Galaxy-E: Ecological data analysis, citizen science and biodiversity indicators production!

@ColineRoyaux #PNDB @Yvan2935

Coline Royaux

Yvan Le Bras

yvan.le-bras@mnhn.fr
Context – We need Atomization

Currently, in ecology ...

One R script for one input datafile
Context – We need Atomization

Currently, in ecology ...

One R script for one input datafile

With Galaxy...

Several atomized R scripts for several input datafiles
Context – We need Sharing & Generalization
Context – We need Sharing & Generalization

<table>
<thead>
<tr>
<th>carre</th>
<th>annee</th>
<th>espece</th>
<th>abond</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2016</td>
<td>ACCGEN</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>2017</td>
<td>ACCGEN</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>2018</td>
<td>ACCGEN</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>2019</td>
<td>ACCGEN</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>183</td>
<td>ACCGEN</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>183</td>
<td>ACCGEN</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>183</td>
<td>ACCGEN</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>183</td>
<td>ACCGEN</td>
<td>0</td>
</tr>
</tbody>
</table>

Toolshed
And some trainings! => https://training.galaxyproject.org/

Climate
Learn to analyze climate data through Galaxy.

You can view the tutorial materials in different languages by clicking the dropdown icon next to the slides (⚠) and tutorial (🔍) buttons below.

Requirements
Before diving into this topic, we recommend you to have a look at:

- Introduction to Galaxy Analysis

Material

<table>
<thead>
<tr>
<th>Lesson</th>
<th>Slides</th>
<th>Hands-on</th>
<th>Recordings</th>
<th>Input dataset</th>
<th>Workflows</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction to climate data</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Functionally Assembled Terrestrial Ecosystem Simulator (FATES)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Functionally Assembled Terrestrial Ecosystem Simulator (FATES) with Galaxy Climate ksysfinlab</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Getting your hands-on climate data</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rangeo ecosystem 101 for everyone - Introduction to Xarray Galaxy Tools</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rangeo Notebook in Galaxy - Introduction to Xarray</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Visualize Climate data with Panoply netCDF viewer</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
And some trainings! => https://training.galaxyproject.org/

**Climate**
Learn to analyze climate data through Galaxy.

You can view the tutorial materials in different languages by clicking the dropdown icon next to the slide and tutorial buttons below.

**Requirements**
Before diving into this topic, we recommend you to have a look at:
- Introduction to Galaxy Analysis

<table>
<thead>
<tr>
<th>Material</th>
<th>Slides</th>
<th>Hands-on</th>
<th>Recordings</th>
<th>Input dataset</th>
<th>Workflows</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction to climate data</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Functionally Assembled Terrestrial Ecosystem Simulator (FATES)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Functionally Assembled Terrestrial Ecosystem Simulator (FATES) with Galaxy Climate Kangaroot</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Getting your hands-on climate data</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rangeo ecosystem 101 for everyone - Introduction to Xarray Galaxy Tools</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rangeo Notebook in Galaxy - Introduction to Xarray</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Visualize Climate data with Panoply netCDF viewer</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Ecology**
Learn to analyze Ecological data through Galaxy.

You can view the tutorial materials in different languages by clicking the dropdown icon next to the slide and tutorial buttons below.

**Requirements**
Before diving into this topic, we recommend you to have a look at:
- Introduction to Galaxy Analysis

<table>
<thead>
<tr>
<th>Material</th>
<th>Slides</th>
<th>Hands-on</th>
<th>Recordings</th>
<th>Input dataset</th>
<th>Workflows</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biodiversity data exploration</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Compute and analyze biodiversity metrics with PAMPA toolsuite</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Metabarcoding/eDNA through Obitools</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RAD-Seq de-novo data analysis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RAD-Seq Reference-based data analysis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RAD-Seq to construct genetic maps</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regional GAM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Species distribution modeling</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Galaxy-E Killer workflows

• Biodiversity exploration tools
• Biodiversity metrics & indicators production
• Dealing with GIS and netcdf files on Galaxy-E
Biodiversity exploration tools

- Abundance of Molluscs of Australian east coast in the environment
- rarefaction curves
- correlation analysis
- survey_date distribution
Biodiversity exploration tools
Dealing with GIS and netcdf files

Fouilloux et al. EGU22 Pangeo for everyone with Galaxy

A “Classical” data processing:
Sampling sites information in GIS data file (often shapefile)
Environmental information in netCDF file

Create a file with environmental information on sampling sites!
Visualize maps of environmental parameters on sampling sites

Until now: R + QGIS + a lot of manual manipulation
Now: a Galaxy workflow mixes scripts, GDAL & Xarray tools making it easily accessible and (re)-runnable.
Galaxy open-source platform for FAIR data analysis offers:

- Pangeo notebook deployment (local dask) available to everyone (free registration);
- Pangeo Galaxy Tools for fully automated workflows;
- GUI for users with no programming skills;
- Self-Paced Learning material and organisation of online training events with the Galaxy Training Network;
- Training Infrastructure as a Service is a free and ready to use with private queues where only training's jobs run.

Pangeo for everyone through Galaxy

Galaxy open-source platform for FAIR data analysis offers:

- Pangeo notebook deployment (local dask) **available to everyone (free registration);**
- Pangeo Galaxy Tools for fully **automated workflows**;
- GUI for users with **no programming skills**;
- Self-Paced **Learning material** and organisation of online training events with the Galaxy Training Network;
- **Training Infrastructure as a Service** is a free and ready to use with private queues where only training’s jobs run.

