Recording data provenance from sample to workflow results with LabID

Laurent Thomas

PostDoc

Data-Science Center - Multimodal Open Data Integration Support (MODIS)

EMBL Heidelberg, Germany

EMBL

The EMBL : European Molecular Biology Laboratory

- 6 sites across 5 countries
- **Diversity of research techniques** and instruments (x-ray, light and electron microscopes, sequencers…)
- Strong **bioinformatics expertise** and **computational infrastructure** (slurm cluster)

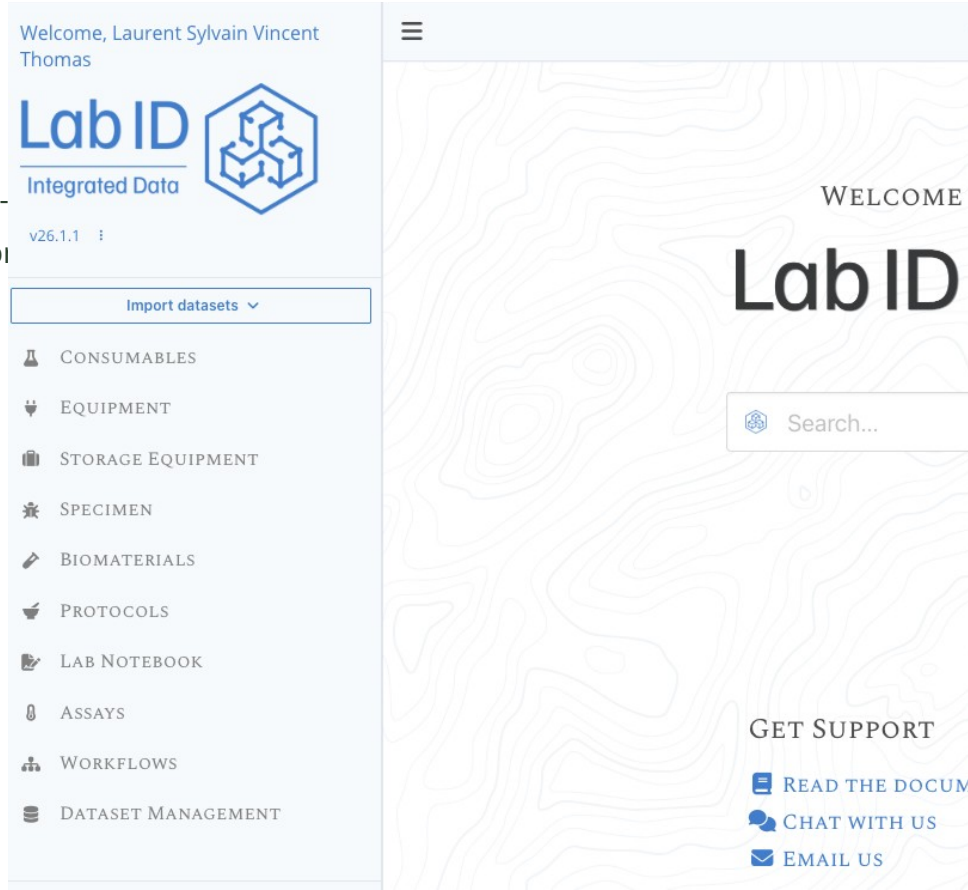
EMBL Data-Science Center

- Training (git, bash..)
- Consulting for data-management / analysis
- Development and maintenance of services : JupyterHub, Galaxy, GitLab, **Lab Integrated Data (LabID)**…

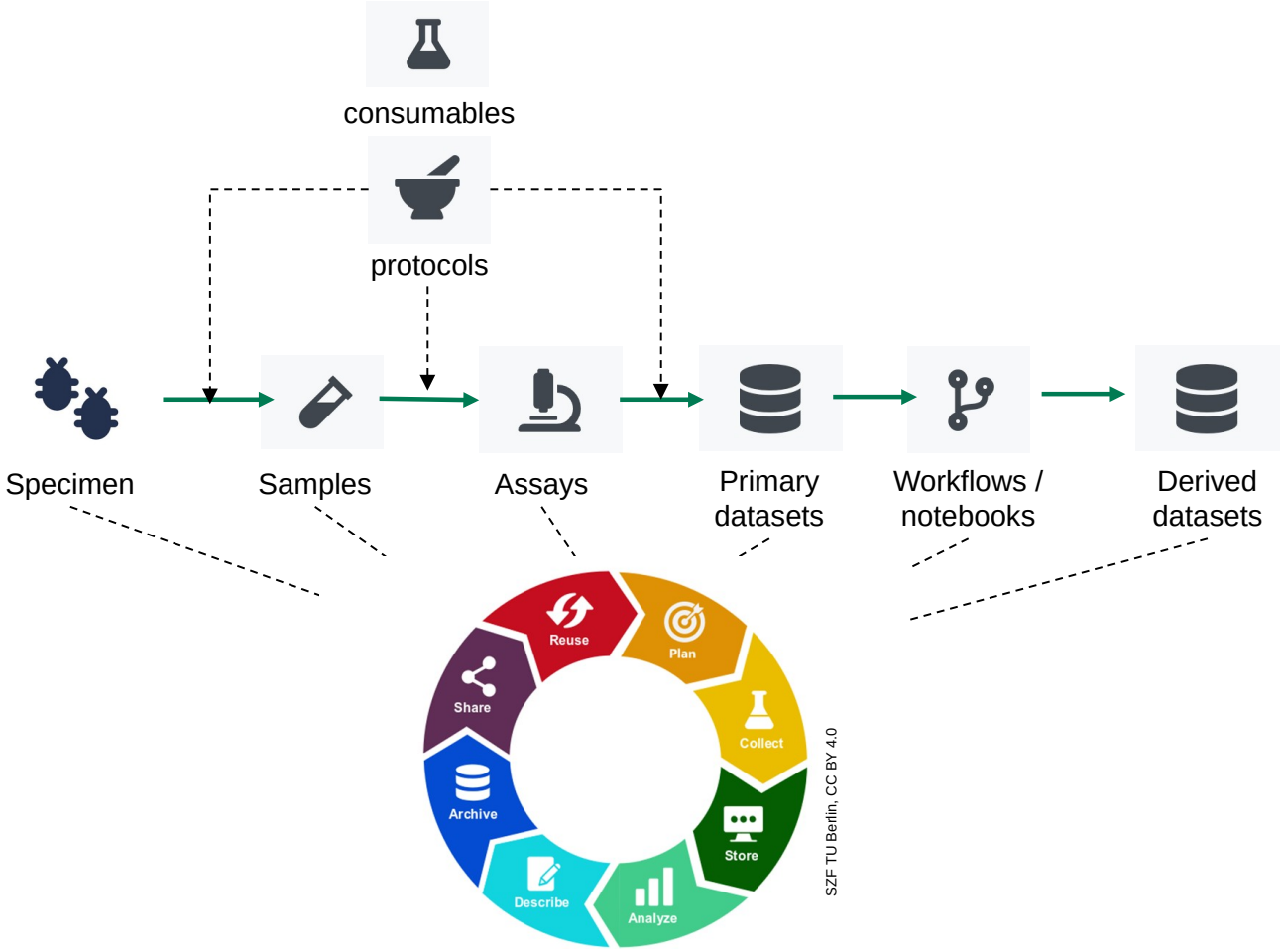
EMBL

What is LabID ?

