

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

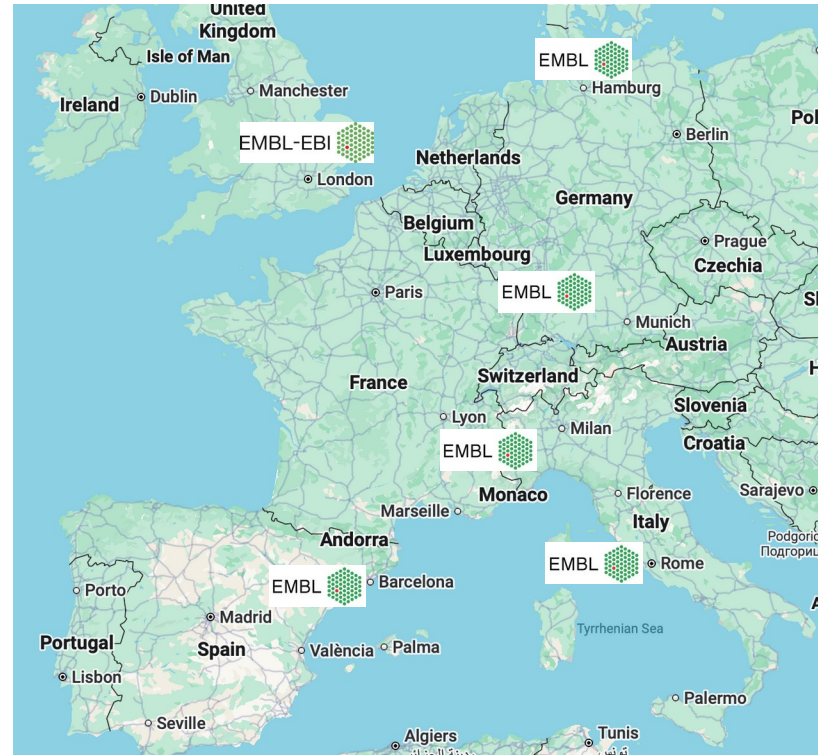


The EMBL : European Molecular Biology Laboratory

- 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)**...



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 pi

Welcome, Laurent Sylvain Vincent
Thomas

Lab ID 
Integrated Data

v26.1.1 ⓘ


Import datasets ▾

-  CONSUMABLES
-  EQUIPMENT
-  STORAGE EQUIPMENT
-  SPECIMEN
-  BIOMATERIALS
-  PROTOCOLS
-  LAB NOTEBOOK
-  ASSAYS
-  WORKFLOWS
-  DATASET MANAGEMENT




☰

WELCOME

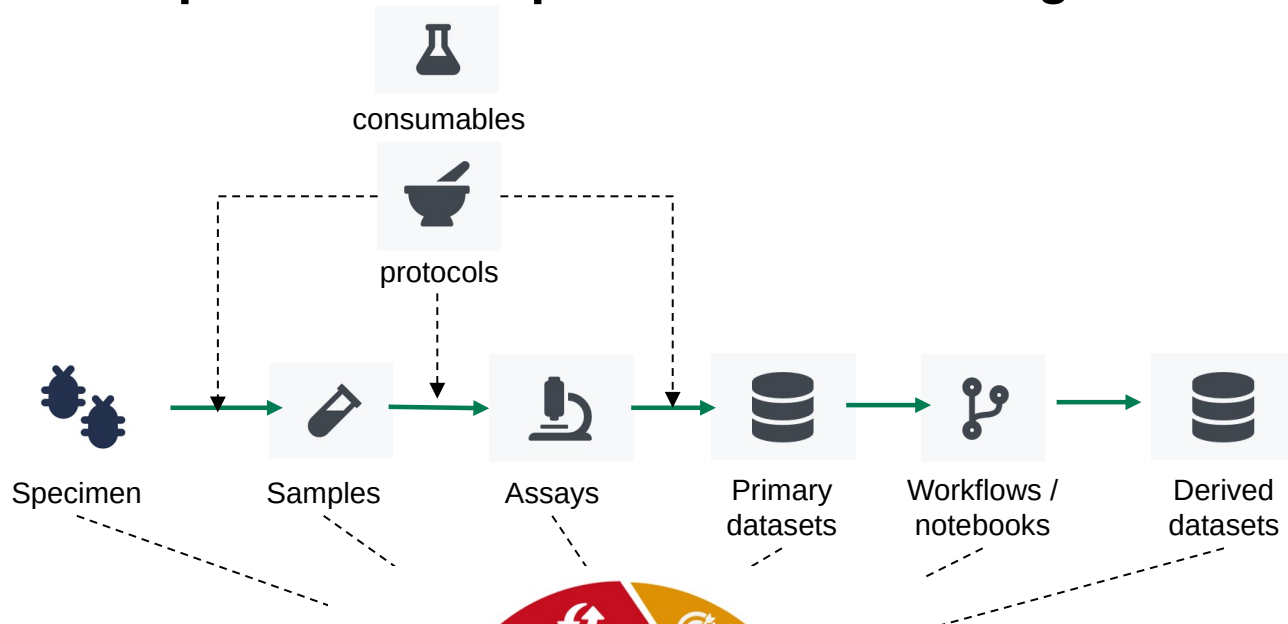
Lab ID

 Search...

