MAJ la plus récente : 31/07/2019

Contacts :
caroline.abela@cnrs.fr
martine.courbin@inria.fr
sabrina.granger@u-bordeaux.fr

Repenser la robustesse et la fiabilité en recherche : les chercheurs face à la crise de la reproductibilité. Compte-rendu de la journée d'étude du 29 mars 2019

Sommaire

La journée a été organisée par l'Unité régionale de formation à l'information scientifique et technique (Urfist) de Bordeaux en partenariat avec : le CNRS (délégation Aquitaine), l'Inria (Centre Bordeaux Sud-Ouest), l'Université de Bordeaux (pôle Ressources humaines et développement social). La journée a permis de réunir des chercheurs et des doctorants de toutes disciplines, ainsi que des personnels de soutien à la recherche.

Compte-rendu rédigé par Caroline Abela (UMR Passages CNRS), Martine Courbin-Coulaud (Inria) et Sabrina Granger (Urfist de Bordeaux) notamment à partir du texte fourni par Christine Buisson (LICIT ; IFSTTAR / ENTPE).

MAJ la plus récente : 31/07/2019

Contacts :
caroline.abela@cnrs.fr
martine.courbin@inria.fr
sabrina.granger@u-bordeaux.fr

1/Argumentaire de la journée

1.A/ Reproductibilité de la recherche, de quoi parle-t-on ?

Les chercheurs sont confrontés au fait de ne pouvoir obtenir les mêmes résultats soit en reprenant les mêmes méthodes et/ou les mêmes données soit en s'appuyant sur de nouveaux jeux de données et/ou d'autres méthodes poursuivant le même objectif de recherche. Le problème de la fidélité et de la réitération des résultats se pose à l'échelle collective (i.e. unité de recherche, spécialistes d'une même discipline) comme individuelle (i.e. : reproductibilité spatio-temporelle (Desquilbet 2018). Il n'existe pas de définition standard de l'expression "recherche reproductible" dans la mesure où chaque discipline va définir avec ses propres critères ce qu'est un résultat. Ainsi que le souligne Leonelli (Leonelli 2018), selon la discipline envisagée, un résultat peut être un objet, un modèle ou encore, un protocole.

1.B/ En quoi et pourquoi le reproductibilité de la recherche est-il un sujet d'actualité?

Le sujet de la reproductibilité est ancien (Barba 2018) et d'aucuns considèrent qu'il vaut mieux évoquer un changement de paradigme de la recherche plutôt qu'une crise (Fanelli 2018). **Il serait tentant de s'en référer à la conception poppérienne de la science selon laquelle l'erreur et sa réfutation résident au cœur même du processus scientifique. Mais le phénomène prend une ampleur telle que les notions mêmes de résultat et de fiabilité sont remises en cause.** Si les réplications ne sont pas possibles, quelle est la valeur des travaux précédents s'ils s'avèrent non reproductibles (Zwaan et al. 2017) ?

Les causes de cette crise ne relèvent pas forcément de manquements délibérés à l'intégrité scientifique (i.e. : p-hacking, HARK, selective reporting, etc.) : des méthodes statistiques mal employées, notamment en raison de la sophistication croissante des méthodes (Wilcox et Rousselet 2018) ; des jeux de données mal proportionnés ; des interprétations problématiques de la p-value (Lakens et al. 2017) et plus généralement, un problème de puissance statistique qui perdure depuis des décennies (Lilienfeld et Waldman 2014; Lakens et Albers 2017). La liste n'est pas exhaustive.

1.C/ Un système éditorial peu adapté

Au-delà des problèmes inhérents à la production des données et de leur analyse, les chercheurs doivent composer avec un système éditorial qui n'incite ni à publier les résultats négatifs ni les auto-rétractations. Les dead-ends sont souvent omises et les travaux présentant les résultats positifs bénéficient des faveurs des éditeurs des revues qualifiantes. **Les pratiques éditoriales scientifiques sont jugées inadaptées pour faire face aux défis de la reproductibilité** (Cornelius 2018; Yale Law School Roundtable on Data and Code Sharing 2010), y compris lorsque les éditeurs affichent une politique volontariste de partage des données (Frankenhuis et Nettle 2018; Stodden 2011). Enfin, le système d'évaluation actuel n'accorde pas une place majeure aux études de réplication.

MAJ la plus récente : 31/07/2019
Contacts :
caroline.abela@cnrs.fr
martine.courbin@inria.fr
sabrina.granger@u-bordeaux.fr

1.D/ Quelle(s) réponse(s) de la part des chercheurs face à ce phénomène ?