- “All-in-one” FAIR **data management platform** for **life-science** institutes and research groups
- Collect **data and metadata,** interconnect it and share
- Developed at the EMBL Heidelberg
- **Used across EMBL sites**, expanding to other institutes (open-
- Tailored to life-science but many generic data management pr



LabID : a platform to capture metadata along the data lifecycle



ENA European Nucleotide Archive · PRIDE · BioStudies · MetaboLights · BioImage

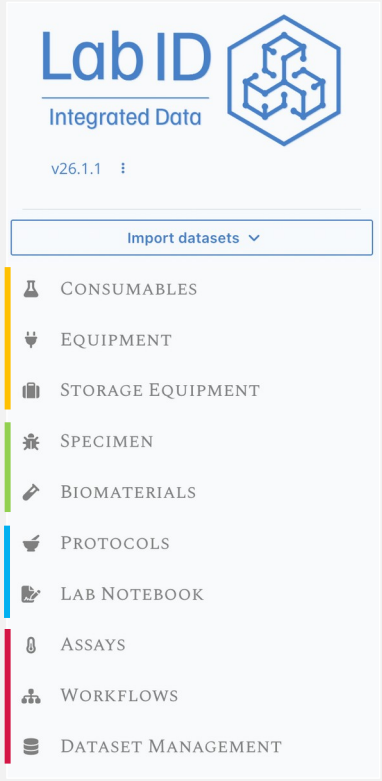
Share with Collaborators

Track information in a lab

Consolidate datasets (AI)

EMBL

4 interconnected modules



Lab Collection Management
(Consumables, Equipment, Storage)

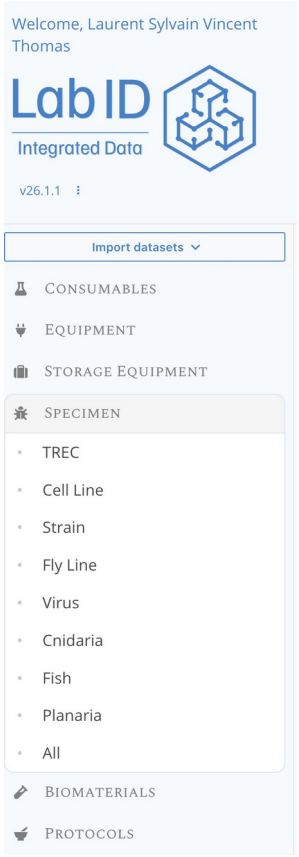
Dataset Management
(Assays, Datasets, Workflows & Studies)

Electronic Lab Notebook (ELN)
(Project, Experiment & Protocols)

Biomaterial Management
(Sample & Specimen, *Biobanking*)

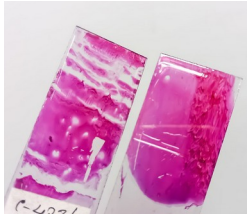
EMBL

Example : tissue imaging assay

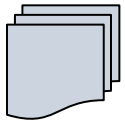


Specimen
Ex : Mouse line *xxx*

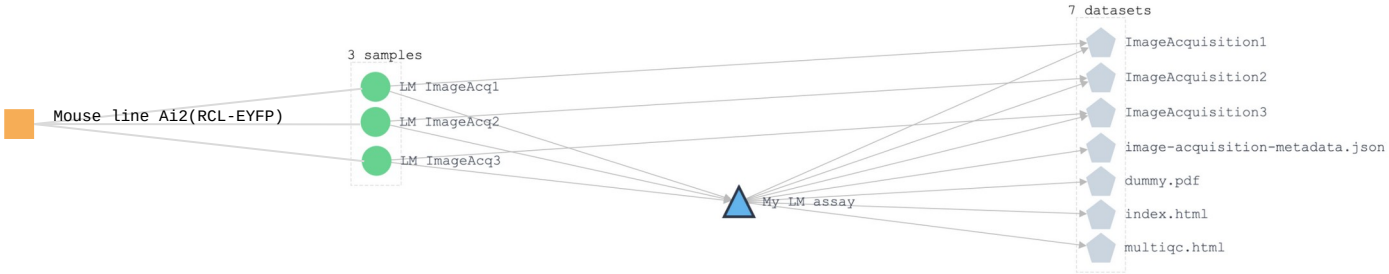
Biomaterial / Sample
ex : tissue section