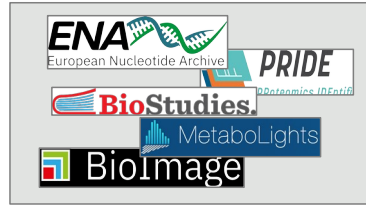
GET SUPPORT

-  READ THE DOCUMENTATION
-  CHAT WITH US
-  EMAIL US

LabID : a platform to capture metadata along the data lifecycle



SZF TU Berlin, CC BY 4.0



Share with Collaborators

Track information in a lab

Consolidate datasets (AI)

4 interconnected modules



v26.1.1

Import datasets ▾

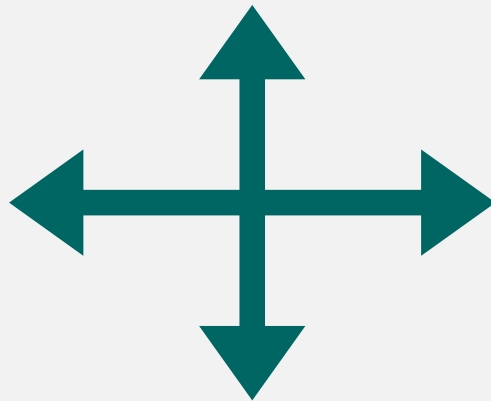
- CONSUMABLES
- EQUIPMENT
- STORAGE EQUIPMENT
- SPECIMEN
- BIOMATERIALS
- PROTOCOLS
- LAB NOTEBOOK
- ASSAYS
- WORKFLOWS
- DATASET MANAGEMENT

Lab Collection Management
(Consumables, Equipment, Storage)

Dataset Management
(Assays, Datasets, Workflows & Studies)

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

Biomaterial Management
(Sample & Specimen, Biobanking)



Example : tissue imaging assay

Welcome, Laurent Sylvain Vincent
Thomas



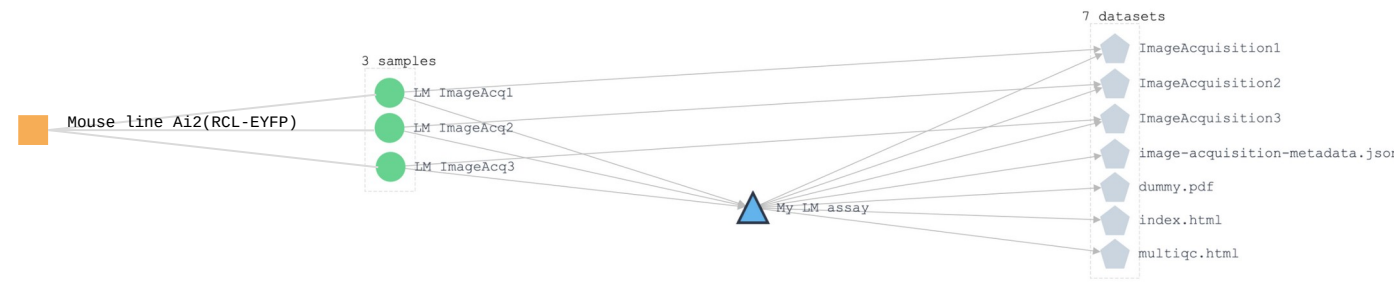
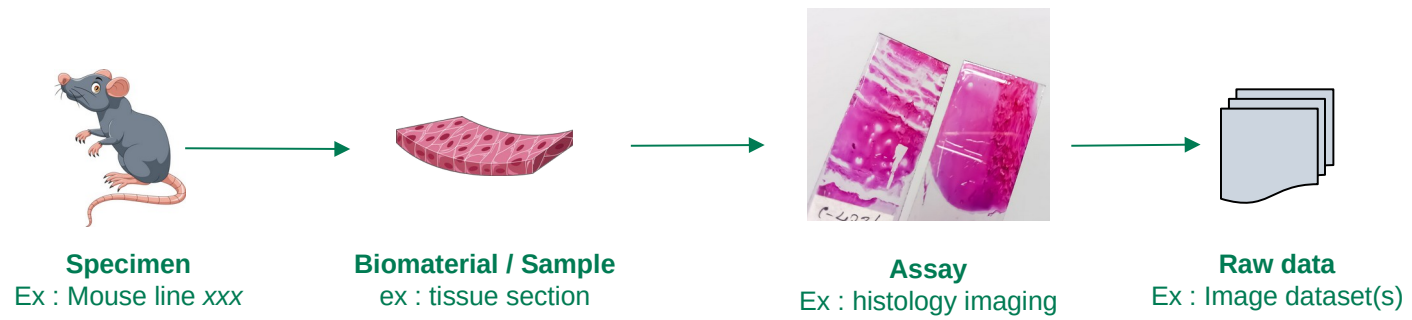
v26.1.1