Mais de nombreux projets voient le jour, y compris dans des domaines réputés imprenables eu égard à la nature des données observées (Milcu et al. 2018), de nouvelles formes éditoriales émergent et l'environnement même de production des connaissances scientifiques évolue.

2/ Rappel des interventions et liens vers les supports de présentation

- "A simple cure to the p < 0.05 disease", Guillaume Rousselet, Université de Glasgow ; **@robustgar**
- "Curate Science : Nutritional Labels for Scientific Transparency", Etienne LeBel, Université de Louvain ; découvrir Curate science en un visuel ; **@eplebel** ; **@curatescience**
- "Assisted Authoring for avoiding inadequate claims in scientific reporting", Patrick Paroubek, LIMSI (Laboratoire de recherche en Informatique pluridisciplinaire), CNRS ; **@LimsiLab**
- "Reproducibility in ecological research: do we need to worry?", Alexandru Milcu, Centre d'écologie fonctionnelle et évolutive, CNRS ; **@EcotronCNRS**
- "rOpenSci, revues de packages R par les pairs pour une meilleure science", Maëlle Salmon, rOpenSci ; **@ma_salmon**; **@rOpenSci**
- "ReScience X : projet de revue dédiée à la reproductibilité expérimentale en psychologie", Etienne Roesch, Université de Reading ; **@etienneroesch**

3/ Quelles solutions pour une recherche plus reproductible ?

Nous avons choisi pour ce compte-rendu de nous focaliser sur les solutions proposées par les intervenants.

3.A/ S'affranchir de la p-value : repenser l'approche des méthodes statistiques en recherche

Guillaume Rousselet a expliqué les défauts méthodologiques de l'utilisation du test statistique basé sur la *p-value* (*probability value*)[1].

D'une part, ce test vise à évaluer la possibilité de rejeter l'hypothèse nulle (par exemple « le traitement n'a aucun effet »), nullement de valider avec raison une hypothèse non nulle (« le traitement a un effet »). Pourtant, c'est cette dernière utilisation qui est faite le plus couramment ; ce qui aboutit à des résultats arbitraires.

D'autre part, une unique expérience ne permet pas d'obtenir une puissance statistique suffisante. Seules de nombreuses répétitions de cette expérience, dans des conditions très voisines, peuvent affermir la confiance que l'on a dans ses résultats. En opposition au manque de méthode de la part de non statisticiens, qui aboutit à une puissance statistique insuffisante Guillaume Rousselet

[1] *« The p-value is a number between zero and one, representing a probability based on the assumption that the null hypothesis is actually true. Given that assumption,* ***the p-value indicates the frequency with which the researcher, if he repeated his experiment by collecting new data, would expect to obtain data less compatible with the null hypothesis than the data he actually found****.» (Randall & Welser, 2018, p. 19).*

MAJ la plus récente : 31/07/2019
Contacts :
caroline.abela@cnrs.fr
martine.courbin@inria.fr
sabrina.granger@u-bordeaux.fr

recommande d'utiliser un ensemble beaucoup plus riche d'analyses de données : utiliser une analyse par déciles, par exemple.
Ainsi Guillaume Rousselet exhorte les chercheurs à ne pas se focaliser sur l'utilisation de la p-value et même à s'en affranchir, car elle ne devrait pas constituer un objectif en soi.

