Optimization of a Galaxy pipeline in the Magerit supercomputer for analysis of the sugarcane microbiome

Loïc BOURGEOIS

April 2015

Previously...

What has been done

- Enabling the Galaxy toolshed for Magerit (chroot)
- Enabling heavy files upload through ftp
- Implementing Nuria's mothur workflow on galaxy

What I wanted to do

- Populate the Galaxy instance with more tools
- Allow users to access data libraries in a simple way
- Improve Mothur's Galaxy wrappers

Things don't always go as expected

Sample of errors

- List index out of range error on some tools installation
- Mothur was giving segmentation faults while trying to rerun some analysis
- Galaxy would not accept Amalia's email, for some reason

Things don't always go as expected

martenson: loic_bourg: how old is your Galaxy?
loic_bourg: martenson : its nearly one year old (revision is c458a0fe1ba8+)
martenson: the file that throws the exception does not even exist in the distribution now :(

Galaxy reinstallation

Pros

- Compatibility with newer tools
- Maintenance
- New functionnalitys
- Debugging (narrow down errors)

Cons

- Possible incompatibility with Magerit installation
- Could take time

Main steps of the Galaxy reinstallation

- Back up the old version
- Install and configure the newest version of Galaxy
- The newer version of Galaxy implements a CLI job runner for SLURM which needed to be tweaked
- All the software installed along Galaxy did not need to be reinstalled

In the end

- Reinstalling Galaxy was actually easier than what I expected
- It solved errors I was getting on some tools installation
- It will be easier to keep the Galaxy instance up to date
- Mothur does not give anymore the segmentation faults that I could not explain
- A mothur's pipeline has been successfully ran for Amalia's data

New tools

- Preprocessing tools : Trimmomatic, FastX-Toolkit, etc.
- Tools for RNAseq analysis from the Tuxedo protocol[4]: TopHat, Cufflinks, etc.
- Plotting tools...