Import datasets

- CONSUMABLES
- EQUIPMENT
- STORAGE EQUIPMENT

- SPECIMEN**
 - TREC
 - Cell Line
 - Strain
 - Fly Line
 - Virus
 - Cnidaria
 - Fish
 - Planaria
 - All

- BIOMATERIALS
- PROTOCOLS



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

≡ DATASET LOADER

+



Start



Select Data



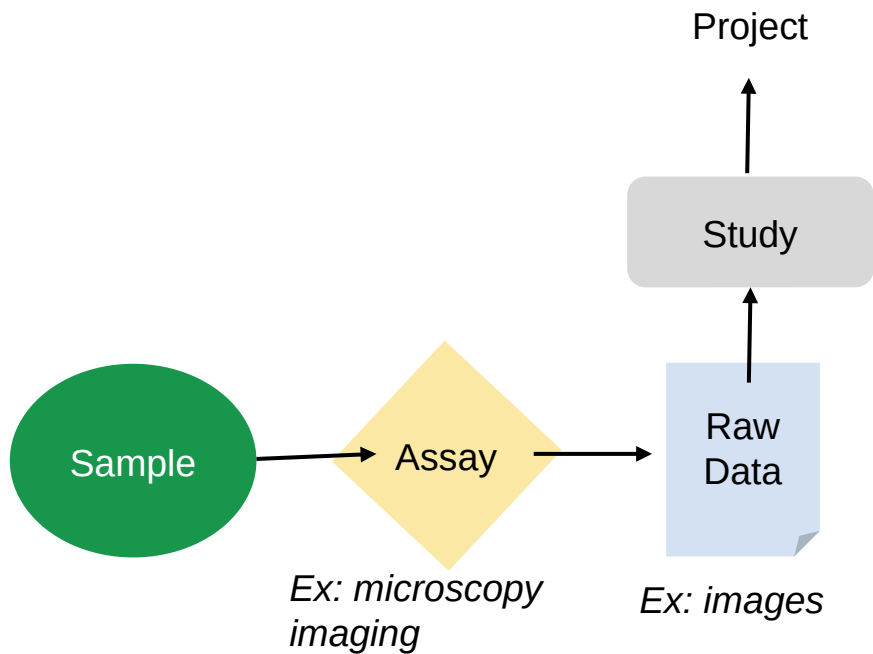
Assay details



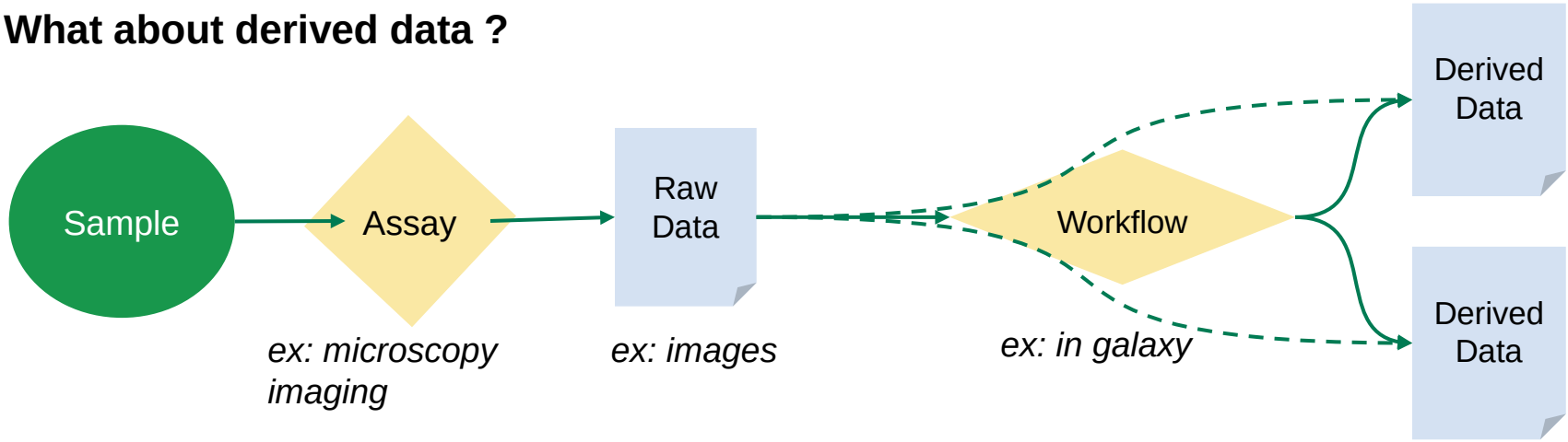
Build datasets




Verify



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**



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

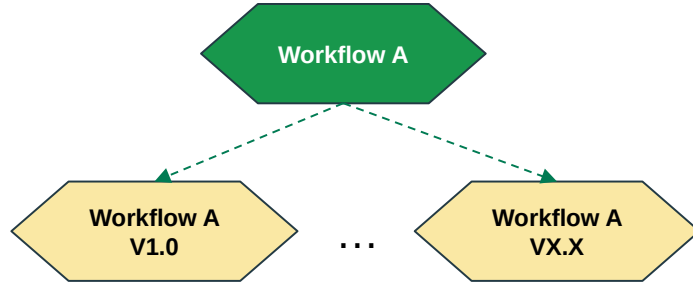
- 1) use LabID to document workflow executions (version, parameters, data used...)
- 2) extend the graph of data provenance to the derived data generated by workflows
- 3) integrate with workflow repositories (WorkflowHub) and workflow engines (Galaxy, Nextflow, Snakemake)
- 4) leverage the **RO-Crate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

Expected benefits

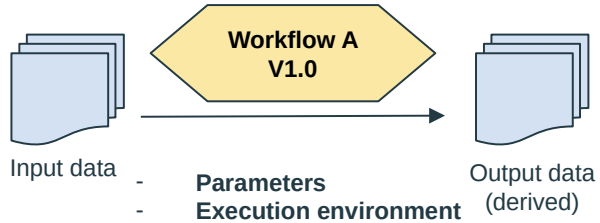
- Foster FAIR practices around workflows
- Facilitate collaborations between users and developers of workflows
- Encourage the publication of workflows and workflow runs

Modelling workflow entities

Workflow



Workflow
Versions



Workflow
Run

Welcome, Laurent Sylvain Vincent
Thomas



v26.1.1

Import datasets

- CONSUMABLES