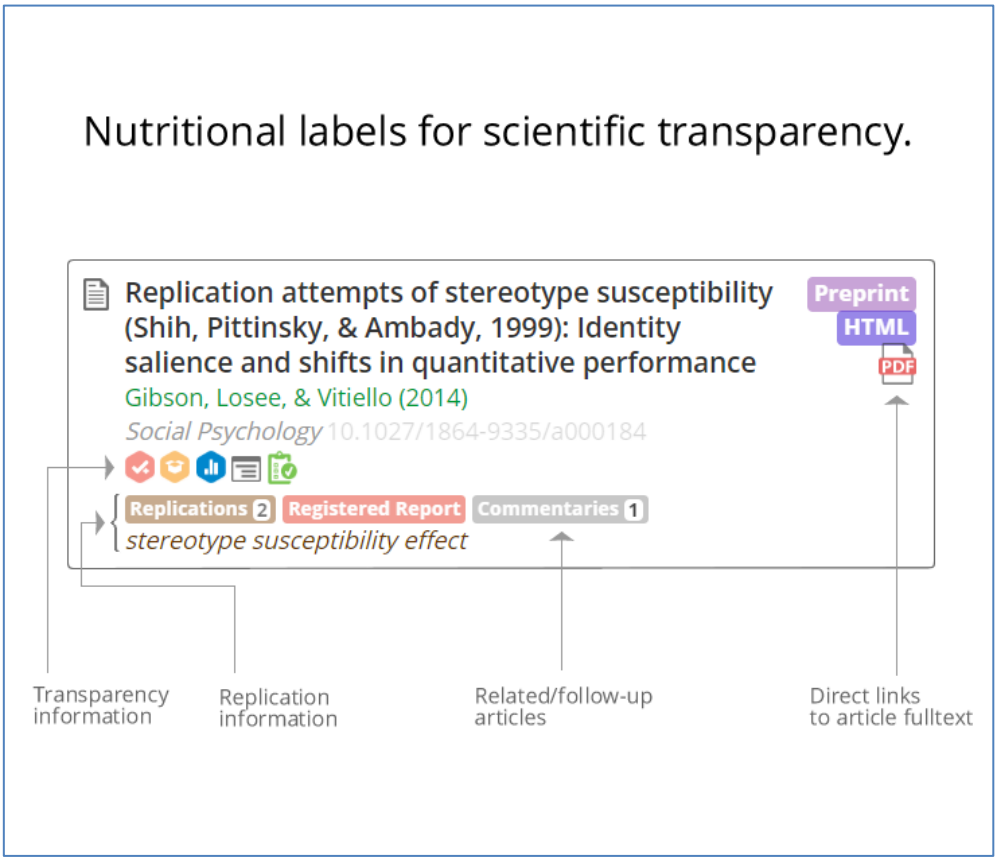
3.B/ Focus sur le rôle de la visualisation de données

Guillaume Rousselet nous encourage également à quantifier nos erreurs de mesure, à utiliser des représentations graphiques illustratives et détaillées, à partager nos codes pour permettre la reproductibilité, à pratiquer la répétition d'expériences déjà conduites de manière à faire face à l'incertitude expérimentale, à considérer les apports de l'épistémologie à nos pratiques quotidiennes, enfin à nous conduire avec honnêteté et modestie quant à nos méthodes d'analyse. La significativité statistique doit en effet être questionnée pour faire de la science.

3.C/ Partager efficacement codes et données

L'objectif est de vraiment viser une plus grande transparence et fiabilité des travaux. Etienne Roesch démontre tout l'intérêt de l'évaluation d'un papier sur la méthode et non sur les résultats ; cela oblige le scientifique à très bien documenter sa méthode. L'expérience présentée par Alexandru Milcu montre l'importance de la collaboration dans le partage afin d'évaluer les effets de variabilité, surtout dans un domaine comme celui de l'écologie.
Si la mise à disposition des données et des codes constitue un premier pas vers une recherche plus reproductible, rendre ces matériaux véritablement ré-exploitables (par d'autres chercheurs, comme par l'auteur initial) demeure encore un défi à l'heure actuelle. Il existe de nombreux dispositifs pour améliorer la transparence de la recherche : les *registered reports*, les logiciels de versionnage, les sites de *social coding*, les entrepôts de données de recherche, etc. Mais ces informations à haute valeur ajoutée demeurent disséminées dans des silos. L'ambition de la plateforme *Curate Science*, présentée et co-fondée par Étienne Le Bel dans le cadre d'un projet financé par l'Union européenne, est de réunir ces différents niveaux d'information *via* une même interface. L'article est-il issu d'une procédure de pré-registration? Où trouver les données et les codes? Quelles sont les analyses des rapporteurs ? Un système de tags synthétise les différentes informations liées à chaque article. On distingue d'une part, ce qui relève d'une transparence "primaire" : liens vers les données en accès libre et standards de compatibilité. D'autre part, l'utilisateur accède aux éléments inhérents à la transparence dite "secondaire" : *open peer review*, mais aussi déclaration de conflits d'intérêt, résumés, figures, etc. La version beta de la plateforme est annoncée pour l'automne 2019 et la mise en production est prévue pour l'été 2020.

