Reproducible Open Science with EGI Notebooks and Replay services

Giuseppe La Rocca
Community Support Team Lead. / EGI Foundation

24 March 2023
ISGC 2023

TLP: GREEN Limited disclosure
Vision
All researchers have seamless access to services, resources and expertise to collaborate and conduct world-class research and innovation

Mission of the EGI Federation
Deliver open solutions for advanced computing and data analytics in research and innovation

Mission of the EGI Foundation
Enable the EGI Federation to serve international research and innovation together

EGI Council Participants:
- 29 countries
- +5 international research organisations
- +1 institutional representative
International Partnerships with MoUs

- GEANT Association
  - RENAM (Moldova)
  - GRENA (Georgia)
  - IMCS UL (Latvia)

- Open Science Grid (USA)
  - CLAF (Latin America)
  - Compute Canada

- CSIR Meraka Institute (South Africa)

- SSTIR (China)
  - CNIC (China)
  - IHEP (China)
  - ASGC (Taiwan)
EGI-ACE – EGI flagship project

Consortium:
- Coordinator – Stichting EGI
- 33 Partners, 23 third parties
- Most of EGI members + Several RIs

Services:
- EGI Services for Research
- EGI Services for Federation
- EGI Services for Business

Scope:
- 49% service delivery (Virtual Access)
- Co-development of services with RIs

Duration:
- Jan 2021 – June 2023 (30 months)
Reproducibility: beyond sharing code and data

Peng, Science, 2011
Reproducible Open Science in EGI/EOSC

1. Notebooks
2. DataHub
3. Share
4. GitHub
5. Zenodo
   - DOI: 10.5281/zenodo.3469564
6. Reproduce
7. Notebooks + Binder
8. DataHub
The Jupyter Notebook is an open-source web application that allows you to **create and share documents** that contain live code, equations, visualizations and narrative text.

Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.
Documentation
Text formatted using Markdown/LaTeX

Interactive
browser based environment

Code
Use your favourite language

Output
Results of the code execution(e.g. plots)
Jupyter hosted in the EGI Cloud
- Offers Jupyter notebooks ‘as Service’
- One-click solution: login and start using

Main Features:
- Easy access: Login with the EGI AAI Check-In service
- Persistent storage for notebooks
- Use EGI computing and storage resources from your notebooks
1. Runs on EGI Cloud providers (Cloud Compute / Online Storage)
2. Uses Check-in for authentication + access to other services
3. Makes user-level software available from CVMFS
   - DIRAC client for submission of jobs to Workload Manager / High Throughput Compute
   - Fedcloudclient for interaction with Cloud Compute
4. Transparent access to Datahub spaces
5. Access to 3rd party services: B2DROP
Customisable options

EGI Notebooks offers different computing environments:

- **Default**: Python, R, Julia, Octave and a wide range of data science libraries
- **MATLAB (Basic/Full)**: Run MATLAB on EGI resources (requires a license!)
- **Community environments**: tuned to meet the needs of specific user communities

Server Options

- **Default EGI environment – 6 GB RAM / 2 core**
  The Default notebook environment includes Python, R, Julia and Octave kernels

- **MATLAB Environment (Basic) – 4GB RAM / 4 cores**
  The MATLAB environment (requires a valid license), includes Python and MATLAB kernels

- **MATLAB Environment (Full) – 4GB RAM / 4 cores**
  The MATLAB environment (requires a valid license), includes Python, MATLAB kernels and additional MATLAB packages

B2DROP connection

Start
Accessing data and code

- **Persistent home**
  - Can be used to store data (10GB limit)
  - Files will be kept even if the notebook server dies
- **DataHub**
  - Access to your accessible spaces in datahub
  - Share data and assign PIDs to shared spaces
- **B2DROP**
  - Access data stored in EUDAT’s B2DROP service
- **CVMFS**
  - Selected CVMFS repositories available
  - Easy to add community specific ones
- **nbgitpuller**
  - Get code from any git repository from a single URL
  - [https://hub.jupyter.org/nbgitpuller/link.html](https://hub.jupyter.org/nbgitpuller/link.html)
1. Get an EGI account (https://aai.egi.eu/signup)

2. Enroll to one of the supported VOs (Virtual Organisations):
   • vo.notebooks.egi.eu
   • vo.access.egi.eu
   • auger
   • biomed
   • vo.reliance-project.eu
   • eiscat.se
   • vo.environmental.egi.eu (https://go.egi.eu/8Hspz)
Binder: reproducing execution environments

An open-source web application to turn repositories in interactive notebooks

It uses Modern technology in cloud orchestration (Kubernetes), interactive computing (Jupyter), scientific computing (the open-science ecosystem)
BinderHub hosted by EGI

- Offered ‘as Service’
- Same access conditions as EGI Notebooks

Main Features:

- Use any binder-compatible repository
- Reproduce your notebooks with access to EGI resources (e.g. datahub)
- No hard limits on sessions duration, customisable resource limits for users/communities
A code repository that contains:

- Code to reproduce (i.e. set of notebooks)
- + description of the software runtime (e.g. a conda environment)
- + any auxiliary files needed to run the code

From https://github.com/enolfc/isgc-2023-enes
Making repositories citable

Turn a Git repo into a collection of interactive notebooks

Have a repository full of Jupyter notebooks? With Binder, open those notebooks in an executable environment, making your code immediately reproducible by anyone, anywhere.

Build and launch a repository

Zenodo DOI (10.5281/zenodo.3242074)

Zenodo DOI ➜ 10.5281/zenodo.7124666

Git ref (branch, tag, or commit) Path to a notebook file (optional)
HEAD Path to a notebook file (optional)

Copy the URL below and share your Binder with others:


Expand to see the text below, paste it into your README to show a binder badge.
EGI Notebooks and EGI Replay services

1. Perform data analysis and visualisation

2. Publish notebook and Generate DOI

3. Cite DOI in publication

4. Discover DOI

5. Resolve DOI to repository

6. Recreate environment

7. Reproduce analysis

EGI Notebooks

EGI DataHub

EGI Replay

GitHub

Your repository

zenodo

DOI: 10.5281/zenodo.3242074

Fellow researchers
Start using the services!

https://notebooks.egi.eu/
https://replay.notebooks.egi.eu/

Documentation: https://docs.egi.eu/users
Support: support@egi.eu
ISGC 2023 - 21st. March at 14:00
EGI Tutorial: Regional infrastructure for reproducible open science in Asia Pacific

June 19th – June 23rd 2023, Poznań, Poland
Call for Contributions is open!
https://indico.egi.eu/event/6071/abstracts/
Deadline extension to March 30th.
Thank you

Giuseppe La Rocca

e giuseppe.larocca@egi.eu

www.egi.eu

This work is partially funded by the EU research
and innovation programme