Data discovery, analysis and reproducibility in Virtual Research Environments

Enrique García[1], Giovanni Guerrieri[1], Rubén Pérez[1], Michael Zengel[1], Georgy Skorobogatov[2], Andrés Tanasijczuk[3], Hugo Gonzalez[1] and Xavier Espinal[1]

[1] CERN, [2] ICCUB Barcelona, [3] UCLouvain

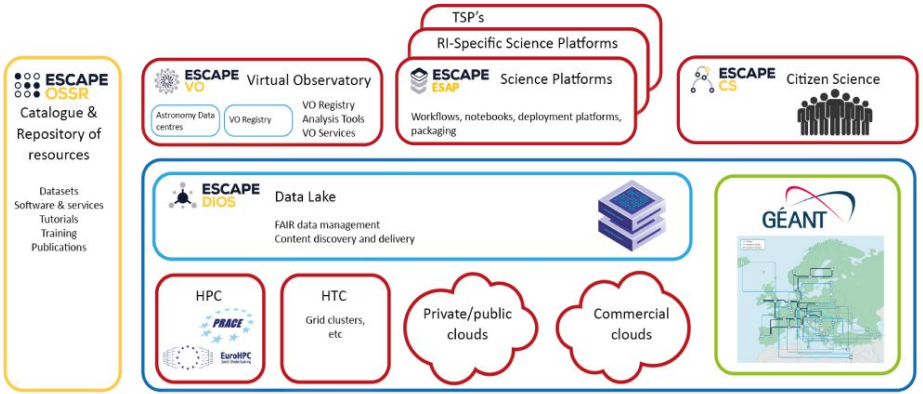
CHEP 2024 | Track 9 | 24 October

Outline

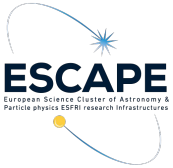
- Context
- The CERN Virtual Research Environment
- Demo video
- Summary & next steps

Context - ESCAPE EU Project

- Starting in 2019, the ESCAPE collaboration brought together various physics communities (ESFRIs, RIs, ERIC...)
  - Tackle and solve common computing issues
    - Data Management
    - Software and Service repository
    - Science platforms
    - Citizen Science
    - IVOA
  - From an Open Science perspective
- Project ended in 2023 → ESCAPE Open Collaboration
  - In addition to main project outcomes
  - Various initiatives and communities were naturally established



Consolidate ESCAPE outcomes



- It is fundamental to keep the momentum and communication to foster collaboration on common projects.

- Cascading grants via OSCARS project.
  - Understand the outcome of the first grant's call.
  - Put communities together to build a common basic roadmap.
  - Identify and aggregate specific community needs.
  - Both within a cluster (ESCAPE) and a between clusters (PaNOSC, SSHOC, ENVRI-FAIR, EOSC-Life).

Disclaimer

What this presentation **IS** going to show

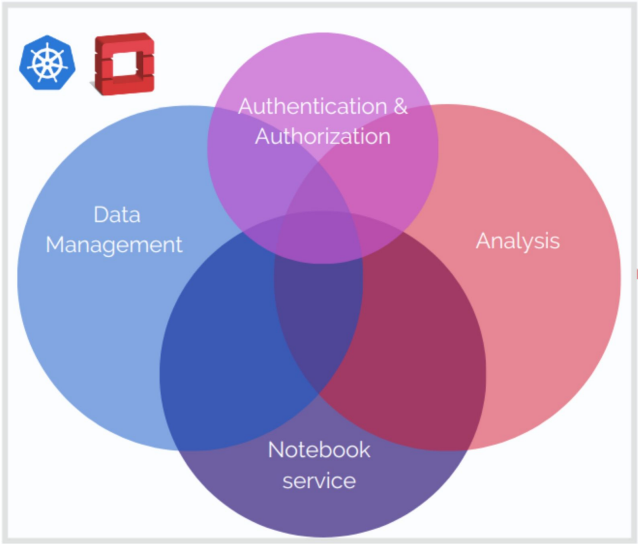
- One of many analysis facility "concepts"
- Results of a bottom-up approach from (ESCAPE) user needs
- Built fully Open Source

What this work **IS NOT** trying to do

- Solve everybody's needs (collaborations, experiments etc...)
  - Might not be compatible with certain computing architectures (bottom-up vs top-down).
- Provide an unique / monolithic solution

CERN Virtual Research Environment

- ESCAPE outcome
  - The place where most of the data-related R&D activities happened.
  - PoC platform provisioning different solutions for users:
    - Built on top of a common scientific tool: Jupyter
    - Software environment & repositories
    - Connection with resources: EOS, CVMFS
    - Ease interacting with underlying services



- To date CERN VRE acts as a **middleware interface** allowing scientific users to access different services and computing resources.