- EQUIPMENT
- STORAGE EQUIPMENT
- SPECIMEN
- BIOMATERIALS
- PROTOCOLS
- LAB NOTEBOOK
- ASSAYS

- WORKFLOWS
 - Workflow
 - Workflow Version
 - Workflow Run

DATASET MANAGEMENT

Import, creation and export of workflows versions in LabID

Version 2026-03-19T12:31:04.794Z Save & exit Cancel

Item Details

ID: f9d88363-b4bf-4cd1-857e-9a31525c20af

NAME: Version 2026-03-19T12:31:04.794Z

DESCRIPTION:

WORKFLOW: my test wf

COMPUTER LANGUAGE: KNIME (KNIME workflow language)

LICENSE: MIT (MIT License)

PUBLISHED AS: —

WORKFLOW RUNS: —

COMMIT HASH: —

COMMIT DATE: —

IS LATEST: Yes No

Ownership and Lifecycle

Sharing

Annotations

Pick an annotation type to add

Attachments

Drag & Drop or [Browse](#) files (max 500MB)

Powered by FilePond

- Other ✓
- Config
- DAG PNG
- Input
- License
- Log
- Main
- Output

Select workflow file type

Existing (public)

Add files

Drag & Drop your workflow files or [Browse](#) local files

Files

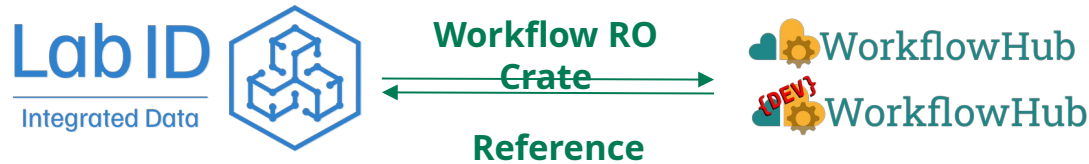
- test_rocrate.py
Added by lathomas, 6 days ago

From

- main
- config
- readme

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)



So far...

LabID workflow integration

Import workflow
versions



git



Galaxy



WorkflowHub

Lab ID
Integrated Data



Export workflows versions



WorkflowHub



Workflow
RO-Crate

What about workflow runs ?...