MAJ la plus récente : 31/07/2019
Contacts :
caroline.abela@cnrs.fr
martine.courbin@inria.fr
sabrina.granger@u-bordeaux.fr



L'initiative pilotée par Patrick Paroubek constitue l'un des volets du projet européen MIROR (bourse Marie Sklodowska Curie) et traite quant à elle des erreurs de *reporting* dans le domaine biomédical. Il s'agit de fournir aux auteurs des outils pour identifier les éventuelles allégations inadéquates, en s'appuyant sur les techniques de traitement du langage naturel. La tâche est ardue car les vecteurs d'erreurs sont de nature variée : problème de citation, erreur d'interprétation des résultats ou des protocoles de l'étude citée, présentation déformée des résultats, etc. En outre, les erreurs de *reporting* peuvent ne pas être identifiées par les rapporteurs et se propager.
Enfin, le *peer review* de packages contribue aussi à une plus grande fiabilité des données scientifiques et *in fine*, à une plus grande reproductibilité. L'initiative de rOpenSci, présentée par Maëlle Salmon, permet d'évaluer la qualité des packages R soumis, à l'instar de ce qui se pratique pour les articles. Les rapporteurs évaluent ainsi si le package est accompagné d'une documentation exhaustive, si l'on constate un haut niveau de couverture des tests, ou encore, le degré de lisibilité et d'utilisabilité du code. Outre ce travail d'évaluation, rOpenSci permet aux développeurs de rendre leur code citable et identifiable en attribuant des DOI.
Au-delà de faire connaître son travail dans sa communauté, le procédé de révision comme décrit par Maëlle Salmon permet un échange entre scientifiques et "réviseurs". Chacun peut adopter de meilleures pratiques. Même si les modes d'évaluation actuels n'en tiennent pas assez compte, ce type

Contacts :
caroline.abela@cnrs.fr
martine.courbin@inria.fr
sabrina.granger@u-bordeaux.fr

de partage ne doit pas être vu comme une perte de temps mais plutôt comme une valorisation du code, une reconnaissance et au final, un moyen de favoriser un changement dans les mentalités en faisant comprendre l'erreur et non pas en culpabilisant.

3.D/ Pratiquer la répétition d'expériences

Pratiquer la répétition d'une même expérience (dans des conditions similaires ou très voisines) apporte également un gage de confiance dans l'obtention des résultats, et constitue même un passage obligé pour certaines disciplines.
Mais comment appréhender la reproductibilité de résultats lorsque tous les facteurs de l'expérimentation ne peuvent être contrôlés. Alexandru Milcu, en présentant une expérience d'écologie conçue pour tester l'impact de variables généralement non contrôlées dans les expériences de ce domaine, montre comment chercher à appréhender la variabilité comme un élément constitutif de la démarche scientifique. Ainsi, 14 laboratoires de recherche publique de 5 pays européens ont associé leurs efforts pour reproduire une même expérimentation, mais avec différents niveaux d'hétérogénéité. Les résultats de cette expérimentation multi-laboratoires révèlent que l'introduction délibérée d'hétérogénéité génétique dans le protocole expérimental réduit la variation des résultats entre laboratoires.

4/ Intégrer une réflexion épistémologique aux pratiques scientifiques : la reproductibilité de la recherche appelle un changement culturel, pas seulement technique

En guise de conclusion, nous retenons que les intervenants exhortent à des changements de pratiques, mais surtout, à une évolution culturelle. Guillaume Rousselet a mis en lumière les conséquences d'un usage dévoyé de la p-value et appelle les chercheurs à ne plus considérer qu'une étude suffit à prouver un effet ou corroborer une hypothèse. Citant le célèbre statisticien Andrew Gelman, Guillaume Rousselet considère que la variabilité des résultats et l'incertitude doivent retrouver une place dans le processus de recherche. Par ailleurs, l'intervention de Patrick Paroubek a donné à voir les perspectives offertes par les techniques d'intelligence artificielle pour une recherche plus traçable. Etienne Roesch et Alexandru Milcu soulignent quant à eux la nécessité de se déprendre d'un modèle plaçant le résultat et non plus le processus d'analyse au cœur de la démarche scientifique. Les études de réplication constituent ainsi une forme éditoriale adaptée pour appréhender les nuances de variabilité d'un résultat. La question de la reproductibilité participe ainsi à faire émerger un nouvel *ethos* scientifique, tourné vers plus de transparence comme le souligne Etienne Lebel ou favorisant une démarche bienveillante de coopération entre chercheurs comme le montre Maëlle Salmon.

MAJ la plus récente : 31/07/2019
Contacts :
caroline.abela@cnrs.fr
martine.courbin@inria.fr
sabrina.granger@u-bordeaux.fr

Sources

Barba, Lorena A. 2018. « Terminologies for Reproducible Research ». *arXiv:1802.03311 [cs]*, février. http://arxiv.org/abs/1802.03311.

Benureau, Fabien, et Nicolas Rougier. 2017. « Re-run, Repeat, Reproduce, Reuse, Replicate: Transforming Code into Scientific Contributions ». *arXiv:1708.08205 [cs]*, août. http://arxiv.org/abs/1708.08205.

