Conclusion

ANF Workflow et reproductibilité en Bioinformatique

Laurent Jourdren, Claire Toffano-Nioche, Thomas Denecker, and all the trainers

november 2025

Points d’intérêt des formations reçues

Organisers' point of view !

To enrich by the learners !!!

Niveau de reproductibilité : Repeat/Replicate/Reproduce/Reuse ⟹ Science Cumulative

Scientific Workflow Management et partage des Worfklows

- Vieux concepts qui existent depuis 15-20 ans
- Passe de GUI ⟹ code based
- Taverva ⟹ Nextflow & Snakemake
- Academique Info ⟹ Bioinformatique (Données Massives)
- Webservices ⟹ Bibliothèques (Conteneurs)
- Pb réseau/disparition service ⟹ Facile car réutilisation de code

Reproductibilité du code par la maitrise de l'environnement logiciel :

- Environnement Conda/Pixi (Spack/Guix)
- Containers : Docker/Apptainer

Outils basés sur le code en ligne de commande

- Nextflow
- Snakemake
- Galaxy (même s'il est un survivant de la première génération de moteur de workflow Galaxy est toujours actuel car il a pu beaucoup évoluer !)

⟶ Ces outils (Nextflow/Snakemake/Galaxy) fournissent des fonctionnalités similaires et il y a une émulation entre les projets

L'important, c'est de partager avec la communauté afin de ne pas dupliquer les efforts et d'améliorer les workflows existants

Les workflows permettent de

- s'affranchir de l'infrastructure sur laquelle ils tournent (local, Cluster, Cloud)
- simplifier le passage du dev à la production

NF-core : reflète cette importance de la communauté : Contribute and participate ! (grâce à Nf-core : Nextflow semble prendre le pas sur Snakemake et Galaxy)

Grâce à l'IA, on peut facilement migrer son workflow vers les nouvelles versions du DSL Nextflow et se conformer aux bonnes pratiques

Importance:

- des bonnes pratiques (carte de métro, formatage de code, changelog, documentation, ...)
- de commencer le plus tôt possible à partager son workflow pour agréger une communauté
- des tests pour s'assurer que le workflow fonctionne dans le temps et résiste aux changements des versions des outils tiers qu'il utilise
- de travailler sur les données de référence (domaine peu pris en compte actuellement). Ex.: calcul des index des génomes modèles pour tous les mappeurs et toutes leurs versions (réduction de consommation de ressources)

▸ Knowledge Graphs are instrumental for FAIR principles

By-design, built for being both human and machine-readable

⟶ Interoperability ✅

▸ Semantic Web technologies provide open and standard protocols: URLs / HTTP / RDF format / SPARQL query language (W3C standards)

⟶ Findability ✅ (URIs for identifying things on the web)

⟶ Accessibility ✅

⟶ Interoperability ✅

Enrichir le code en métadonnées pour le rendre plus FAIR

- Ontologies are community-aggreed controlled vocabularies

→ Interoperability ✅

→ Reuse ✅

Ces registres/catalogues (e.g. bio.tools, workflowHub, etc.) sont importants car ils exposent automatiquement ces métadonnées pour nous

EDAM and FAIR-Checker

EDAM: A domain-specific ontology for Bioinformatics for annotating

- usage context of bioinformatics tools (EDAM Topics)
- what does the tool (EDAM Operations)
- input and output data (EDAM Data, EDAM Formats)

FAIR-Checker: permet d'estimer la qualité des métadata des sites web:

1. en évaluant les principes FAIR
2. en vérifiant la réutilisation de vocabulaires contrôlés / ontologies (e.g. DCterms, Schema;org, EDAM ...)
3. en identifiant des métadonnées importantes pour certains type de ressources (profils Bioschemas) mais qui pourraient manquer

Mission : **sauvegarde du code** ; robustesse financière (Unesco + boites privées)

Equivanlance DOI pour du code : **SWHID** (SoftWare Hash) persistent et intrinsèque (hash SHA calculé à partir du code source déposé)

SH archive volumineuse : 27 milliards de fichiers, 400 millions de projets ⟹ compression avec dédoublonnage (ex. fichiers de LICENSE toujours les mêmes, enregistrés qu'une fois)

récolte automatique depuis les dépôts usuels (GitHub, GitLab, ...) et de paquets (CRAN, Bioconductor), archive toutes les releases/tag annotés
Cependant maj de GitHub trop rapide ⟹ retard dans la récolte de SH
savegarde personnelle possible : https://archive.softwareheritage.org/save
fréquence : pas à chaque commit mais à chaque release/tag

Quel code déposer ? tous ! se décomplexer à partir du moment où c'est du code pour la science ! Si pas de licence associée : archivage ok mais pas de réutilisation ...