Importing/creating workflow runs in LabID



Galaxy API / BioBlend



Built-in LabID CLI

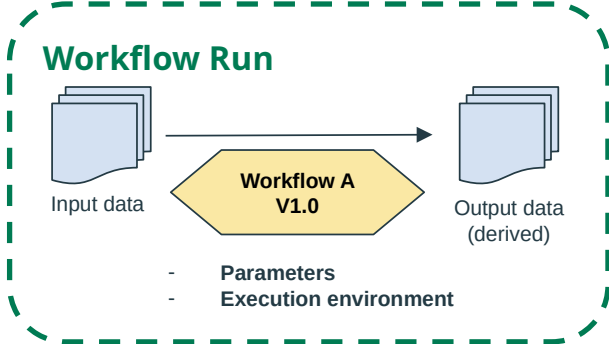
LabID CLI customizable "sniffer"

Workflow run directory

LabID web interface (manual approach)

Collection of datasets

Lab ID
Integrated Data



Example : image-conversion as Workflow runs

WORKFLOW RUN

conversion_czi_to_zarr_20260121_162346_20260121_162405

Details Command Line Config Inputs Outputs Reports Logs

Success

ID 856883fb-b5d6-45ea-a3bc-20cb33a3093d

Name conversion_czi_to_zarr_20260121_162346_20260121_162405

Description —

Workflow Euro-Biolmaging/EuBI-Bridge

Workflow Version Version jdk-v11 (4b2d1f7)

Workflow Manager OTHER

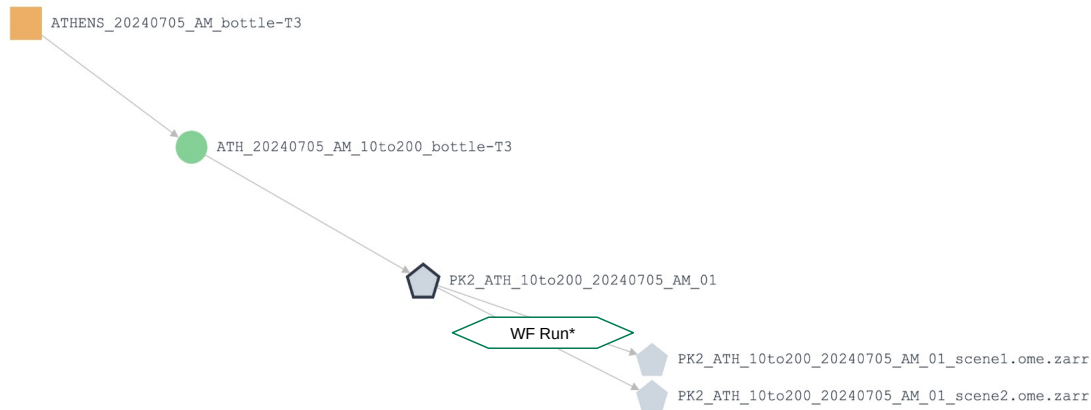
Status SUCCESS

Started at —

Completed at —

Is Imported ✗ no

Is Imported ROCrate ✗ no



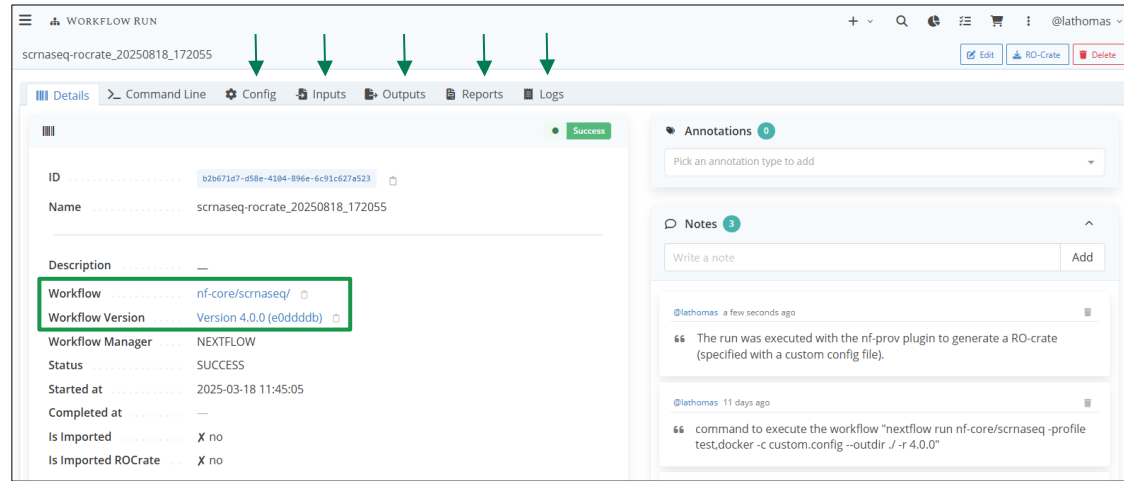
(* not yet depicted in the current version of LabID)