Data libraries

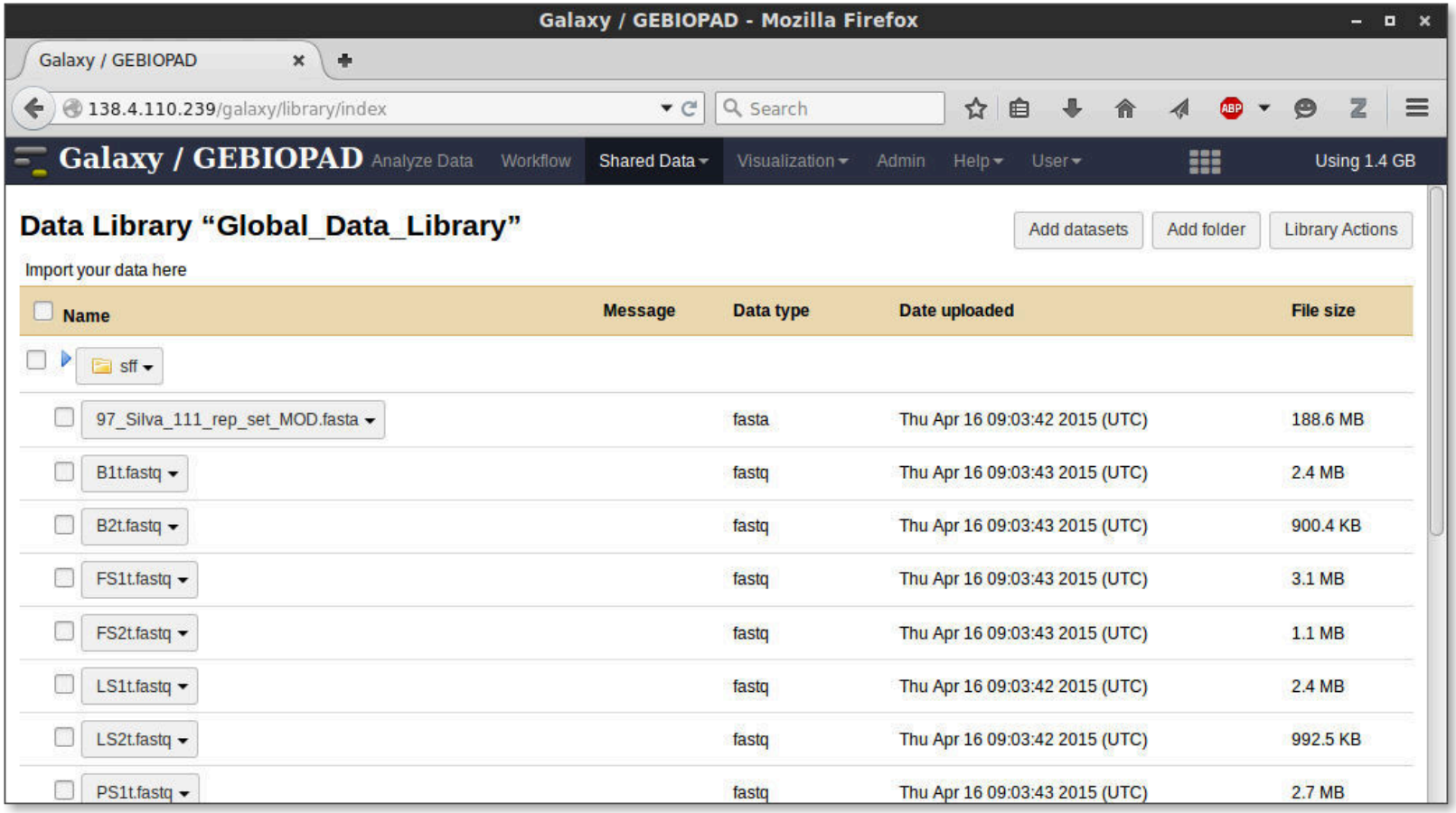

Figure : An example of a data library

Why is it better?

Benefits

- Users can store and organize their data easily in them [2]
- It is possible to reimport easily the data for future analysis
- Data can be linked into galaxy history, instead of copied
- It stores the files "physicly" in a single folder

But...

Only Galaxy adminitrators can create them

Workaround

After discussing with the Galaxy developers, it appears that the only way to avoid requiring an admin is to use the galaxy API.

What is an API?

GUIs are nice and user friendly, but it is impossible to automatize tasks with them. An API (Application Programming Interface) is an interface allowing to access programmatically to a given software.

The galaxy API and Bioblend

Galaxy offers two ways to do so :

A classic REST API

- Classic HTTP requests
- Complete
- Heavier in term of scripting

Bioblend[1][3]

- Python wrapper on the REST API
- Not all the REST API functions are implemented
- Easier to use

Workaround

First solution that worked, but with some problems :

One Data Library for everyone with the right permissions

- Each new user in Galaxy would need to have permissions added to the data library automatically which would require a script running in background
- If a user do mistakes while importing a file, and admin would be required to fix it

Second solution, which I adopted :

Write a Galaxy Tool allowing the user to create a data library using the API

- The user has full control on his data library
- Lighter in term of ressources, as the script is launched only when needed by the user
- Was not working until yesterday because of a "bug"