Assay
Ex : histology imaging

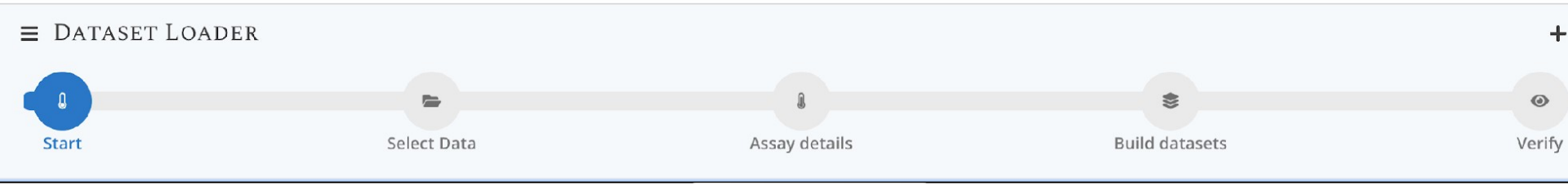


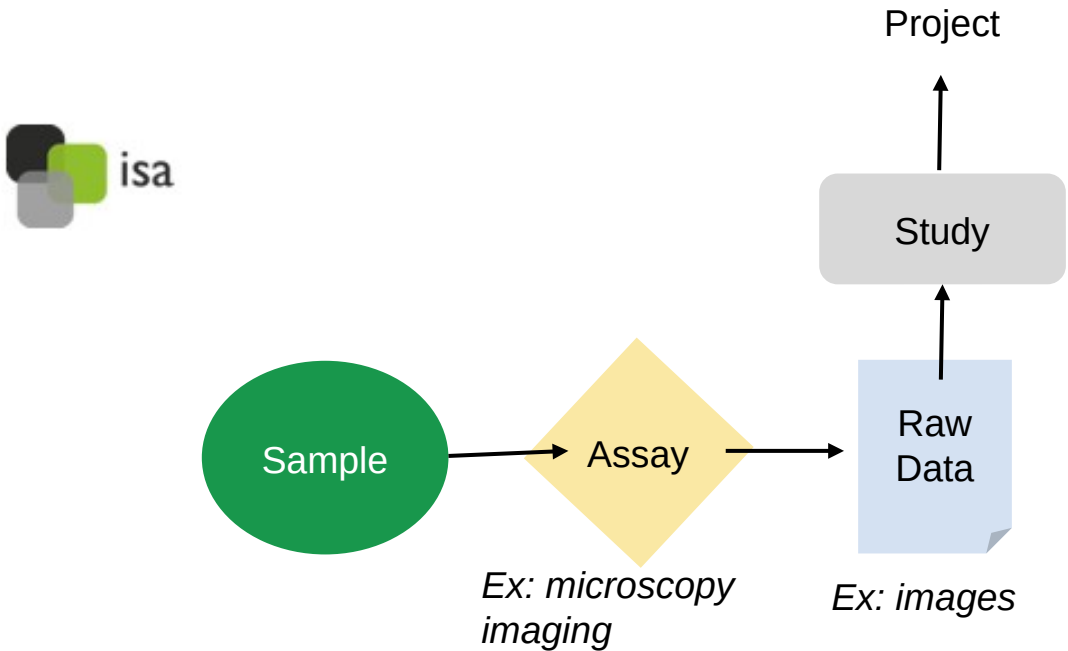
Raw data
Ex : Image dataset(s)



Tissue section : "Image provided by Servier Medical Art (https://smart.servier.com/), licensed under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/)."

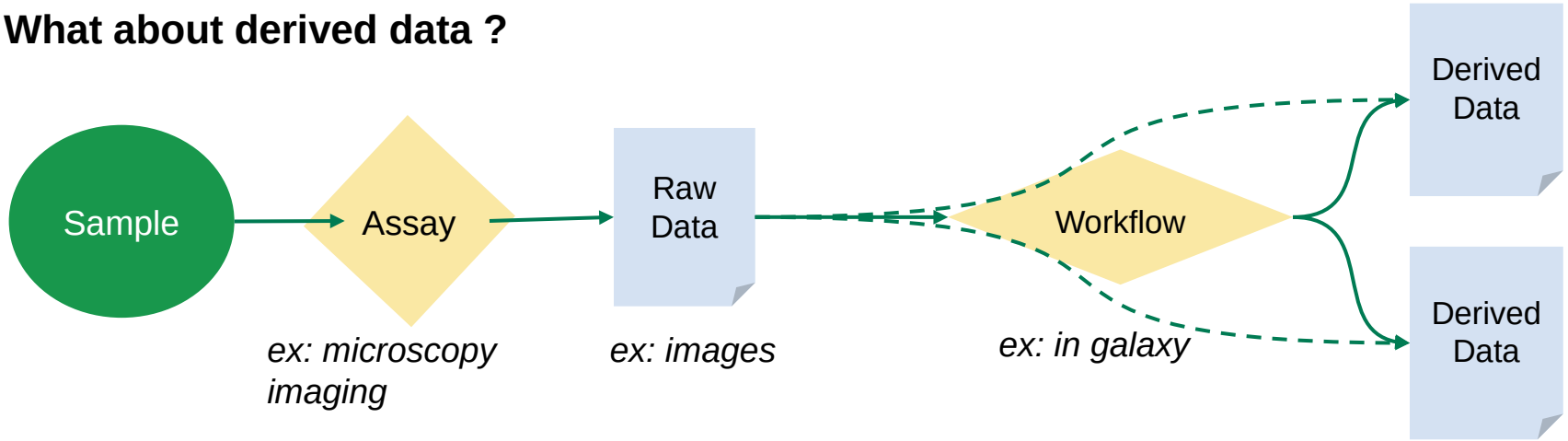
Recording data provenance at data registration





EMBL

What about derived data ?



possible to link raw to derived data

BUT missing all the information about the workflow execution / how the data was produced !

LabID workflow integration

Documenting the provenance of “derived” data

OSCARS Project :

LabID-PROV: Tracking and Sharing Data Provenance with RO-Crate in Lab Integrated Data

Funded by
the European Union

EMBL

LabID PROV - motivation and goals

Workflows are commonly used to process research data but...

1) there is **no central solution to track workflow executions**
2) workflows are not systematically published
3) sharing derived data is not common practice (community dependant)
4) shared data often **lacks metadata** (research context, biological provenance...)

> How to simplify tracking and sharing of derived data with full data provenance?

Proposal

1) use LabID to document workflow executions (version, parameters, data used...)
2) extend the graph of data provenance to the derived data generated by workflows
3) integrate with workflow repositories (WorkflowHub) and workflow engines (Galaxy, Nextflow, Snakemake)
4) leverage the **RO-Crate specification** as an **interoperable format capturing biological and computational data provenance**

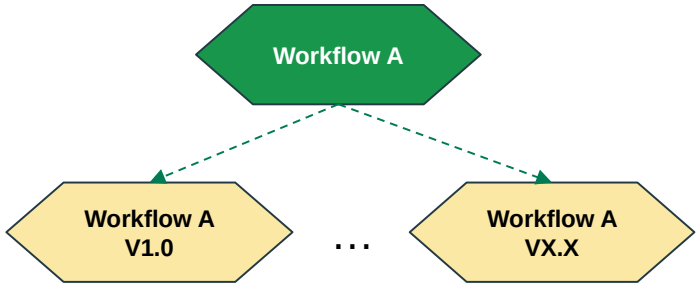
Expected benefits

- Foster FAIR practices around workflows
- Facilitate collaborations between users and developers of workflows
- Encourage the publication of workflows and workflow runs

OSCARS Open Science Clusters' Action for Research & Society
Funded by the European Union
EMBL

