Taking tools out of their laboratory: Methods and practices to make Bioinformatics tools accessible through Galaxy

Valentin Marcon1, Alexis Dereeper2, Sarah Maman-Haddad2, Melanie Petera4, Luc Jouneau2, Marie Tremblay-Franco6, and Olivier Inizan2

A.R.T. Accessibility, Reproducibility and Transparency are the guidelines of the web-based platform Galaxy. These three principles have lead the Galaxy Community to create Good Practices for tool integration into their platform. Today, a developer has clear guidelines to wrap a tool under Galaxy (xml file, packaging).

FRENCH COMMUNITY The French community has been among the early users of Galaxy. Several tools wrappers have been developed since 2010. Wrappers have a heterogeneous non-standardized architecture because the stabilization of methods took a consequent time. After this significant stabilization time, the French community has adopted the Galaxy good practices.

BEST PRACTICES They have been established by the community to improve the quality of wrappers development. Respect of these practices provides a harmonization of tools architecture, an easy tool maintenance and an insurance of the tool quality for users.

PROGRESS For the past 21 months we have set up methods for the overall project. And we principally applied it on 4 use cases (subprojects).

TECHNOLOGIES In front of this process we apply some technologies commonly used by the community:
- Planemo, an assisting software for Galaxy tools building. It eases development by providing wrappers skeletons (template) and checking the good syntax of the xml files. Also, it can automate tools testing and publication on the toolshed.
- Conda, a dependencies and environment management technology used by Galaxy developers to manage tools packaging. It is now the privileged way for dependencies management in Galaxy last versions. There is also a conda repository for all the biological tools packages named “BioConda”.
- Docker, a container manager, that provide light virtual machines. The interest of this technology is to easily share tools in a close environment, ready to use.

RESULTS The first use case is the suite “REPET” to detect, annotate and analyze repeats in genomic sequences (specifically designed for transposable elements). For this tool we finalized wrappers that allow the two “REPET” pipelines to run on Galaxy (in a lighter way). Difficulties came from the large number of dependencies and the need of a HPC cluster for these pipelines. We succeed to install it on galaxy instances on the IFB cloud and on a URGI virtual machine. We have therefore carried out a training on the URGI galaxy machine, it allows us to validate the tool, and make a list of new features to improve the current version. We made a second development cycle, with the delivery of a new version of REPET. This version has been successfully used in a second training session.

The second use case “BIOS4BIOL” was a set of wrappers, build by members of a statistical working group at INRA. We standardized the wrappers and created a new one that achieve normalization on data, a common feature of the old wrappers. We worked with the original developers of the tools and to facilitate our collaborations we made a training on the versioning tool “git”. We also have managed dependencies with Conda packages.

The tools are now available in the “Genotool” Galaxy instance and in the Galaxy Toolshed.

DELIVERABLES For each use case we have 4 main deliverables: the packaged tools (wrapper for Galaxy and packages to install dependencies), an instance easily accessible for users; a use case with someone having a good knowledge of the tool; documentation and/or training support ready to use.

WHAT'S NEXT We are currently working on a new use case “miRDeep2”. It is a pipeline made from the miRDeep2 tool but never finished. We are installing it on a Test Galaxy instances and then we will reproduce the work done for the other project that is to: ensure the wrappers follow the best practices; be sure they can be shared with Conda package dependencies; publish them on the Toolshed; Install them on a Galaxy instance and maybe create a container and organize a training about it. And for SNIPay we are managing some others pipelines like the Haplophylies analysis one.