CERN Virtual Research Environment

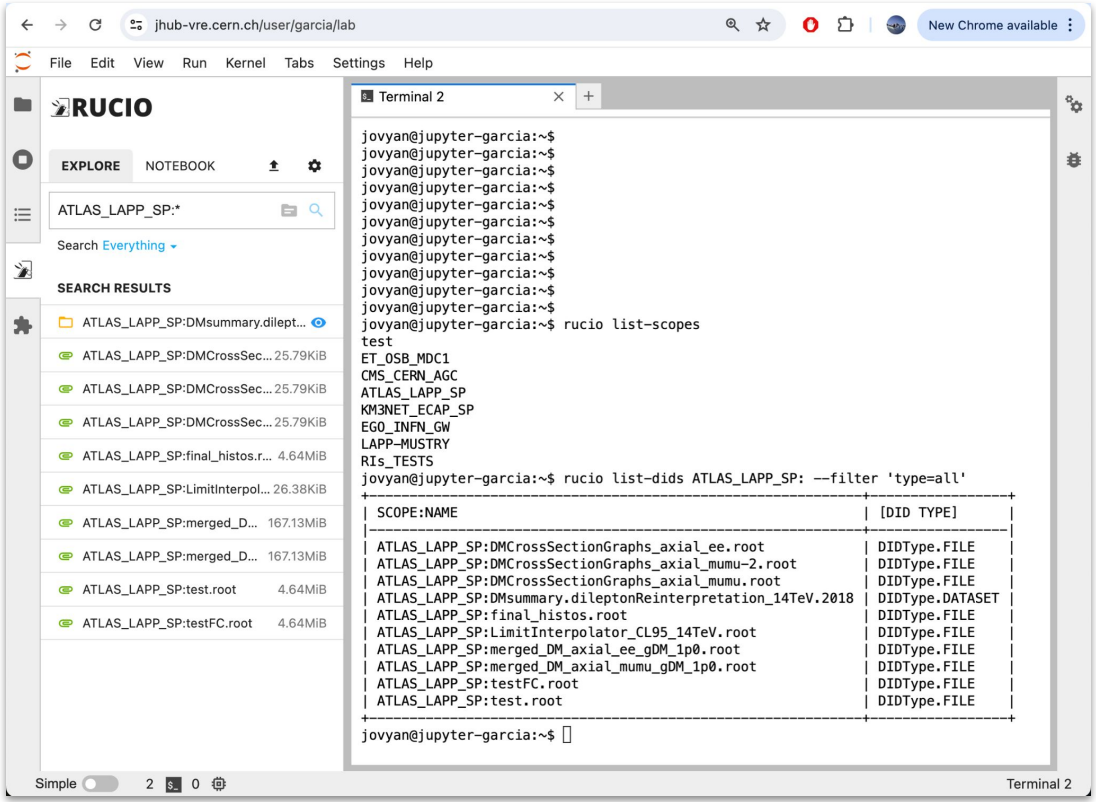
- K8s cluster using OpenStack @ CERN
- Services installed via helm
  - ESCAPE Rucio Instance - ESCAPE Data Lake
  - Reana VRE instance
  - JupyterHub
- AAI: ESCAPE Indigo IAM
- Pre-loaded software environments via Docker images
- CVMFS
- EOS and CEPHFS storages