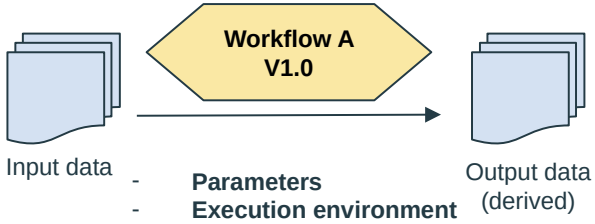
Modelling workflow entities

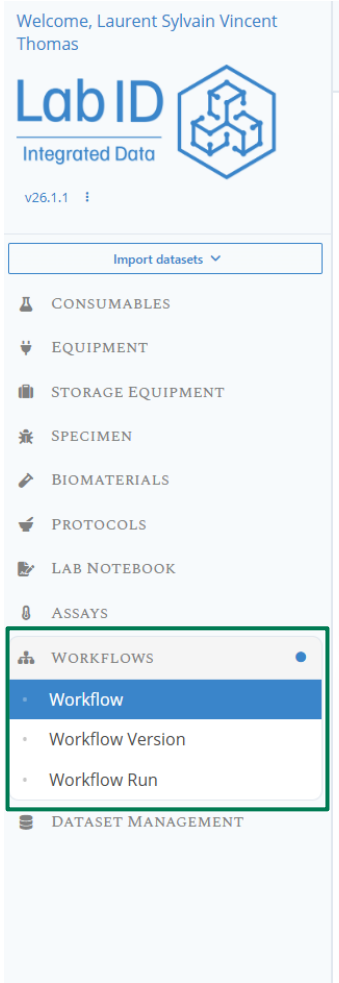
Workflow



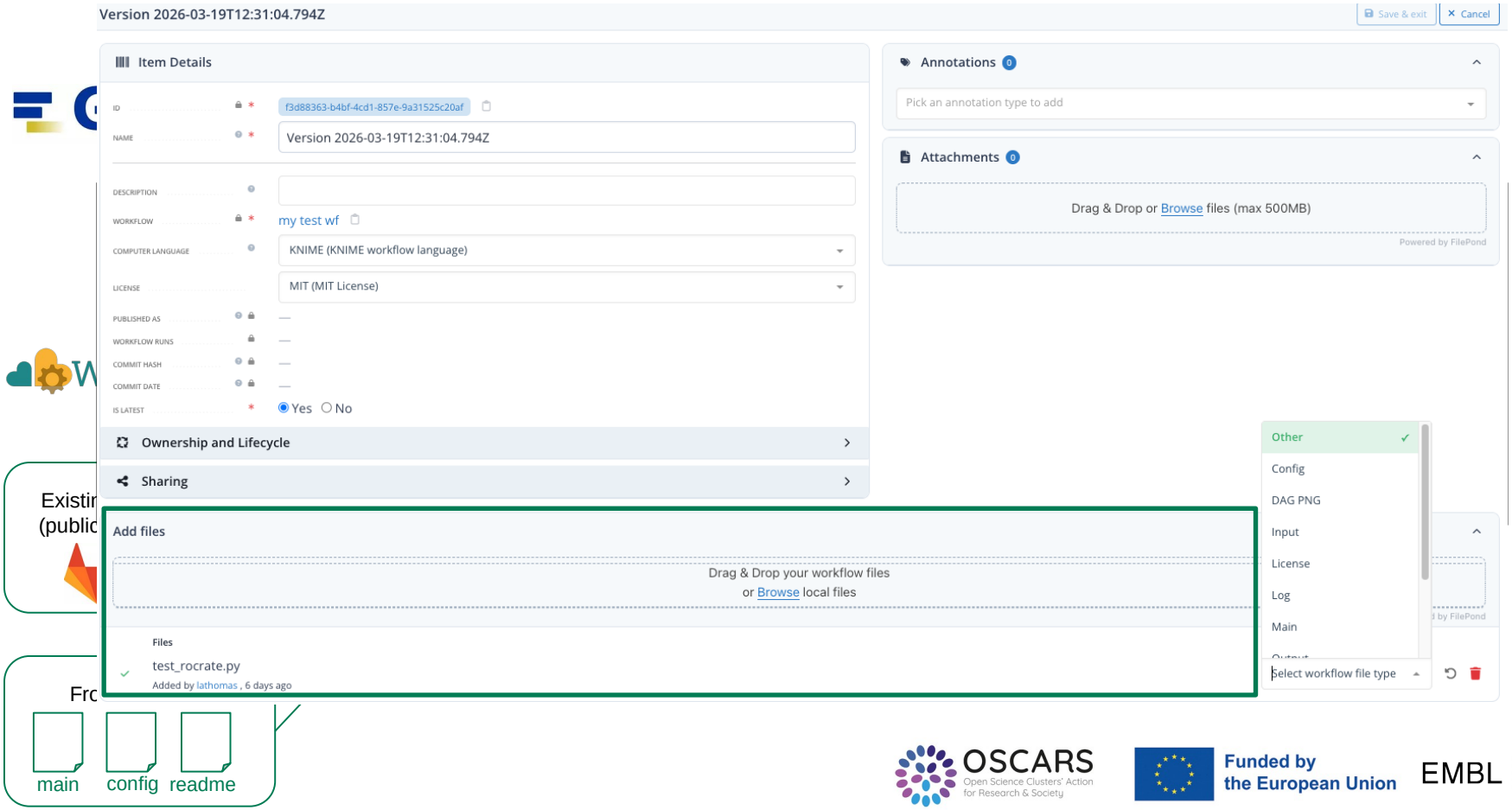
Workflow Versions

Workflow Run





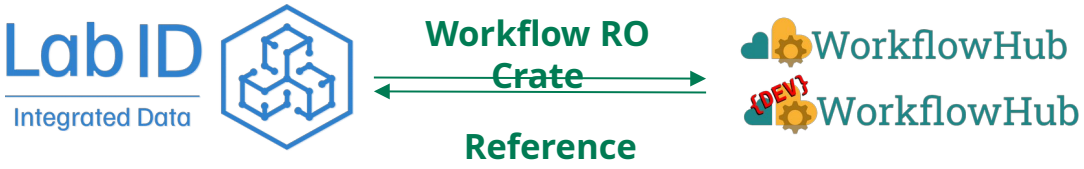
Import, creation and export of workflows versions in LabID



OSCARS Open Science Clusters' Action for Research & Society
Funded by the European Union
EMBL

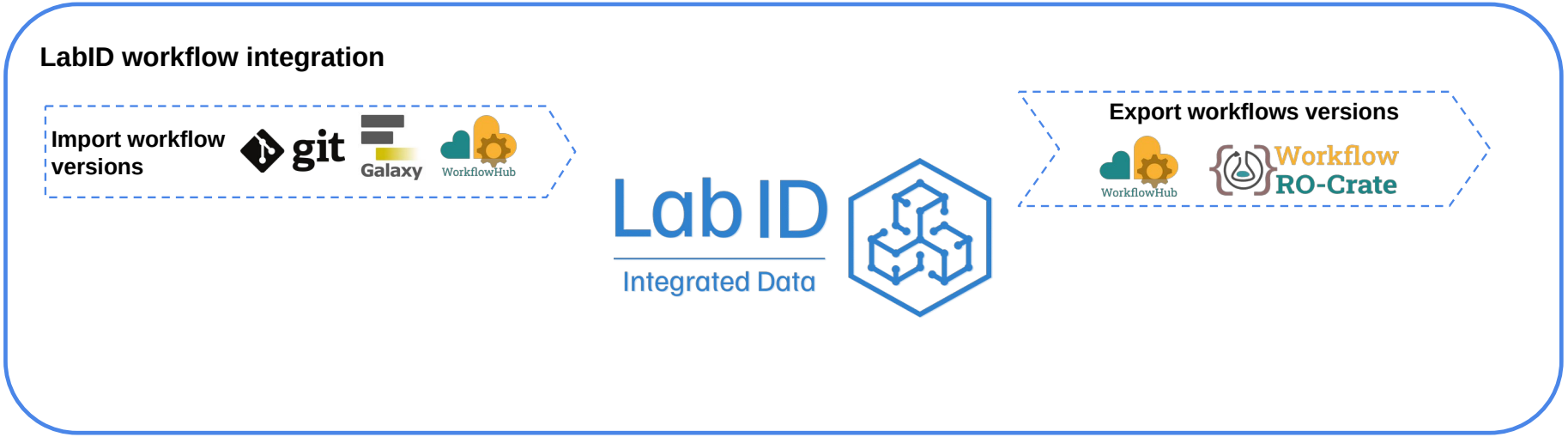
1-click export of workflows to WorkflowHub