Il faut des métadonnées : pour les humains (README, AUTHORS, ...) et pour les machines (...)
Extension dans le navigateur: vert (déjà enregistré et dernière version) jaune (pas la dernière version), gris (absent), rouge (dépôt logiciel privé)

SWHID : utilisé dans les publis + HAL (avec le "dir"), chercher le bouton "permalink" sur le navigateur

- Absence de Licence == on ne sait pas comment utiliser le code == on ne peut donc pas le réutiliser
- Semantic commit == convention pour les commits git
- Release: Major.Minor.Patch
- Automatic/Semantic release: si Semantic commit alors possible de faire des publications automatiques des releases, patchs, documentation dans le changelog, etc, nécesite juste un fichier `gitlab.ci.yml`

Parfois les noms de releases se basent sur la date au moment de la release, apporte la notion de "fraîcheur" de la release, est-ce une bonne pratique ? perte des "niveaux" (major,minor,patch), possible de faire les 2

⟹ cf. sur le moodle en bonus : bcp de liens (github, gitlab, ...)

Generative AI

Based on

M.Aliouat, C.Cadiou , … A.Thomas 2024

Recommandations INRAE pour l'usage de l'IA générative comme assistant personnel

Large Language Model (LLM): result of a process that enables text prediction based on statistics (from a huge volume of texts).

Various techniques used to control the responses/texts returned by the LLM:

- **fine-tuning**: additional but lightweight learning process to specialize responses in a field
- **Retrieval Augmented Generation** (RAG) or **prompt-engineering**: allows users to ask questions in natural language (no computational cost). The prompts from a session are cumulative (the prompt size increases): the business model == based on the number of words
- ...

Typical usages

- equivalent of a web query but in natural langage
- writing assistance (time saving)
- generate or verify computer codes

Note on next advantages and risks sessions:

rapidly evolving field $\Longrightarrow$ limited comprehensiveness

Advantages and new opportunities

- increased **productivity**
- identify **new hypotheses**, create protocols, conduct experiments (e.g. AlphaFold)?
- increase **accuracy** by the amount of information taken into account?
- **learning** aid tool
- enable new **interactivity** with computer and automatic systems, e.g. control machines (agricultural, drones, etc.) through human-machine dialogue
- communicate with each other, e.g. to negotiate contracts for access to information (increased **interoperability**)

Risks associated with exploitation

- need of a considerable **amount of data**
- **environmental** impact during the construction of the foundation model, lighter during operation (but much more).
- **social & ethical** impacts, alienation of part of the community (exploitation of underpaid workers during training phase)
- degree of **openness** of the model, few information on data or biases in the corpus (data in English, biased towards American culture)
- **sovereignty issue** when used for research, impact on **reproducibility** (service interruption)
- massive increase in article production $\Longrightarrow$ longer peer-reviewing times $\Longrightarrow$ **decrease in the production of scientific articles**

Risks associated with use

- factual errors (**hallucination**), based on probabilities ⟹ flexibility associated with errors; LLM does not understand either the question asked or the answer it gives (what is plausible from a probabilistic point of view could be wrong). Result stated without any hint of doubt ⟹ check!
- **outdated** information (date of training not specified)
- **legal**: **loss of novelty** required for intellectual property rights, **counterfeiting** without the user's knowledge (data submitted under licenses or copyright), loss of confidentiality of **sensitive data** (personal data present in the models but also provided by the user)
- reproduction of **discriminatory social norms** present in learning data
- **scientific integrity**: total delegation of tasks to AI, loss of skills in the long term

INRAe recommandations

- continue monitoring LLM technology and its uses as personal assistants
- develop expertise in evaluating LLM
- master LLM specialization: not improving foundation models (only a few companies can achieve this) but refining them (fine-tuning or RAG)
- controlling data flow (training, specialization but also usage), sensitive/ personal data $\Longrightarrow$ encouraging the use of LLMs locally
- raise awareness, train, and support users: the quality of the response depends on the prompt; the user becomes responsible for the result and must ensure compliance with regulations
- need to structure AI governance due to the risks associated with its use