Scientific analysis stages

+

VRE@ CERN

Services accessible via the CERN VRE



Scientific analysis stages

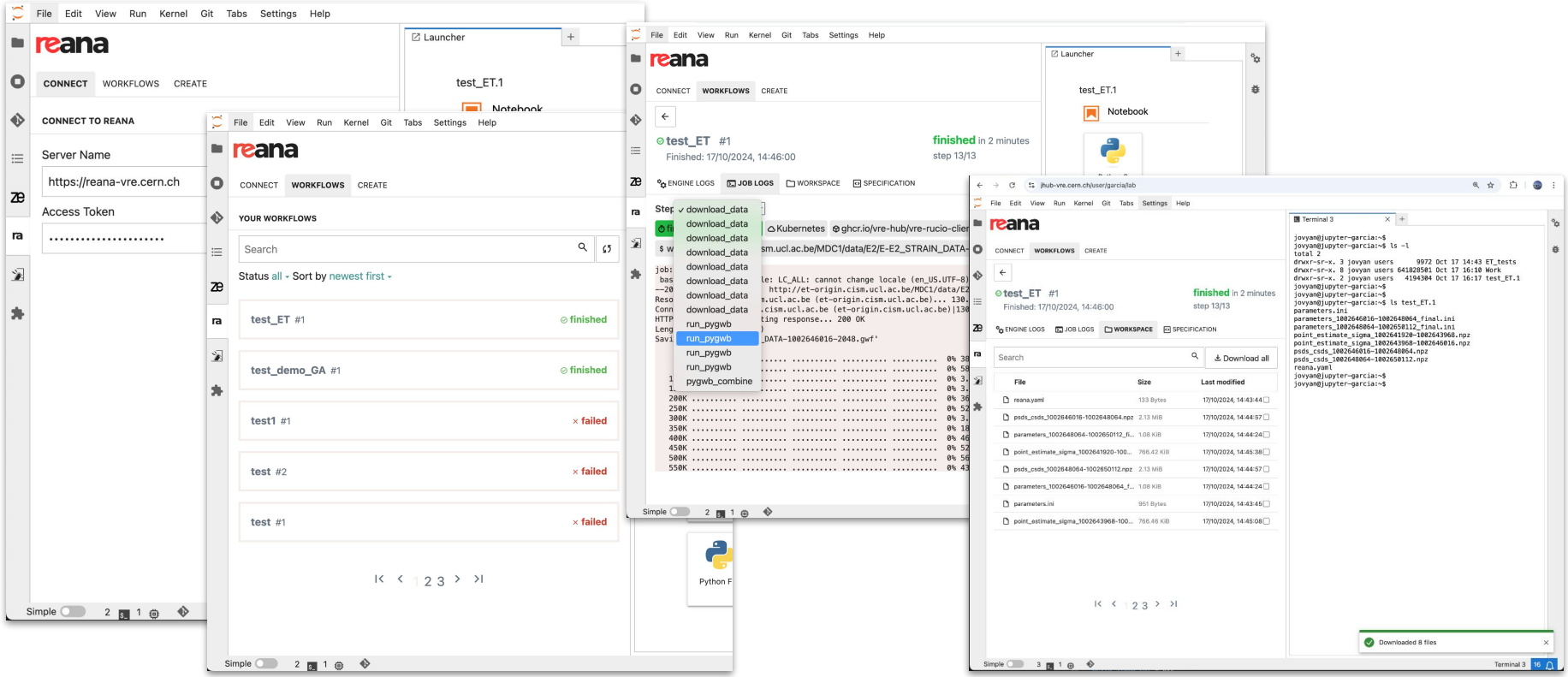
+

VRE@ CERN

RUCIO

Data discovery → Data access → Analysis of data ↔ Data reproduci-bility ↔ Data preserva-tion