- **Workflow RO Crate** generated by LabID if not existing
- Support for test instance (dev.workflowhub.eu)
- Reference to published workflow stored in LabID (prevent double-publication)
- Leverage **WorkflowHub API authentication** (API key of user stored in LabID)



OSCARS
Open Science Clusters' Action for Research & Society
Funded by the European Union
EMBL

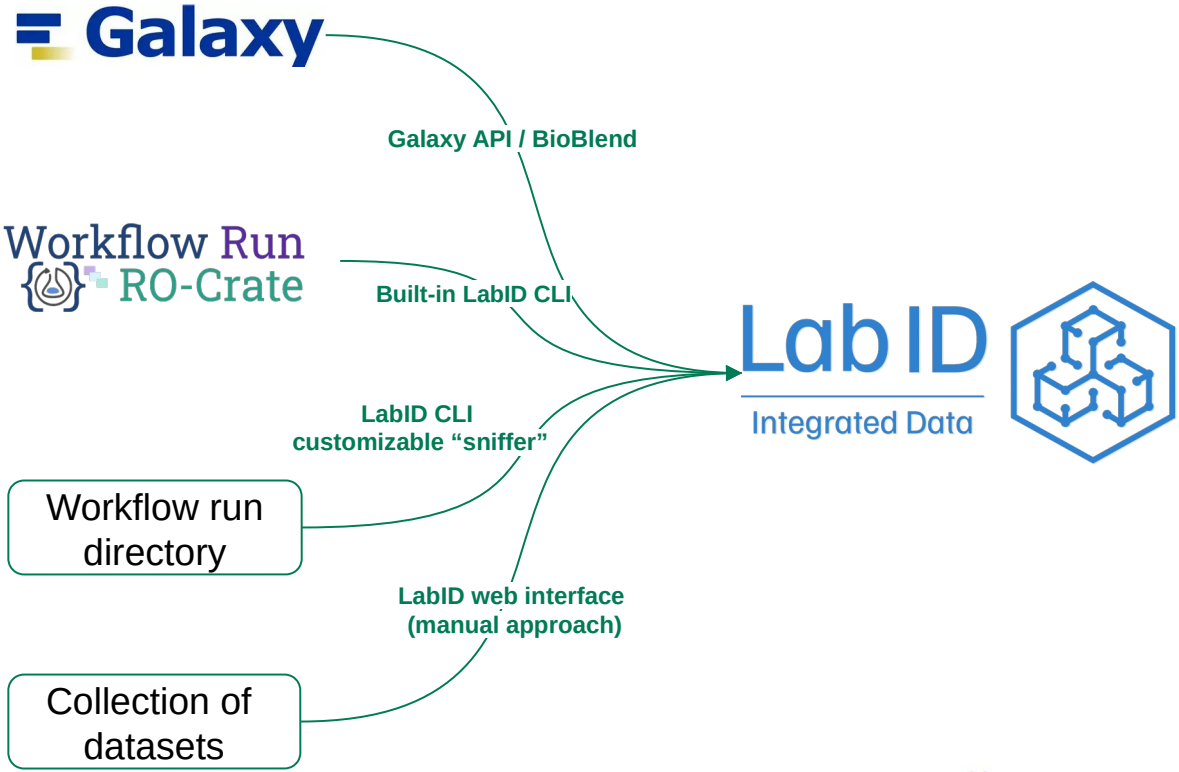
So far…

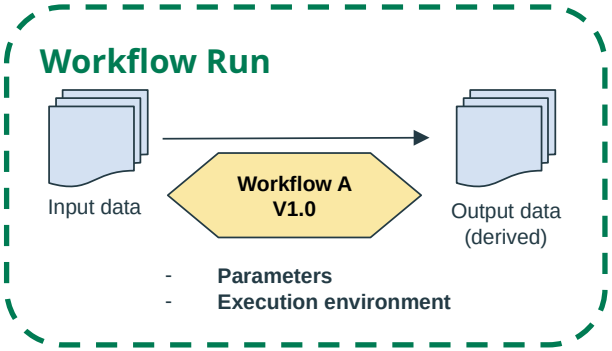


What about workflow runs ?...

OSCARS
Open Science Clusters' Action for Research & Society
Funded by the European Union
EMBL

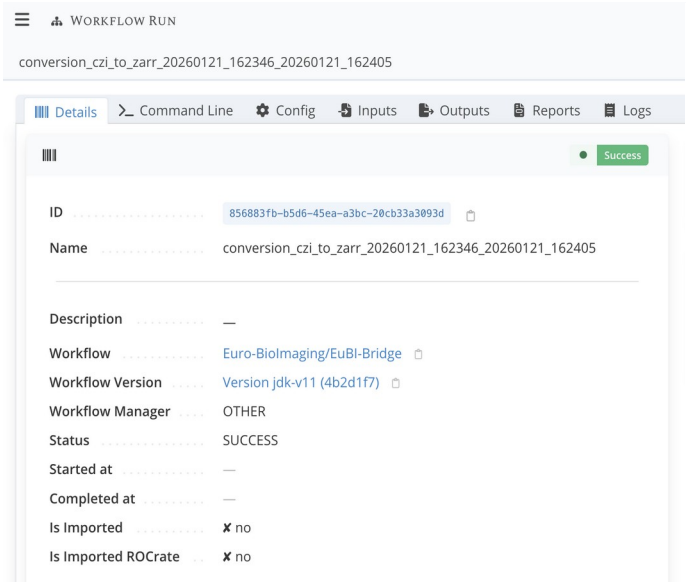
Importing/creating workflow runs in LabID