- ✓ 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)

Example workflow run

Workflow version executed



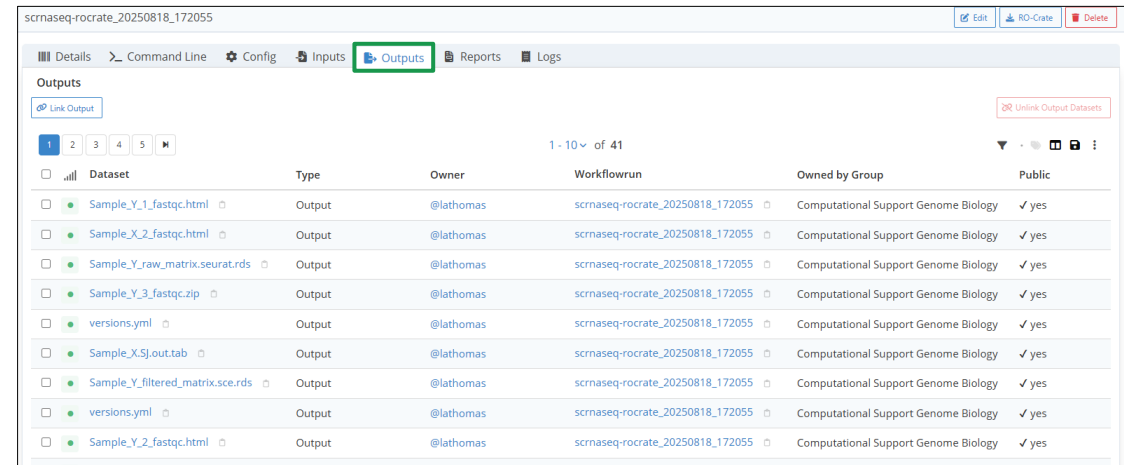
Workflow Run Details for `scrnaseq-rocrate_20250818_172055`

Workflow: `nf-core/scrnaseq`
Workflow Version: `Version 4.0.0 (e0d4db)`

Status: SUCCESS
Started at: 2025-03-18 11:45:05

Notes:
The run was executed with the nf-prov plugin to generate a RO-crate (specified with a custom config file).
command to execute the workflow "nextflow run nf-core/scrnaseq -profile test,docker -c custom.config --outdir ./-r 4.0.0"

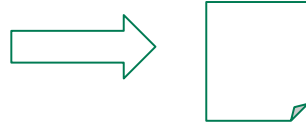
Data used as input, config and generated by the run (output, logs...)



Outputs

Dataset	Type	Owner	Workflowrun	Owned by Group	Public
Sample_Y_1_fastqc.html	Output	@lathomas	scrnaseq-rocrate_20250818_172055	Computational Support Genome Biology	✓ yes
Sample_X_2_fastqc.html	Output	@lathomas	scrnaseq-rocrate_20250818_172055	Computational Support Genome Biology	✓ yes
Sample_Y_raw_matrix.seurat.rds	Output	@lathomas	scrnaseq-rocrate_20250818_172055	Computational Support Genome Biology	✓ yes
Sample_Y_3_fastqc.zip	Output	@lathomas	scrnaseq-rocrate_20250818_172055	Computational Support Genome Biology	✓ yes
versions.yml	Output	@lathomas	scrnaseq-rocrate_20250818_172055	Computational Support Genome Biology	✓ yes
Sample_X_5j.out.tab	Output	@lathomas	scrnaseq-rocrate_20250818_172055	Computational Support Genome Biology	✓ yes
Sample_Y_filtered_matrix.sce.rds	Output	@lathomas	scrnaseq-rocrate_20250818_172055	Computational Support Genome Biology	✓ yes
versions.yml	Output	@lathomas	scrnaseq-rocrate_20250818_172055	Computational Support Genome Biology	✓ yes
Sample_Y_2_fastqc.html	Output	@lathomas	scrnaseq-rocrate_20250818_172055	Computational Support Genome Biology	✓ yes