Services accessible via the CERN VRE



Scientific analysis stages

+

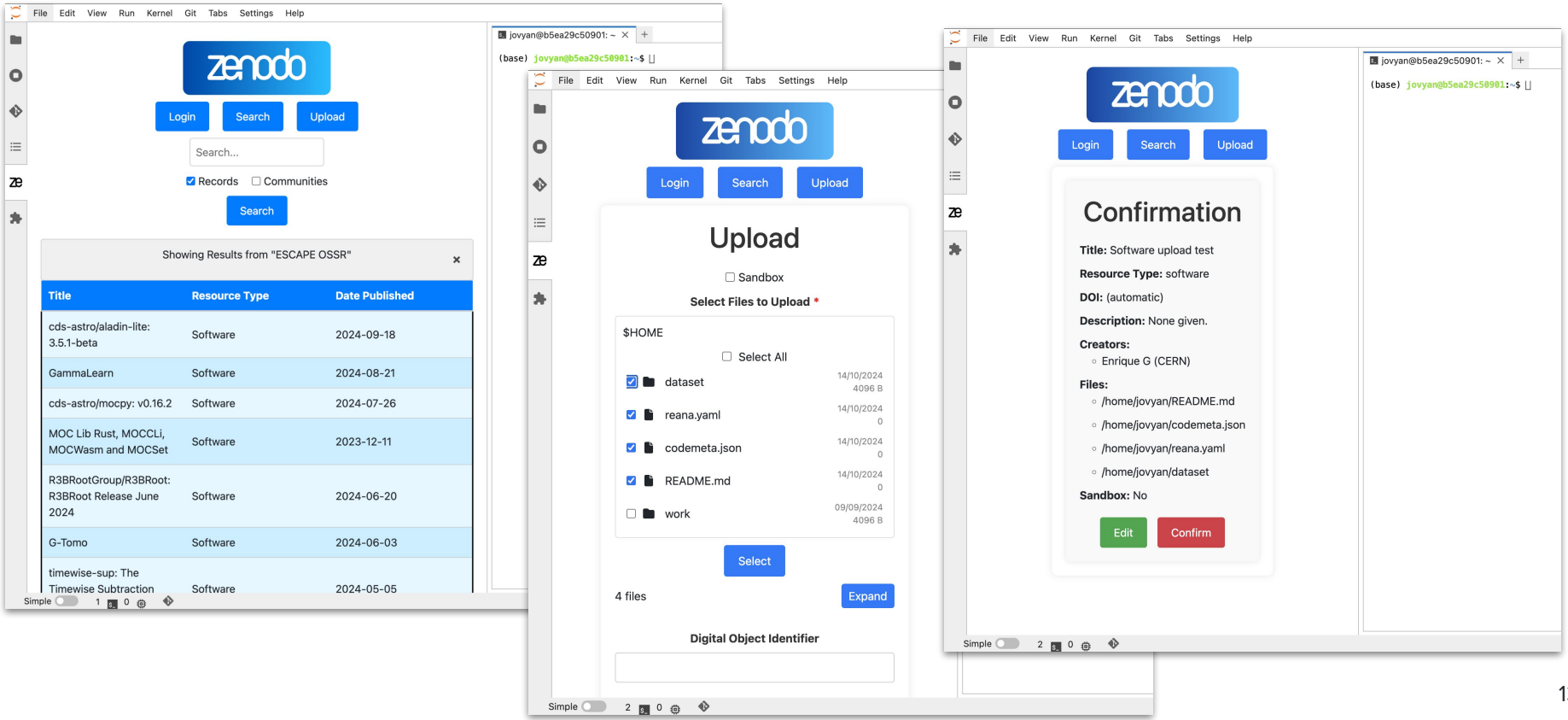
VRE@ CERN

RUCIO

reana

Data discovery → Data access → Analysis of data ↔ Data reproduci-bility ↔ Data preserva-tion