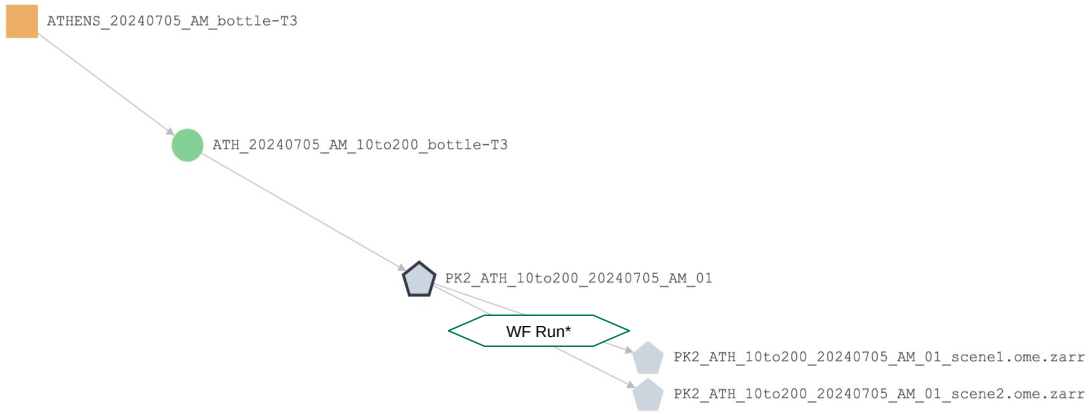




OSCARS
Open Science Clusters' Action for Research & Society
Funded by the European Union
EMBL

Example : image-conversion as Workflow runs





(* not yet depicted in the current version of LabID)

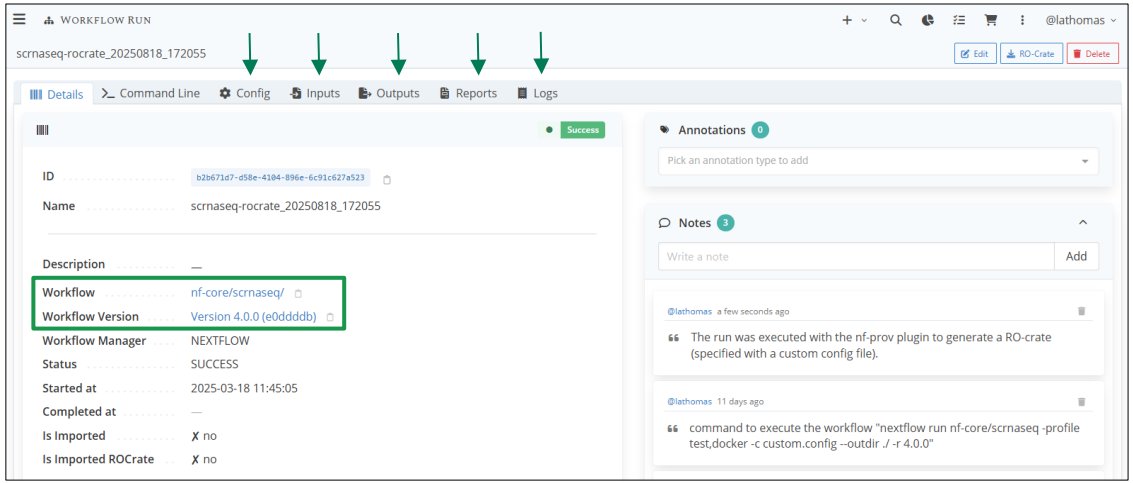
✔️Capture OME-Zarr conversion metadata e.g. tool version, parameters and execution environment
✔️Complete provenance information, up to original biological material

Perspective: **Automate** image conversion to OME-Zarr **upon registration of raw data** (czi image)

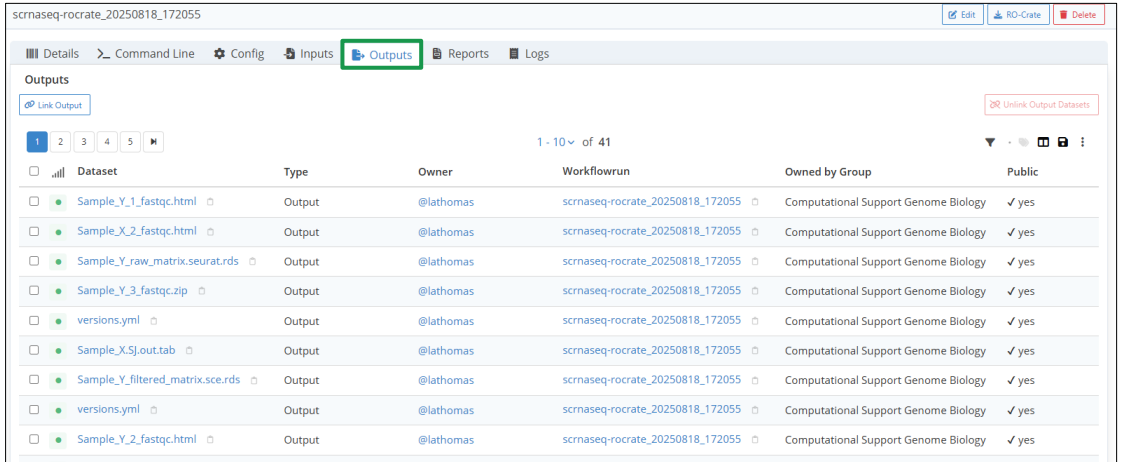
OSCARS Open Science Clusters' Action for Research & Society
Funded by the European Union
EMBL

Example workflow run

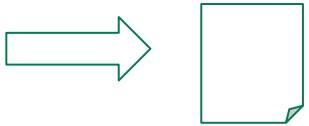
Workflow version executed



Data used as input, config and generated by the run (output, logs...)



How to share/export workflow runs including all the provenance ?

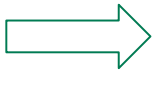


Tabular / MAGE-TAB / XML (ENA-EGA)

- simple but limited support for derived data
- no widely adopted standard (mainly suitable for genomics)

