

Figures reproductibles bit à bit

Contexte

La reproductibilité d'une expérience scientifique est l'une des conditions nécessaires pour que des observations faites lors de cette expérience puissent entrer dans le processus d'amélioration perpétuelle des connaissances scientifiques. Malheureusement, beaucoup d'expériences scientifiques publiées dans des articles scientifiques ne sont pas reproductibles. Ce constat est souvent nommé crise de la reproductibilité. La science informatique n'y fait pas exception.

Les figures sont une partie très importante des articles scientifiques expérimentaux, puisqu'elles sont la manière principale de présenter une analyse de données aux lecteurs de l'article. Différentes piles logicielles capables de générer de telles figures existent. Malheureusement ces piles logicielles génèrent le plus souvent des figures au contenu indéterministe par défaut. Certaines de ces piles logicielles peuvent être configurées et exécutées de manière contrôlée afin de générer des figures au contenu visuellement déterministe, mais même comme ça, elle n'arrivent pas à générer de figures au contenu strictement identique quand elles sont exécutées sur des machines différentes.

Pouvoir générer des figures au contenu strictement identique est crucial pour automatiser différentes parties des traitements nécessaires à une expérience scientifique. En particulier on va s'intéresser ici à la génération de fichiers identiques bit à bit, qui ont l'avantage d'être extrêmement simples et peu coûteux à comparer (typiquement via le calcul d'un hash cryptographique de leur bits).

Objectifs du projet

Le premier but de ce projet est d'évaluer les capacités de reproductibilité bit à bit de différentes technologies populaires de génération de figures scientifiques. Des environnements logiciels complètement reproductibles (Nix) seront utilisés pour chaque technologie. Des exemples de figures simples seront implémentés dans chaque technologie. Différents formats d'images seront étudiés pour les figures, en particulier le format d'image matricielle PNG et les formats d'image vectorielle PDF et SVG. Différentes technologies d'isolation de l'exécution pourront être utilisées telles que des conteneurs ou des machines virtuelles.

Le second but de ce projet est d'améliorer les piles logicielles existantes pour les rendre reproductibles bit à bit. En particulier, la quasi totalité des technologies de génération de PDF sont indéterministes, il sera attendu de vous que vous identifiez le problème et que vous proposiez une solution pour le résoudre.

Candidature

Contactez-moi par mail avec les informations suivantes.

- Très courte motivation par rapport au sujet (2 phrases)
- Bulletins de notes (master et licence)
- CV court (parcours d'études, expériences professionnelles, compétences)