

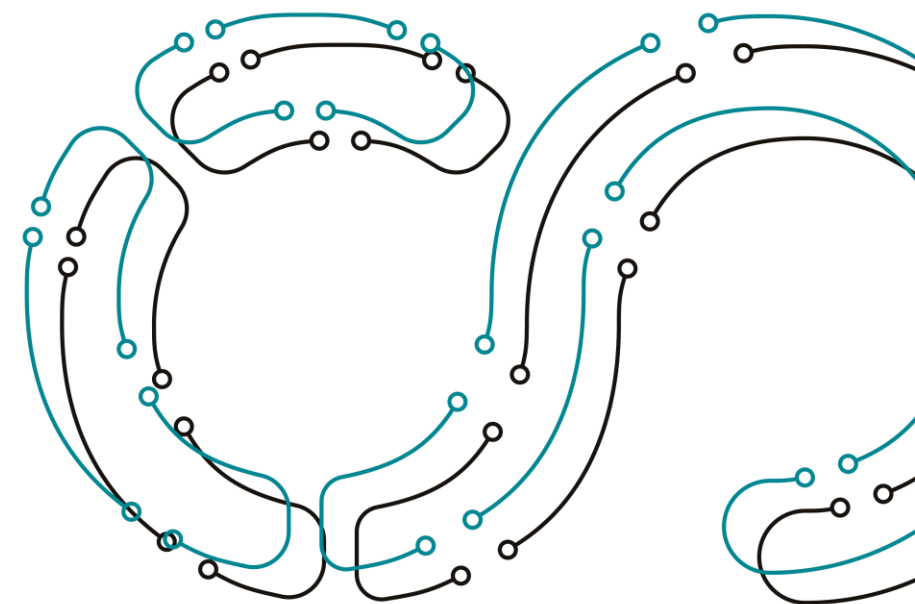
FORGE – Federated Open Research Genomics in EOSC