- **Graph** of data-provenance
- associated metadata

Lab ID
Integrated Data



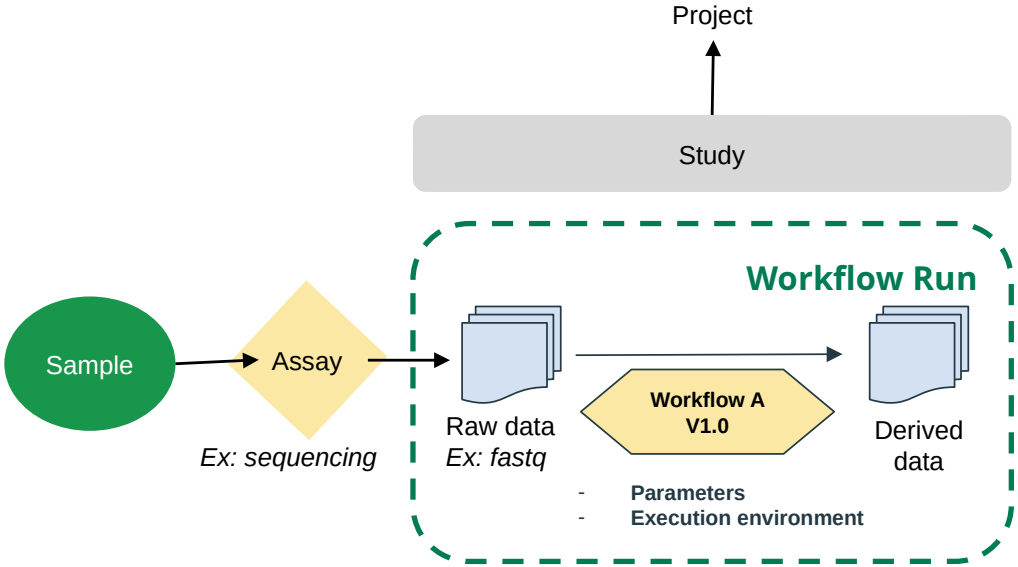
Workflow Run RO-Crate

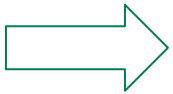
- JSON describing a **graph** of **data** and **context entities**
- **standardised** with ontologies / controlled vocabularies
- **interoperable** / **machine-actionable** (indexing, search)
- **extensible**
  ex : ISA-RO-Crate specs to document biological provenance (assay, sample prep.)
- already **adopted by other platforms and communities** (WorkflowHub, Nextflow, Galaxy, ARC…)

OSCARS
Open Science Clusters' Action for Research & Society
Funded by the European Union
EMBL

Exporting workflow runs from LabID

- LabID generates a custom **Workflow Run RO Crate** (unless imported as RO-Crate)
- Including **computational AND biological data-provenance** (e.g sample LabID identifiers of exported entities)
- Can be published to e.g Zenodo, RO-Hub…



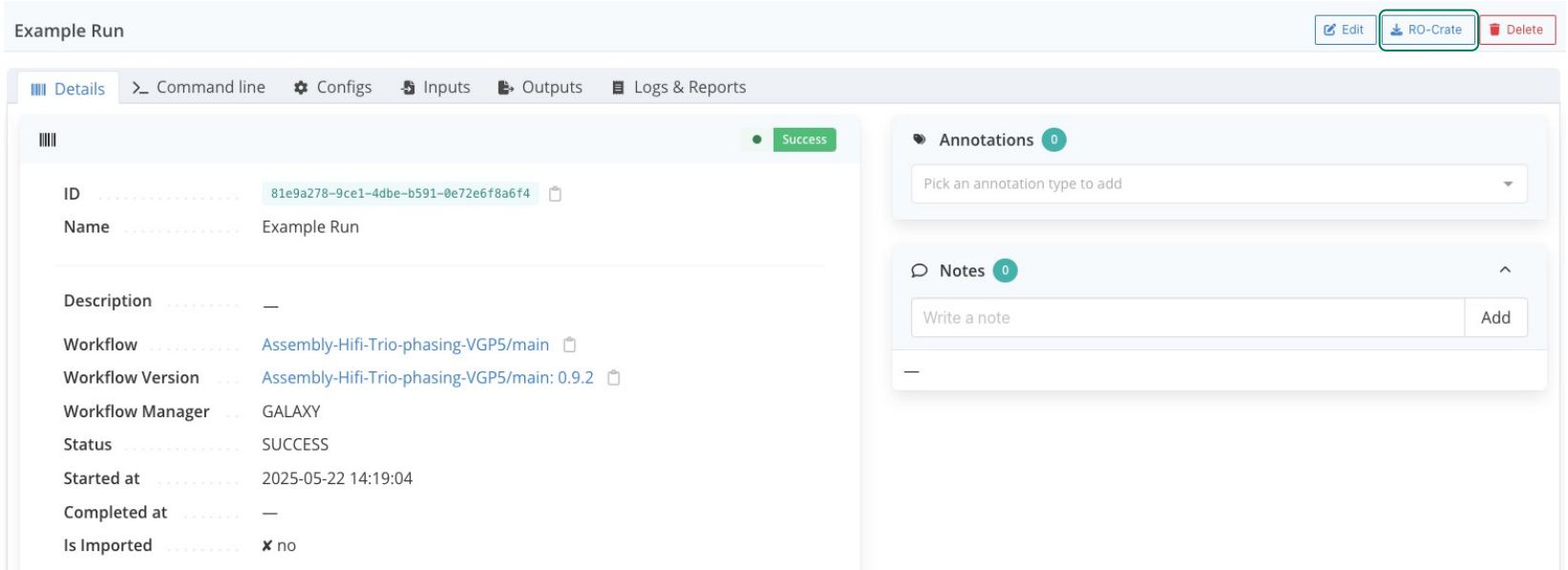


“Augmented”
Workflow Run RO-Crate

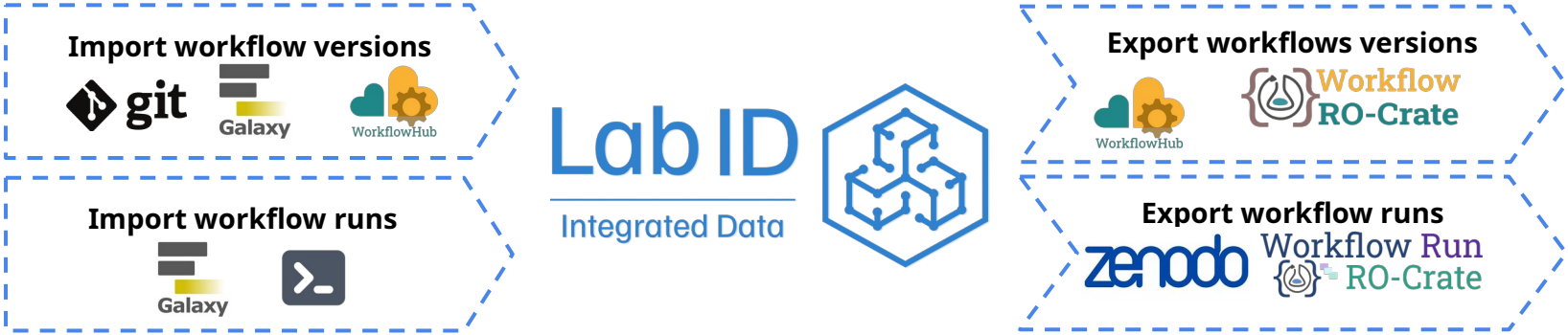
Work in progress…

1-click export of Workflow Run RO-Crate

- agnostic of workflow engine
- transferable to other disciplines



Interoperability with workflow ecosystem



LabID workflow integration - take-aways

Enhance FAIR practices around workflows by

- documenting **biological** and **computational provenance of data**
- integrating with workflow management systems / repositories
- streamlining workflows/runs **publication** and **sharing**
- facilitating **workflow versioning** and **collaborative development**
- encouraging adoption by providing **simple automation mechanisms**

