewalle et al. (2009)
  - Bénéfices : moi, les autres, la communauté

- Peut-on toujours être reproductible ?
  - Non, certaines conditions ne le permettent pas: sujet d'étude, savoir tacites
  - Que fait-on alors: transparence, traçabilité

Huschka, Denis, Why Should We Share Our Data, How Can it Be Organized, and What are the Challenges Ahead? (May 30, 2013). RatSWD_WP_ 216, Available at SSRN: http://dx.doi.org/10.2139/ssrn.2272028

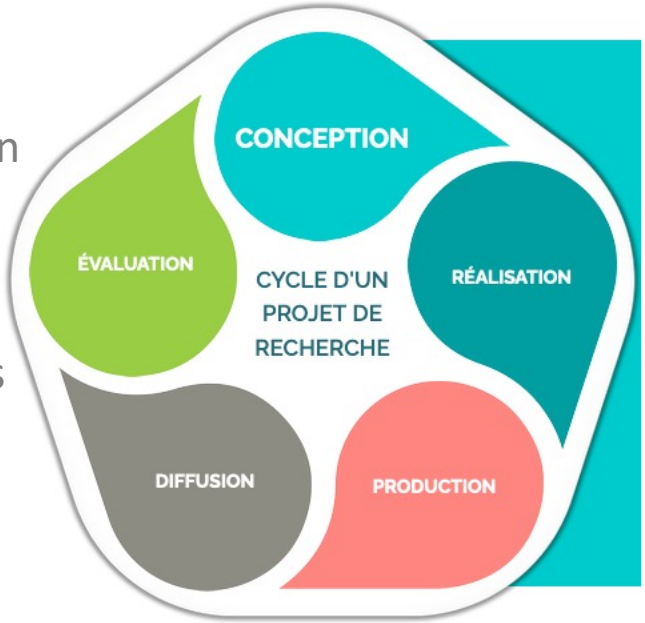
P.Vandewalle, J.Kovacevic and M.Vetterli, "Reproducible research in signal processing," in IEEE Signal Processing Magazine, vol.26, no.3,pp.37-47, May2009
DOI:10.1109/MSP.2009.932122

Être reproductible

- Quand ?
  - Au quotidien, à chaque étape du cycle de vie du projet/des données
  - On ne fait pas du reverse-engineering de la reproductibilité.
- Comment ?
  - Tout au long du cycle de vie du projet

en facilitant l'évaluation par les tiers

en décrivant les règles de diffusion et de réutilisation



en planifiant les étapes, en partageant ses hypothèses et ses sources

en donnant accès, en traçant les différentes étapes

en rendant accessibles l'ensemble des productions

Sur quoi doivent porter nos efforts ?

- Les données
  - utilisées, produites...
- Les codes
  - qui ont servi à produire les données, à mettre en œuvre les modèles, à obtenir les résultats
- L'environnement
  - les méthodes de travail
  - les outils et l'environnement logiciels

Être reproductible dans ses données, c'est

- déposer dans un entrepôt ouvert
- bien renseigner les métadonnées
- penser aux bonnes licences: ODBL, Etalab 2.0, CC-BY
- lors de la collecte:
  - utiliser des standards : unités/outils de mesures, des référentiels
  - utiliser les bons formats : ouverts/pérennes/lisibles par les machine
- lors des traitements
  - ne pas modifier les données brutes
  - éviter les interventions manuelles (copier/coller)
  - privilégier la ligne de commande: des programmes simples, clairs, génériques, sobres → règles de nommage par ex.
  - inutile de sauvegarder des fichiers de données intermédiaires (temp)
  - travailler dans un environnement structuré, partagé et sauvegardé

Valérie Orozco, Christophe Bontemps, Élise Maigné, Virginie Piguet, Annie Hofstetter, et al.. How To Make A Pie: Reproducible Research for Empirical Economics & Econometrics. 2019. .hal-02025843

Être reproductible dans son code, c'est

- décrire, commenter ses codes, et leurs enchaînement

  → logigrammes, notebook

- utiliser des standards de développement : framework (Django, Symfony), langages de programmation courants, libre (R, Python)

  → catalogue du logiciel libre : https://code.gouv.fr/sill/

- travailler dans un environnement partagé et sauvegardé
- choisir les bonnes licences: Apache, GNU GPLv3, MIT
- travailler avec des outils de gestion de version,

  → dans une forge logicielle → intégration continue

- archiver son code → Software Heritage
- publier son code/logiciels, dans des revues ouvertes

Être reproductible dans son environnement de travail, c'est

- bien décrire et préparer à ouvrir
  - l'environnement, les conditions de l'expérimentation
  - les sources de données, bibliographiques et leurs droits d'utilisation, de réutilisation,
  - les outils (matériels, logiciels, langages) utilisés pour le traitement, pour le calcul
  - tracer et versionner sa progression

  → le plan de gestion de données/logiciels du projet
    - des outils : Opidor, Data Stewardship, ARGOS (MA)
    - des trames de PGD : établissement, ANR, Horizon Europe...
    - décrire tous les produits de recherche

  → les cahiers de laboratoire électroniques (ElabFTW, LabGuru) avec horodatage ou des TMS

Être reproductible dans son environnement de travail, c'est

- faciliter la réutilisation des « objets numériques »
  - déposer ses données et codes dans des environnements (entrepôts) indexés (PID), accessibles, ouverts, sauvegardés...
  - travailler dans un environnement ouvert, partagé avec collaborateurs et sauvegardé (système de gestion de tâches, espaces partagés,...)
  - lier les données, le code et les résultats → des documents computationnels (Jupyter, Rstudio NoteBook)
  - indexer les objets : par un vocabulaire contrôlé (liste, taxonomie, thésaurus, ontologie)→ classer, raisonner
  - publier dans des revues ouvertes: ses résultats, des data papers,

- Et malgré tout...

Être reproductible dans son environnement de travail, c'est

- Reproductibilité de l'environnement d'exécution
  - le logiciel utilisé doit se comporter de la même façon quelle que soit la machine qui l'exécute
    - dépend du compilateur, des bibliothèques utilisées, de leurs versions...
  - mettre à disposition avec les données, le pipeline d'analyse de ces données **et** les dépendances logicielles
    - copier tout le pipeline dans une VM ou un conteneur (docker, singularity)
    - utiliser un gestionnaire de paquets (Pip, Conda) qui gère les dépendances logicielles de plusieurs langages (Python et +).
    - déploiement du pipeline (GUIX)

Conclusion: on va dans la bonne direction !

- La reproductibilité n'est plus (ou presque) un pb d'outils
  - le plus gros frein c'est nous (chercheurs)
- Tous les chercheurs (ou presque) la souhaitent et en ont besoin
  - les sources de NR: négligence, erreur, la fraude
  - complexité!
- Des initiatives (dans différents domaines) nous aident à évaluer la reproductibilité de nos productions → confiance
  - des revues: Computo
  - des conférences: Agiles
  - des structures/organisations :
    - de certification: Cascad
    - délivrance de badges : ACM
    - guides pour les auteurs : Sciences Sociales

Merci

Discussions

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

INRAE