Reproducible Open Data analysis with Binder and DataHub

Andrea Manzi, Giuseppe La Rocca, Enol Fernández
Reproducibility: beyond sharing code and data

Reproducibility Spectrum

Publication only

Publication +

Code

Code and data

Linked and executable code and data

Full replication

Not reproducible

Gold standard

Peng, Science, 2011
Reproducible Open Science

- The computational **tools** to solve a problem
  - Python, R, Julia, and wide ecosystem of libraries and tools for science
- An **interface** to facilitate coding/creating
  - Jupyter
- A way to **communicate** your work
  - Notebooks
- Leverage on the EGI Cloud to **scale-up the resources**
- A way to **share** your work
  - GitHub, Zenodo or other similar repositories
- A way to **pack** it all for replication
  - Docker (used by Binder)
- A way to **persistently identify** it
  - DOIs (Digital Object Identifiers)
Reproducible Open Science with EGI

1. Perform data analysis and visualisation
2. Publish notebook and Generate DOI
3. Cite DOI in publication
4. Discover DOI
5. Resolve DOI to Notebook
6. Recreate environment
7. Reproduce analysis

EGI Notebooks → GitHub → Your repository → zenodo

Fellow researchers

EGI Notebooks ← EGI DataHub ← EGI Notebooks

DOI 10.5281/zenodo.3242074
Binder

- 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)
What does Binder do?

**GitHub**

- **repo2docker**: Creates reproducible containers from repositories
- **jupyterhub**: Generates user sessions that serve these containers
- **Kubernetes**: Manages the computing infrastructure

**EGI Binder**: Provides an interface to create, share, and share the sessions

Pulls code from repository
Zenodo

- ‘Open Science’ repository from OpenAIRE
  - Hosted at CERN data centre
  - [http://zenodo.org](http://zenodo.org)
- For
  - Publications, white papers, datasets, software, etc.
- For
  - Long-tail science
  - Projects
  - Communities
  - Research groups
  - ...
- Dropbox, GitHub integration:
  - Web or API access
- Access stats and reports for communities
Demo time
Thank you for your attention.

Questions?