OSCARS
Open Science Clusters' Action for Research & Society
Funded by the European Union
EMBL

Training and outreach

- **Try it !**

  demo server => https://labid-demo.embl.de/

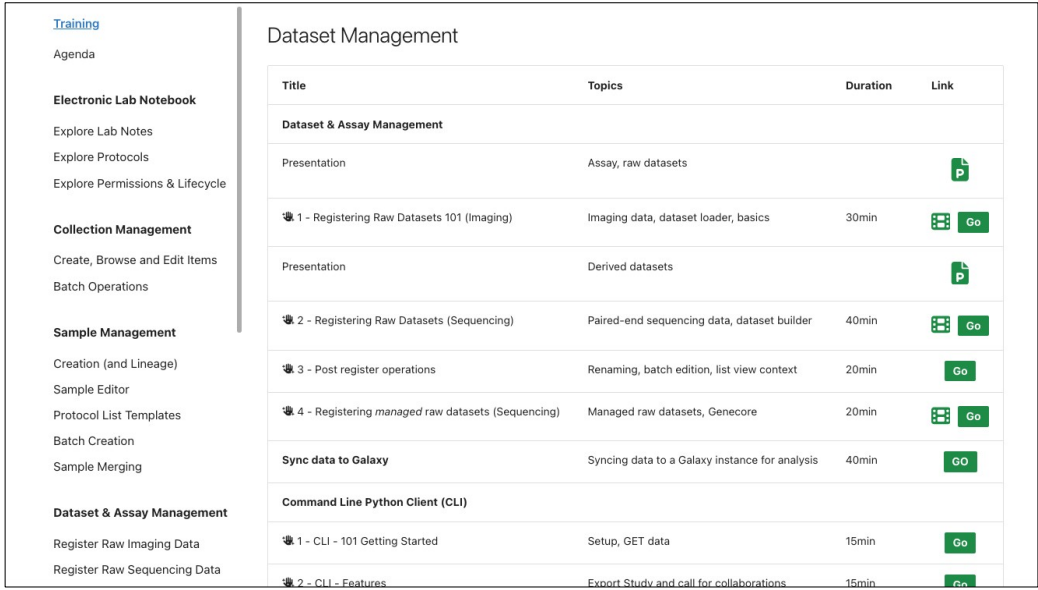
  or on your machine with docker-compose (see server repo)

- **Documentation, video and training tutorials**

  https://grp-gbcs.embl-community.io/labid-user-docs/

- **Upcoming in-person training on workflow integration**

  End of September 2026



OSCARS Open Science Clusters' Action for Research & Society
Funded by the European Union
EMBL

LabID is open source – open to contributions and collaborations

Many ways to contribute

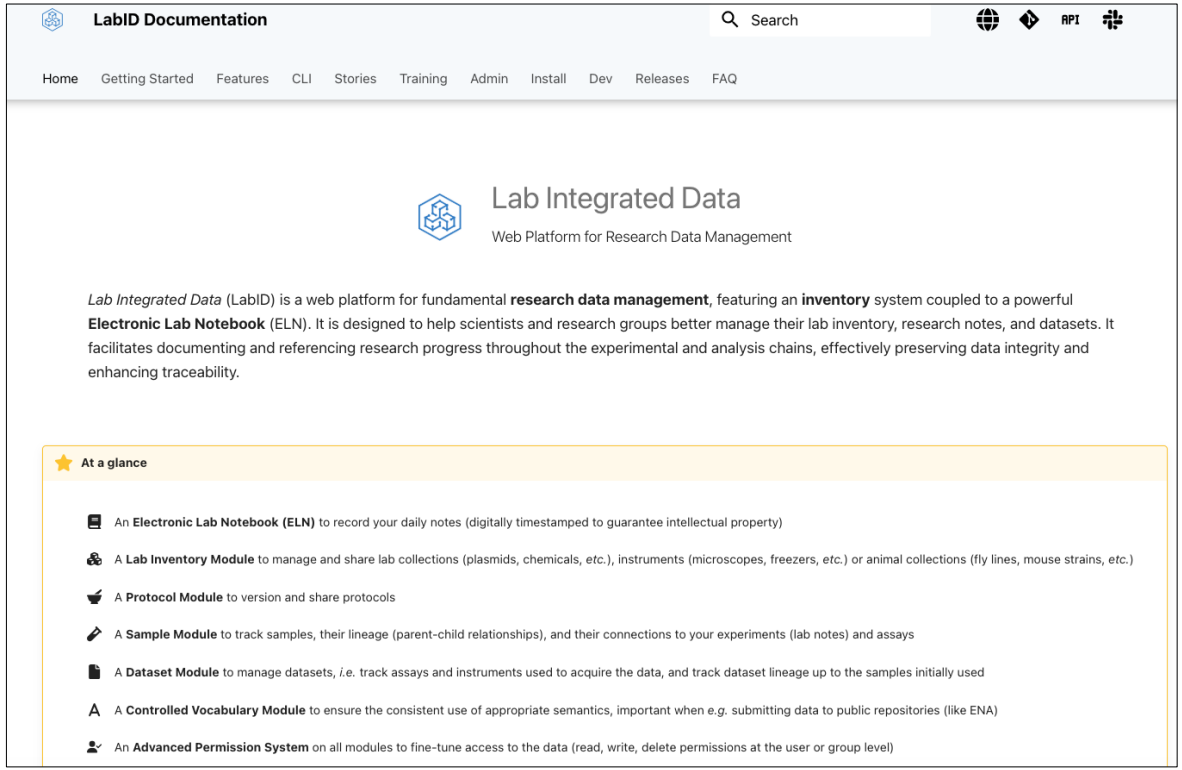
- feedback on user/admin experience
- documentation
- suggestions and bug reports

- **Source code**

https://gitlab.com/lab-integrated-data

- **Reach out**

modis@embl.de or Slack



OSCARS Open Science Clusters' Action for Research & Society
Funded by the European Union
EMBL

TEAM

Multimodal Open Data Integration Support (MODIS)

Charles Girardot
Head of Multimodal Open Data Integration Support
ORCID: 0000-0003-4301-3920

Laurent Thomas
Postdoctoral Fellow
ORCID: 0000-0001-7686-3249

Md Nayeem Reza
Senior Software Engineer
ORCID: 0000-0003-2068-5812

Matthias Monfort
Web Developer / Workflow Management

Marco Wetter
Full Stack Developer

lab rise CONSULTING

Jelle Scholtalbers · 1er
Freelancer / Full stack developer
France · Coordonnées
https://labrise-consulting.com
ORCID : 0000-0002-6090-2482

OSCARS
Open Science Clusters' Action for Research & Society

Funded by the European Union

This work was supported through the Open Science Clusters' Action for Research and Society (OSCARS) European project under grant agreement N°101129751.

EMBL