Veille et développement d'outils pour la recherche reproductible dans le cadre du journal académique Computo

Contexte

Computo (https://computo.sfds.asso.fr/) est un journal scientifique (ISSN 2824-7795) qui a pour objectif de promouvoir des contributions en statistiques et apprentissage (*machine learning*, ML) publié par la *Société Française de Statistique* (https://www.sfds.asso.fr/), qui s'inscrit dans une démarche de science ouverte et de reproductibilité des résultats scientifiques, en utilisant des avancées technologiques en programmation lettrée (*literate programming*) et en *reporting* scientifique. *Computo* a été créé dans un contexte de crise de la reproductibilité en science [1-5], ce qui appelle à une amélioration des méthodes de recherche [6,7] et des standards de publications scientifiques. Il est en accès libre en lecture ainsi qu'en publication (*open access* diamant [8]) et les rapports de relecture sont ouverts, i.e. visible et consultable librement après acceptation de la contribution.

Par ailleurs, assurer la reproductibilité des résultats numériques est une condition nécessaire pour publier dans *Computo*. En particulier, chaque soumission doit inclure toutes les données nécessaires (via un dépôt public comme Zenodo) et les codes sources des logiciels utilisés. Le contenu de l'article et les résultats présentés doivent être produits et être reproductibles grâce à l'utilisation de la programmation lettrée et de l'intégration continue^[Système généralement intégré aux forges logicielles permettant d'automatiser des traitements informatiques.]. Pour les contributions proposant des implémentations de méthodes/algorithmes, la qualité du code fourni est évalué lors du processus de relecture.

Sujet de stage

Afin d'améliorer le processus de publication, et notamment la phase de reproductibilité des résultats et du contenu de chaque article soumis, nous proposons un stage d'ingénieur de 4 à 6 mois. L'objectif est de mettre en place un système d'intégration continue pour la (re)production automatisée des articles destinés à la publication en ligne sur le journal. Le système devra être basé sur des outils libres, ouverts et hébergées dans des institutions françaises (comme une forge gitlab). En effet, le système actuel d'intégration continue utilise la forge Github pour ces processus, laquelle est d'une part non "open source" et d'autre part hébergée aux États-Unis. Par ailleurs, la plupart des outils impliqués repose sur des communautés actives et dynamiques de dévelopeurs, ce qui aménera assez probablement des interactions avec ces communautés.

Par ailleurs, les outils et processus que nous implémenterons seront mis à disposition de manière libre et ouverte, de sorte qu'ils pourront bénéficier à d'autres initiatives similaires dans la communauté scientifique autour des problématiques de science ouverte et de recherche reproductible.

Une autre piste d'amélioration du journal est son référencement, tant dans les bibliothèques (numériques) institutionnelles que via les systèmes propriétaires (Google Scholar, Scopus, Web of Science). Le stagiaire pourra être amené a documenter les mécanismes utilisés par ces systèmes et les intégrer à nos pratiques de publication.

Compétences requises

- Bonnes compétences en programmation (Python/bash, éventullement R/Julia),
- Expérience de travail dans un environnement de recherche reproductible (Git, versioning, intégration continue)
- Connaissances souhaitées en DevOps (Docker)
- Expérience avec (ou intérêt pour) l'intégration continue (Github actions, Gitlab CI/CD)
- Expérience avec (ou intérêt pour) les outils numériques de publication numériques (quarto, latex)

Environnement

Le stagiaire sera accueilli sur le Campus Agro Paris-Saclay, encadré par l'équipe éditoriale et technique de *Computo* (https://computo.sfds.asso.fr/).

Contact

computo@sfds.asso.fr (mailto:computo@sfds.asso.fr)

Références

1. Ioannidis, J P A 2005 Why Most Published Research Findings Are False. PLoS Medicine, 2(8): e124. DOI: https://doi.org/10.1371/journal.pmed.0020124

2. Steen, R G 2011 Retractions in the scientific literature: is the incidence of research fraud increasing? Journal of Medical Ethics, 37(4): 249-253. DOI: https://doi.org/10.1136/jme.2010.040923

3. Allison, D B, Brown, A W, George, B J, et Kaiser, K A 2016 Reproducibility: A tragedy of errors. Nature, 530(7588): 27-29. DOI: https://doi.org/10.1038/530027a

4. Bastian, H 2016 Reproducibility Crisis Timeline: Milestones in Tackling Research Reliability. URL https://absolutelymaybe.plos.org/2016/12/05/reproducibility-crisis-timeline-milestones-in-tackling-research-reliability/. [En ligne; consulté le 22-mars-2023]

5. Whitfield, J 2021 Replication Crisis. London Review of Books, 43(19). URL https://www.lrb.co.uk/the-paper/v43/n19/john-whitfield/replication-crisis. [En ligne; consulté le 22-mars-2023]

6. Desquilbet, L L, Granger, S, Hejblum, B, Legrand, A, Pernot, P, Rougier, N P, Castro Guerra, E de, Courbin-Coulaud, M, Duvaux, L, Gravier, P, Le Campion, G, Roux, S, et Santos, F 2019 Vers une recherche reproductible. Unité régionale de formation à l'information scientifique et technique de Bordeaux. URL https://hal.science/hal-02144142

7. The Turing Way Community 2022 The Turing Way: A handbook for reproducible, ethical and collaborative research. DOI: https://doi.org/10.5281/zenodo.7625728

8. Ancion, Z, Borrell-Damián, L, Mounier, P, Rooryck, J, et Saenen, B 2022 Action Plan for Diamond Open Access. DOI: https://doi.org/10.5281/zenodo.6282403