PAUL SCHERRER INSTITUT

PSI

SDSC

Rok Roškar & Krisztian Pozsa

Integrating computational workflow systems and data repositories

2023

Agenda

- Data Repositories
- Computational Workflow Systems
- What is a Workflow?
- User Stories
- Key Challenges
- Solutions

Data Repositories

- Data repositories in scientific research
  - Data organization and documentation
  - Data storage and backup
  - Data sharing and collaboration
  - Data quality assurance
  - Data security and privacy
  - Data preservation and long-term access
  - Data publishing

- Multiple data repositories in use today within the ETH domain
  E.g.: Zenodo, SciCat, EnviDat, OpenBIS, ETH Research Collection, ERIC

Computational Workflow Systems

- Computational workflow systems in scientific research
  - Organize and automate data analysis, simulation, and modeling
  - Facilitate tracking and documentation of data use
  - Automate execution of computational tasks
  - Allow to adapt and modify workflows easily
  - Enable collaboration
  - Manage computational resources
  - Visualize workflows and monitor execution status

- Multiple computational workflow systems in use today within the ETH domain
  E.g.: Renku, AiiDA

What is a Workflow?

Input — E.g.: proposals, instruments, raw datasets

Process — E.g.: researcher's notes, data transformations

Output — E.g.: derived datasets, visualizations, research papers

User Stories

- A researcher **develops a pipeline** on Renku to construct a dataset, which is public and published as a paper; she wants to directly **export the dataset** to the **ERIC repository** with all the associated metadata together with **information about the workflow** that generated the dataset; she wants other interested researchers to be able to **explore the dataset** within a **compute environment** with a click.
- A researcher **publishes an ML dataset** through the **ETH Library Research Collection**; the dataset is **very large** (many TBs) and difficult to move around; he wants to be able to **mount the dataset** within **interactive compute sessions** so collaborators and others can have **hands-on access** to the methods and data and **reproduce his results**.

Key Challenges

- **Limited publication methods**: data repositories often lack standardized and user-friendly methods for publishing computational workflows
- **Documentation complexity**: current methods often require researchers to manually document and annotate workflows in data repositories
- **Lack of interoperability**: limited integration between computational workflow systems and data repositories makes it difficult to identify previous uses of data or find relevant workflows
- **Fragmented sharing methods**: in the absence of easy-to-use data repositories for workflows researches resort to other sharing methods

Solutions

- **Import/export datasets:** enable data exchange between data repositories and computational workflow systems
- **Interoperable workflows**: extend the computational workflow systems so they can export/import their workflows to a common format
- **Publishing workflows**: create direct integrations between the data repositories and computational workflow systems to share workflows and data provenance metadata and to motivate and enable data re-use