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

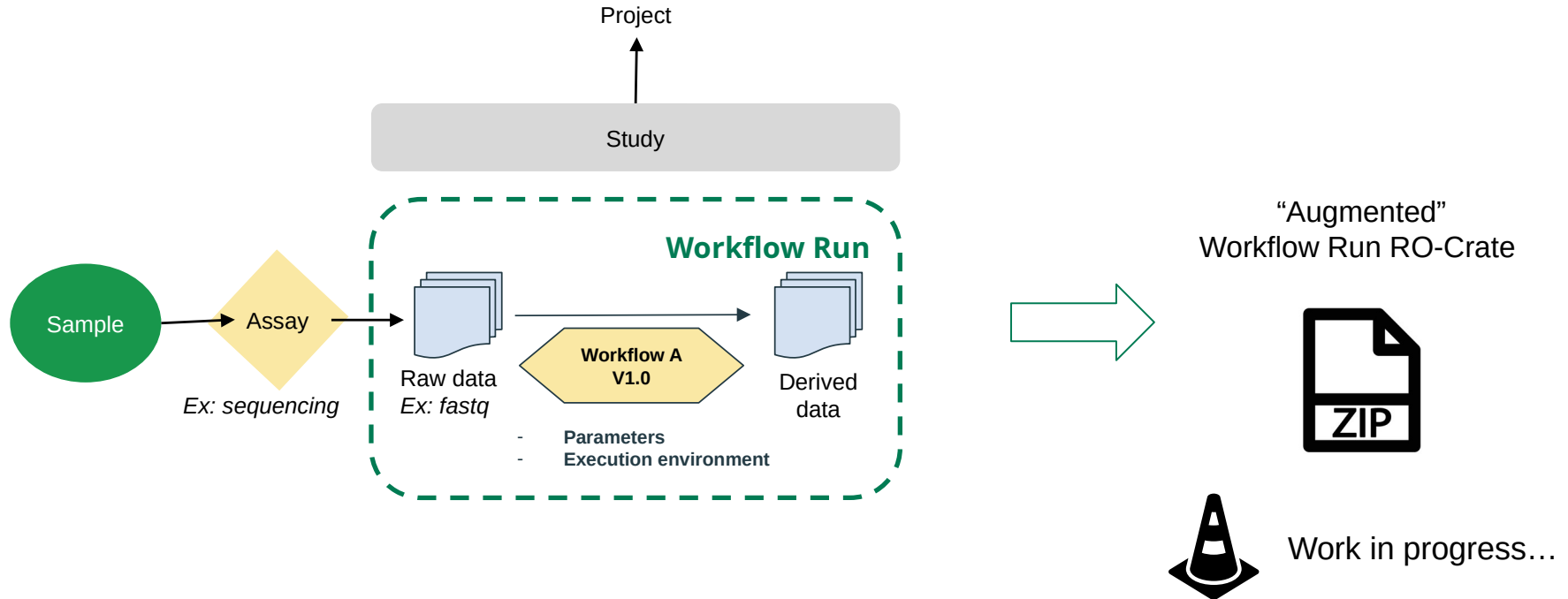


Workflow Run
{ } RO-Crate

- 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...)

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...



1-click export of Workflow Run RO-Crate

- agnostic of workflow engine
- transferable to other disciplines

The screenshot shows a web interface for a workflow run. At the top, there's a header "Example Run" with three buttons: "Edit", "RO-Crate" (highlighted with a green box), and "Delete". Below the header is a navigation bar with tabs: "Details" (selected), "Command line", "Configs", "Inputs", "Outputs", and "Logs & Reports". The main content area is split into two panels. The left panel, titled "Details", shows a list of metadata for a workflow run that has completed successfully. The right panel contains sections for "Annotations" and "Notes", both currently empty.

Property	Value
ID	81e9a278-9ce1-4dbe-b591-0e72e6f8a6f4
Name	Example Run
Description	—
Workflow	Assembly-Hifi-Trio-phasing-VGP5/main
Workflow Version	Assembly-Hifi-Trio-phasing-VGP5/main: 0.9.2
Workflow Manager	GALAXY
Status	SUCCESS
Started at	2025-05-22 14:19:04
Completed at	—
Is Imported	✗ no