Services accessible via the CERN VRE



Scientific analysis stages can be performed in the VRE

+

VRE@CERN

RUCIO

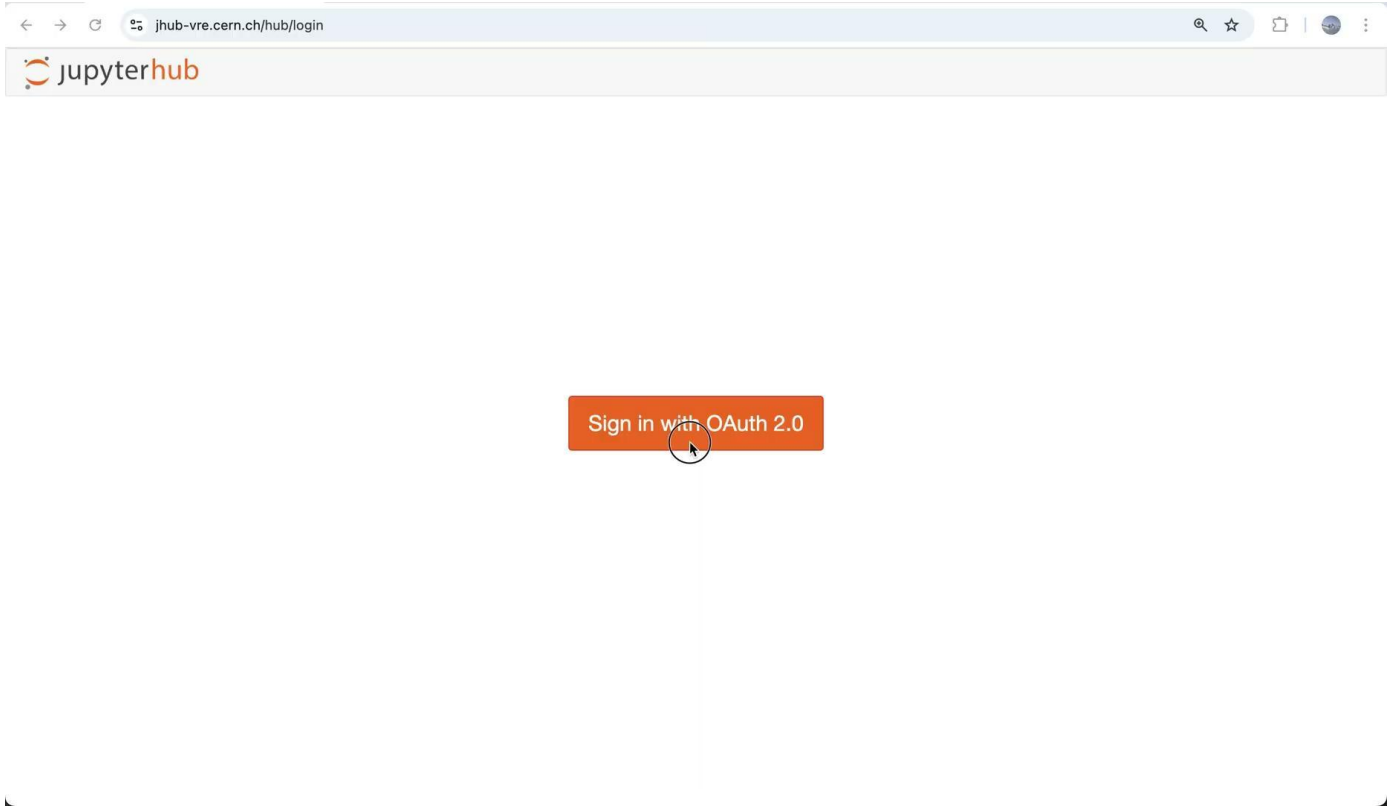
reana

zenodo

Data discovery → Data access → Analysis of data ↔ Data reproduci-bility ↔ Data preserva-tion

Demo (video x1.5)