Cornelius, Stephen. 2018. « Scholarly publishing is stuck in 1999 ». *Stephen Cornelius* (blog). 15 avril 2018. https://medium.com/@stphencornelius/scholarly-publishing-is-stuck-in-1999-507ab9bb06f5.

Desquilbet, Loïc. 2018. « Répétabilité, reproductibilité, et concordance de méthodes de mesure ». https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0ahUKEwiBy6qMy4fbAhXEfFAKHQRBDp8QFgguMAA&url=https%3A%2F%2Feve.vet-alfort.fr%2Fmod%2Fresource%2Fview.php%3Fid%3D13266&usg=AOvVaw1ZbfUQpfig29zCNvqtcvuv.

Fanelli, Daniele. 2018. « Is Science Really Facing a Reproducibility Crisis, and Do We Need It To? » *Proceedings of the National Academy of Sciences* 115 (11): 2628-31. https://doi.org/10.1073/pnas.1708272114.

Frankenhuis, Willem, et Daniel Nettle. 2018. « Open Science is Liberating and Can Foster Creativity ». *Open Science Framework*, février. https://doi.org/10.17605/OSF.IO/EDHYM.

Goodman, Steven N., Daniele Fanelli, et John P. A. Ioannidis. 2016. « What Does Research Reproducibility Mean? » *Science Translational Medicine* 8 (341): 341ps12-341ps12. https://doi.org/10.1126/scitranslmed.aaf5027.

Ioannidis, John P. A. 2005. « Why Most Published Research Findings Are False ». *PLOS Medicine* 2 (8): e124. https://doi.org/10.1371/journal.pmed.0020124.

Lakens, Daniel, Federico G. Adolfi, Casper Albers, Farid Anvari, Matthew A. J. Apps, Shlomo Engelson Argamon, Marcel A. L. M. van Assen, et al. 2017. « Justify Your Alpha: A Response to "Redefine Statistical Significance" ». *PsyArXiv*, septembre. https://doi.org/10.17605/OSF.IO/9S3Y6.

Lakens, Daniel, et Casper Albers. 2017. « When power analyses based on pilot data are biased: Inaccurate effect size estimators and follow-up bias ». *PsyArXiv*, juillet. https://doi.org/10.17605/OSF.IO/B7Z4Q.

Leonelli, Sabina. 2018. 'Re-Thinking Reproducibility as a Criterion for Research Quality'. *History of Economic Thought and Methodology*, January, 19.

Lilienfeld, S.O., and I.D. Waldman, eds. 2014. "Maximizing the Reproducibility of Your Research." *Psychological Science Under Scrutiny: Recent Challenges and Proposed Solutions*, February. https://doi.org/.

Milcu, Alexandru, Ruben Puga-Freitas, Aaron M. Ellison, Manuel Blouin, Stefan Scheu, Grégoire T. Freschet, Laura Rose, et al. 2018. « Genotypic Variability Enhances the Reproducibility of an Ecological Study ». *Nature Ecology & Evolution* 2 (2): 279. https://doi.org/10.1038/s41559-017-0434-x.

MiRoR Project. 2016. « Scientific Programme MiRoR : Methods in Research on Research ». *Projet MiRoR* (blog). 25 janvier 2016. http://miror-ejd.eu/scientific-programme/.

Nuijten, Michèle B., Chris H. J. Hartgerink, Marcel A. L. M. van Assen, Sacha Epskamp, et Jelte M. Wicherts. 2016. « The Prevalence of Statistical Reporting Errors in Psychology (1985–2013) ». *Behavior Research Methods* 48 (4): 1205-26. https://doi.org/10.3758/s13428-015-0664-2.

Stodden, Victoria. 2011. « Trust Your Science? Open Your Data and Code ». *Amstat News*, 2.

MAJ la plus récente : 31/07/2019
Contacts :
caroline.abela@cnrs.fr
martine.courbin@inria.fr
sabrina.granger@u-bordeaux.fr

Wilcox, Rand R., et Guillaume A. Rousselet. 2018. « A Guide to Robust Statistical Methods in Neuroscience ». *Current Protocols in Neuroscience* 82 (janvier): 8.42.1-8.42.30. https://doi.org/10.1002/cpns.41.
Yale Law School Roundtable on Data and Code Sharing. 2010. « Reproducible Research ». *Computing in Science & Engineering* 12 (5): 8-13. https://doi.org/10.1109/MCSE.2010.113.
Zwaan, Rolf A., Alexander Etz, Richard E. Lucas, et M. Brent Donnellan. 2017. « Making Replication Mainstream ». *Behavioral and Brain Sciences*, octobre, 1-50. https://doi.org/10.1017/S0140525X17001972.