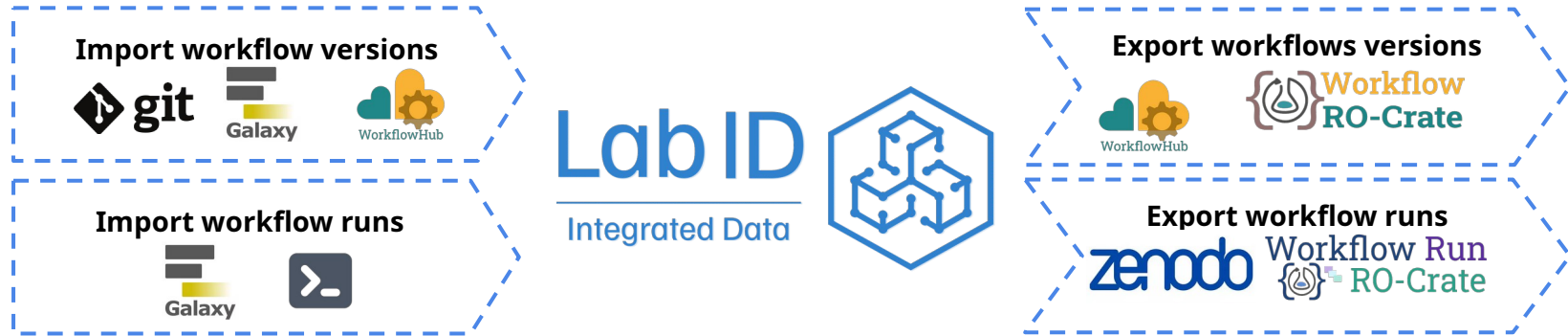
Annotations 0

Pick an annotation type to add

Notes 0

Write a note

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**

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

Training		Dataset Management			
Agenda					
Electronic Lab Notebook					
Explore Lab Notes					
Explore Protocols					
Explore Permissions & Lifecycle					
Collection Management					
Create, Browse and Edit Items					
Batch Operations					
Sample Management					
Creation (and Lineage)					
Sample Editor					
Protocol List Templates					
Batch Creation					
Sample Merging					
Dataset & Assay Management					
Register Raw Imaging Data					
Register Raw Sequencing Data					
Title	Topics	Duration	Link		
Dataset & Assay Management					
Presentation	Assay, raw datasets				
1 - Registering Raw Datasets 101 (Imaging)	Imaging data, dataset loader, basics	30min			
Presentation	Derived datasets				
2 - Registering Raw Datasets (Sequencing)	Paired-end sequencing data, dataset builder	40min			
3 - Post register operations	Renaming, batch edition, list view context	20min			
4 - Registering <i>managed</i> raw datasets (Sequencing)	Managed raw datasets, Genecore	20min			
Sync data to Galaxy	Syncing data to a Galaxy instance for analysis	40min			
Command Line Python Client (CLI)					
1 - CLI - 101 Getting Started	Setup, GET data	15min			
2 - CLI - Features	Export Study and call for collaborations	15min			

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](#)



LabID Documentation

Home Getting Started Features CLI Stories Training Admin Install Dev Releases FAQ

Lab Integrated Data

Web Platform for Research Data Management

Lab Integrated Data (LabID) is a web platform for fundamental **research data management**, featuring an **inventory** system coupled to a powerful **Electronic Lab Notebook** (ELN). It is designed to help scientists and research groups better manage their lab inventory, research notes, and datasets. It facilitates documenting and referencing research progress throughout the experimental and analysis chains, effectively preserving data integrity and enhancing traceability.

★ **At a glance**

- 📄 An **Electronic Lab Notebook (ELN)** to record your daily notes (digitally timestamped to guarantee intellectual property)
- 🧬 A **Lab Inventory Module** to manage and share lab collections (plasmids, chemicals, etc.), instruments (microscopes, freezers, etc.) or animal collections (fly lines, mouse strains, etc.)
- 📄 A **Protocol Module** to version and share protocols
- 🔗 A **Sample Module** to track samples, their lineage (parent-child relationships), and their connections to your experiments (lab notes) and assays
- 📄 A **Dataset Module** to manage datasets, *i.e.* track assays and instruments used to acquire the data, and track dataset lineage up to the samples initially used
- 📄 A **Controlled Vocabulary Module** to ensure the consistent use of appropriate semantics, important when *e.g.* submitting data to public repositories (like ENA)
- 👤 A **Advanced Permission System** on all modules to fine-tune access to the data (read, write, delete permissions at the user or group level)

Multimodal Open Data Integration Support (MODIS)



Charles Girardot

Head of Multimodal Open Data Integration Support

ORCID: [0000-0003-4301-3920](https://orcid.org/0000-0003-4301-3920)



Laurent Thomas

Postdoctoral Fellow

ORCID: [0000-0001-7686-3249](https://orcid.org/0000-0001-7686-3249)



Md Nayeem Reza

Senior Software Engineer

ORCID: [0000-0003-2068-5812](https://orcid.org/0000-0003-2068-5812)



Matthias Monfort

Web Developer / Workflow Management



Marco Wetter

Full Stack Developer



Jelle Scholtalbers · 1er
Freelancer / Full stack developer

France · [Coordonnées](#)

<https://labrise-consulting.com>

ORCID : [0000-0002-6090-2482](https://orcid.org/0000-0002-6090-2482)



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.