permanent link



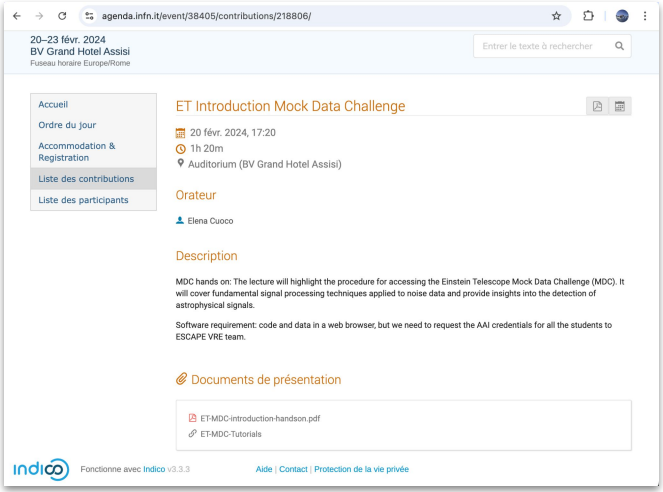
Summary

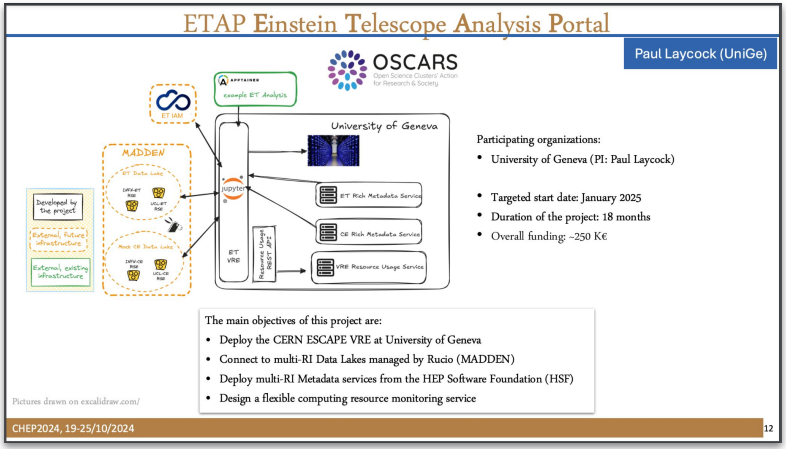
- CERN VRE presents an Analysis Facility concept
  - Allows performing most of the stages of a complete scientific analysis
  - All from the localhost of a Jupyter session - compatible with a remote / cloud-based architecture
    - Accessing and discovering data via the **Rucio** Plugin
    - (re)Triggering a workflow on a **Reana** instance
    - Publishing results on **Zenodo**

- Fully modular
  - Can help establishing the basic tools for certain communities

Future of the CERN VRE

- Demonstrator will continue
  - As a PoC and test place for other communities
  - As a place for R&D activities
  - Potential Open Science platform
    - ET Workshop Feb '24
- Maintain the community active and provide support / expertise
  - Lia L. Track 1 presentation on behalf ET community





Thanks for your attention

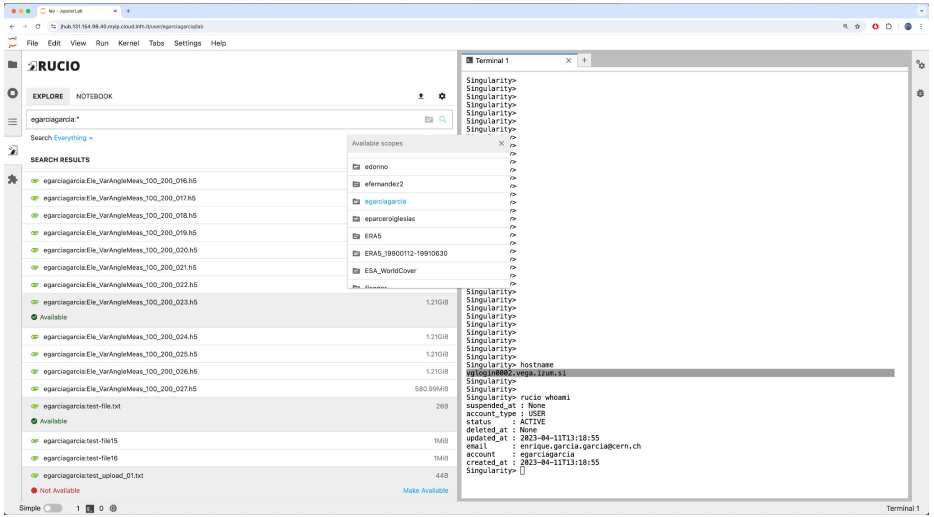
enrique.garcia.garcia@cern.ch; escape-cern-ops@cern.ch

VRE links:

- Access to the VRE https://jhub-vre.cern.ch/
- VRE documentation https://vre-hub.github.io/
- GitHub organisation https://github.com/vre-hub