Amazing basis for eosc | FAIR-EASE

PNDB / Galaxy-E work for next months

Essential Biodiversity Variables workflows

Kissling et al. 2017
EBV workflows: STOC

1. Preprocess data
2. Filter data
3. Analyze community indexes
4. Analyze species abundance

Biodiversity data

Community

Species - population

EBV workflows: PAMPA

Existing accessible & reusable Galaxy tools
convert / concatenate / Column Regex Find and Replace / Merge Columns / Filter / Count / Regex Find and Replace / Advanced Cut

Data selection

Biodiversity data

1. Compute community metrics
   - Community
   - Pre-processed data
   - Species - population
   - PAMPA

2. GLM on community metrics
   - Metric ~ site + year + habitat
   - Time-series plot from GLM results

3. GLM on population metrics

www.ifremer.fr/pampa
Essential Biodiversity Variables workflows

Help BONs to identify gaps & reuse EBV workflows
Essential Biodiversity Variables workflows

Global Open Science Cloud (GOSC)
Case Studies

Help BONs to identify gaps & reuse EBV workflows

Amazing basis for EBVOSC

PNDB / Galaxy-E work for next months

Do you think this can help create a national biodiversity network on your country (Germany, Australian, ..) ? Contact us!
Not only for data analysis

=> Also for Research data management
Ecological research data management

Import Biodiversity occurrences data

Generate metadata

Convert metadata

Upload data an metadata
Ecological research data management

**Import Biodiversity occurrences data**

**Convert metadata**

**Amazing basis for PNDB / Galaxy-E work for next months**
Not only asynchronously

=> Thanks to Galaxy interactive tools GxIT
One specific workflow goal

Galaxy as workflow engine

workflows (the async ones)

Interactive tools

- Interactive Jupyter Notebook
- GPU enabled Interactive Jupyter Notebook for Machine Learning
- Interactive Climate Notebook
- Interactive Pangeo Notebook
- RStudio
- Pyron Interactive Jupyter Notebook
- HiGlass an interactive Hi-C data visualizer.
- QRCode Workbench Working with messy data
- Ubuntu XFCE Desktop
- Panoply interactive plotting tool for geo-referenced data
- AskOmics a visual SPARQL query builder
- Interactive CellXgene Environment
- bam.lobio visualisation
- VCF ( lobio ) Visualisation
- Neo4j ( Graph Database )
- Phinch Visualisation
- Paraview
- Wilson Webbased Interactive Omics visualization
- Wallace Webbased Interactive modeling of species niches and distributions
- geoexplorer An interactive spatial analysis platform using ggvgs and Leaflet
- radiant Data analytics using Radiant R Shiny app
- EtherCalc
- VRM Editor interactive tool for creating Variable Resolution Mesh for NetESM/CEMS
- SimText Interactive shiny app to explore SimText output data
- ISEE
- metashark Metadata Shiny Automated Resource and Knowledge
One specific workflow goal

Galaxy as workflow engine

workflows (the async ones)

Amazing for R Shiny apps!

Cofest « R Shiny » subject?
One specific workflow goal

Galaxy as workflow eng **What is making this guy ?**
One specific workflow goal

Galaxy as workflow engine

What is making this guy?

Analysing audio files on Galaxy!
Not only for scientists

- Crowdsourcing through Galaxy webhooks
- Data / Biodiversity literacy through Galaxy-Bricks
Galaxy-Bricks

Galaxy for pupils!

bricks.vigienature-ecole.fr
Galaxy-Bricks

Galaxy for pupils!

bricks.vigienature-ecole.fr

Amazing basis for Cofest « vue.js » GUI?
Crowd sourcing through Galaxy

MOODA concept (Massively Open Online Data Analysis)

Crowdsourcing with hoverflies (syphres) images from SPIPOLL project

GAPARS project

63 320 classifications in 2.5 years!
Crowd sourcing through Galaxy

MOODA concept (Massively Open Online Data Analysis)

Crowdsourcing with hoverflies (syphres) images from SPIPOLL project

63 320 classifications in 2.5 years!
Crowd sourcing through Galaxy

MOODA concept (Massively Open Online Data Analysis)

Crowdsourcing with hoverflies (syphres) images from SPIPOLL project

63 320 classifications in 2,5 years!

Amazing basis for

Cofest « citizen science » webhook?
Crowd sourcing through Galaxy

MOODA concept (Massively Open Online Data Analysis)

https://tinyurl.com/galaxymooda
Thank you!

PNDB team

Coline Royaux – engineer R / Galaxy dev (workflows to compute biodiversity indicators)

Elie Arnaud – engineer R Shiny / knowledge – metadata dev

Marie Jossé– engineer R / Galaxy dev

Julien Sananikone – engineer DevOps / sys admin / web dev

Olivier Norvez – animation coordinator

Yvan Le Bras – Beta tester

https://www.pndb.fr/

PNDB « bricks »:
MetaShARK Metadata work:
https://youtu.be/OViSMzRgtw
Data metadata portal:
https://youtu.be/STwsYDHet2A
Galaxy Europe demo:
- https://youtu.be/HeIAHgjX6D4
- Essential biodiversity variables on Galaxy: implementing PAMPA
- Producing biodiversity indicators from citizen science projects

https://www.pndb.fr/
yvan.le-bras@mnhn.fr