Data libraries for everyone

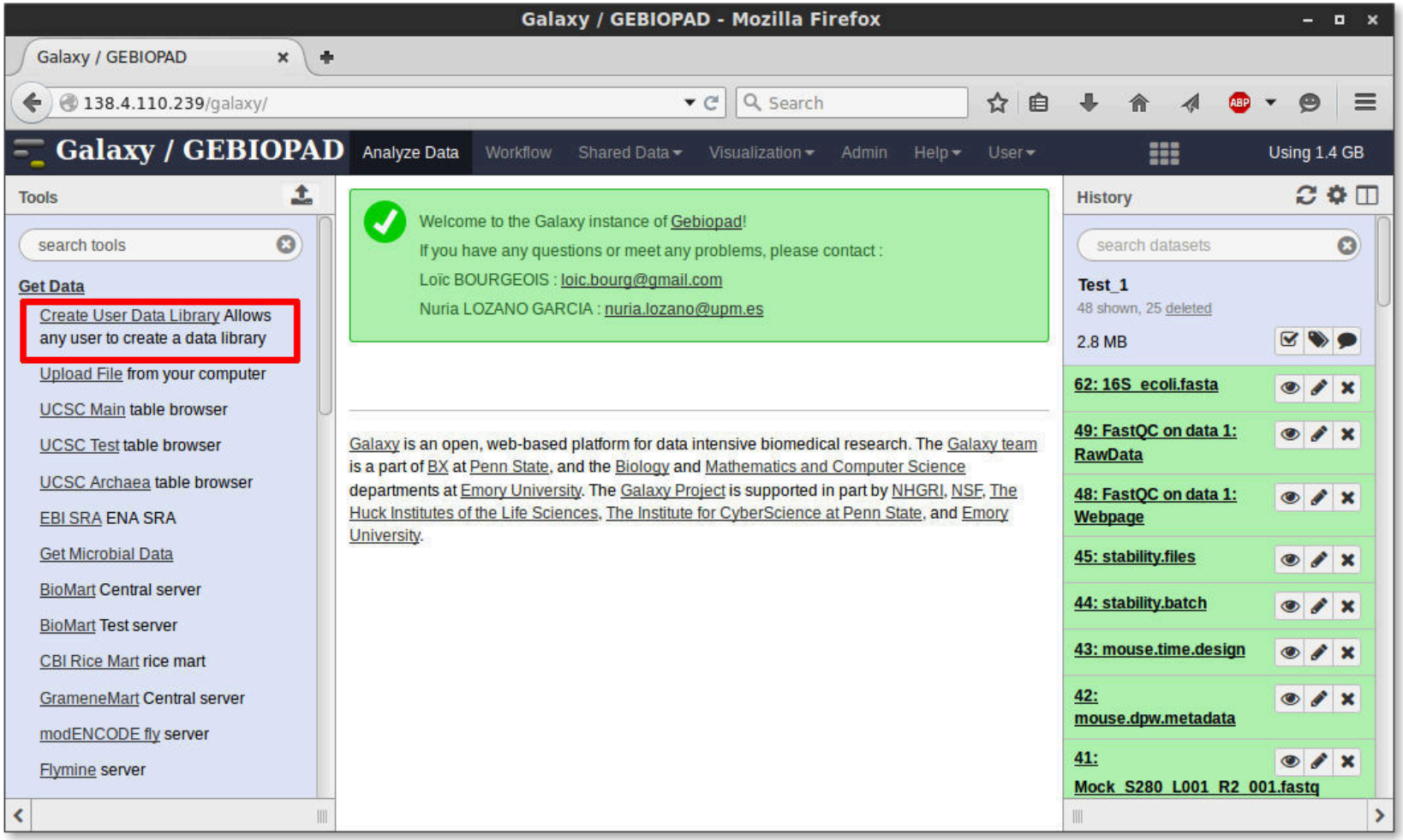

Figure : New tool allowing data library creation

What's next...

- Compare QIIME and Mothur's pipelines [3]
- Setup maintenance scripts and cron jobs
- Enable the alignment visualisation Tool on Galaxy
- Test the Tuxedo protocol in the galaxy instance for RNA seq analysis

Thank you for your attention!

References

[1] Simone Leo, Luca Pireddu, Gianmauro Cuccuru, Luca Lianas, Nicola Soranzo, Enis Afgan, and Gianluigi Zanetti. BioBlend.objects: metacomputing with Galaxy. Bioinformatics (Oxford, England), 30(19):2816–2817, October 2014.

[2] Luisa Maria Seoane, Omar Al-Massadi, Mary Lage, Carlos Dieguez, and Felipe F. Casanueva. Ghrelin: from a GH-secretagogue to the regulation of food intake, sleep and anxiety. Pediatric endocrinology reviews: PER, 1 Suppl 3:432–437, August 2004.

[3] Clare Sloggett, Nuwan Goonasekera, and Enis Afgan. BioBlend: automating pipeline analyses within Galaxy and CloudMan. Bioinformatics (Oxford, England), 29(13):1685–1686, July 2013.

[4] Cole Trapnell, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R. Kelley, Harold Pimentel, Steven L. Salzberg, John L. Rinn, and Lior Pachter. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature Protocols, 7(3):562–578, March 2012.