Back up slides

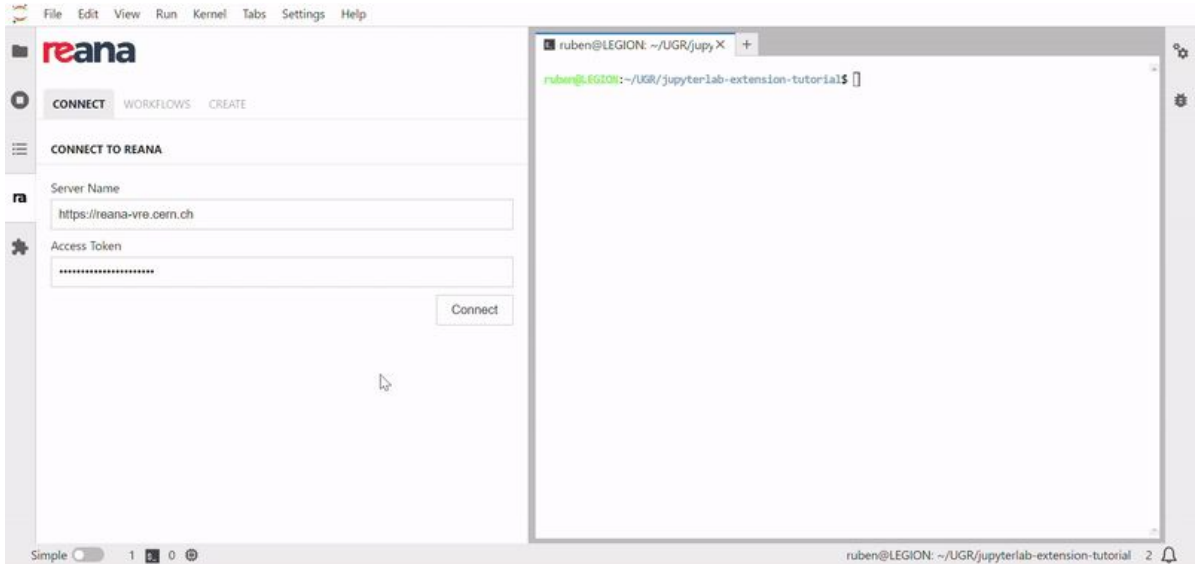
Rucio JupyterLab Extension



- Official Rucio component
- Compatible both
    - jupyterlab v3
    - jupyterlab v4
- Integrated into interTwin - EU project infrastructure
- Code: https://github.com/rucio/jupyterlab-extension
- Documentation: https://vre-hub.github.io/docs/extensions/rucio-jupyterlab/

Reana Extension

- 2024 Summer Student project

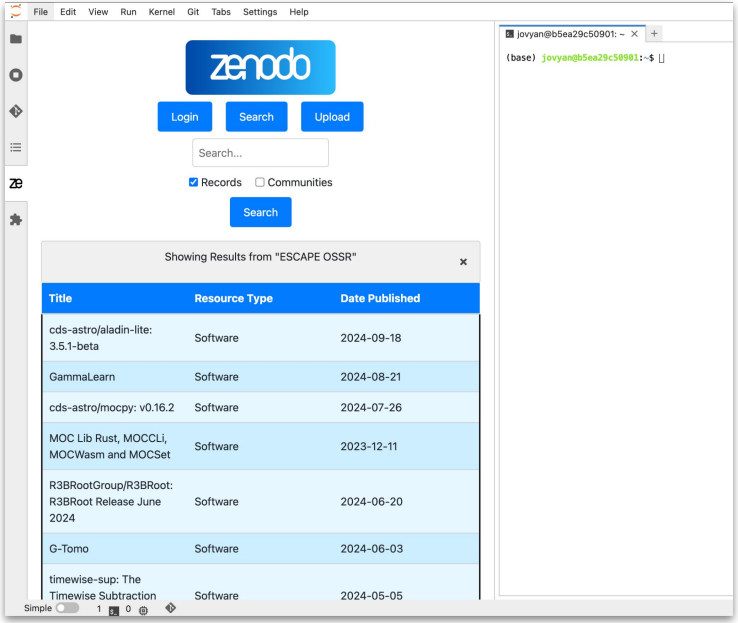


pypi v1.0.0

DOI 10.5281/zenodo.13366643

- Code: https://github.com/vre-hub/reana-jupyterlab-extension
- Documentation: https://vre-hub.github.io/docs/extensions/reana-jupyterlab/

Zenodo JupyterLab Extension



- 2024 Summer Student project
  - Few functionalities still under dev
  - V1.0.0 before the end of the year
- Code: https://github.com/vre-hub/zenodo-jupyterlab-extension
- Documentation: https://vre-hub.github.io/docs/extensions/zenodo-jupyterlab/