Federica Bazzocchi
Mariarita de Luca
Valerio Piomponi

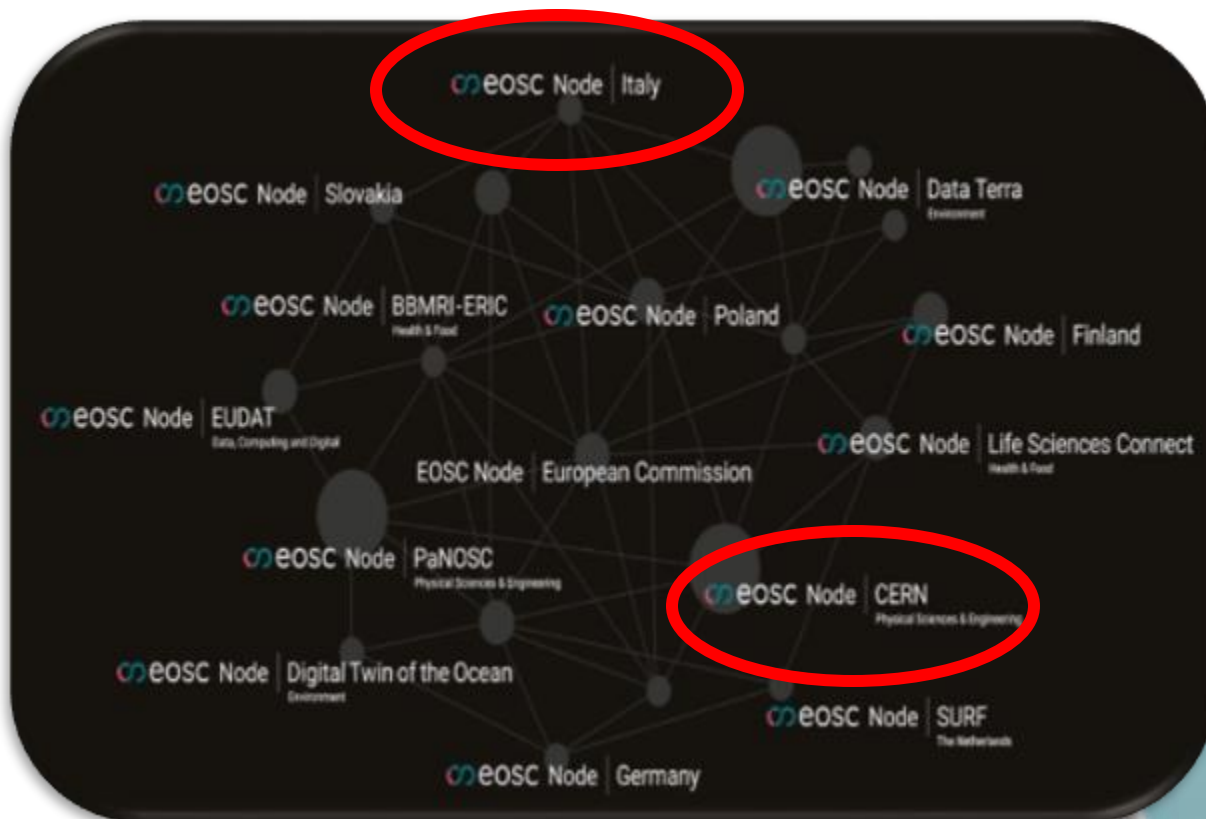
Laboratory of Data Engineering – LADE

Area Science Park

15/10/2026

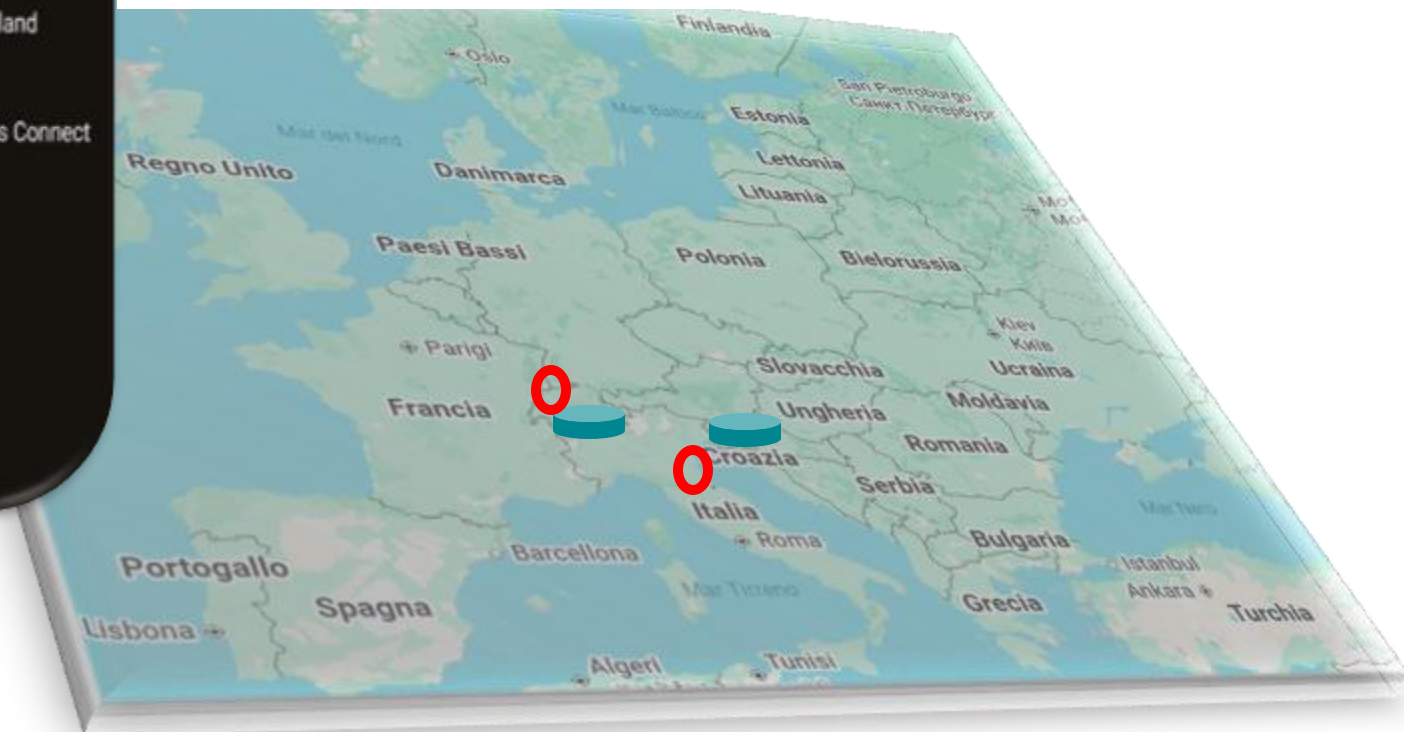


eosc What the EOSC FEDERATION can do for a researcher?



F_{indable} A_{ccessible} I_{nteroperable} R_{eusable}

Door SangyaPundir - Eigen
werk, CC BY-SA 4.0,
<https://commons.wikimedia.org/w/index.php?curid=53414062>

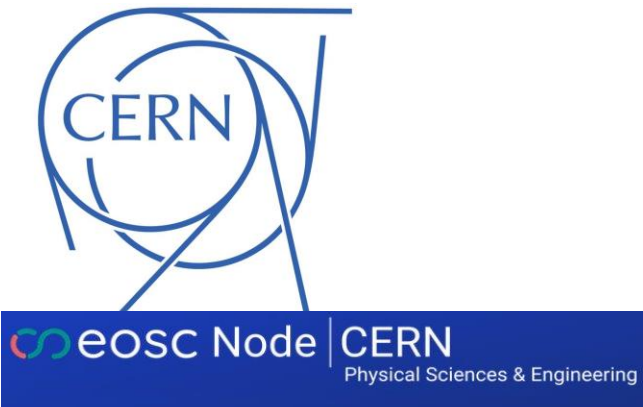


Resources (dataset, tool and computation)

EOSC Nodes

FORGE GOAL and INGREDIENTS

Reuse existing tools and resources and make them interoperate within the EOSC FEDERATION





ORFEO Data Center
(Area Science Park)

Storage (S3)
Endpoint: <https://buckets.area-sciencepark.it>
Bucket: valerio-piomponi-scout
Prefix: SPN01/



Italian EOSC Node
Catalogue / Discovery / Access

- Publishes and exposes the metadata
- Enables discovery of the dataset and services
- Provides a single access point to information



RUCIO

Federated Data Management

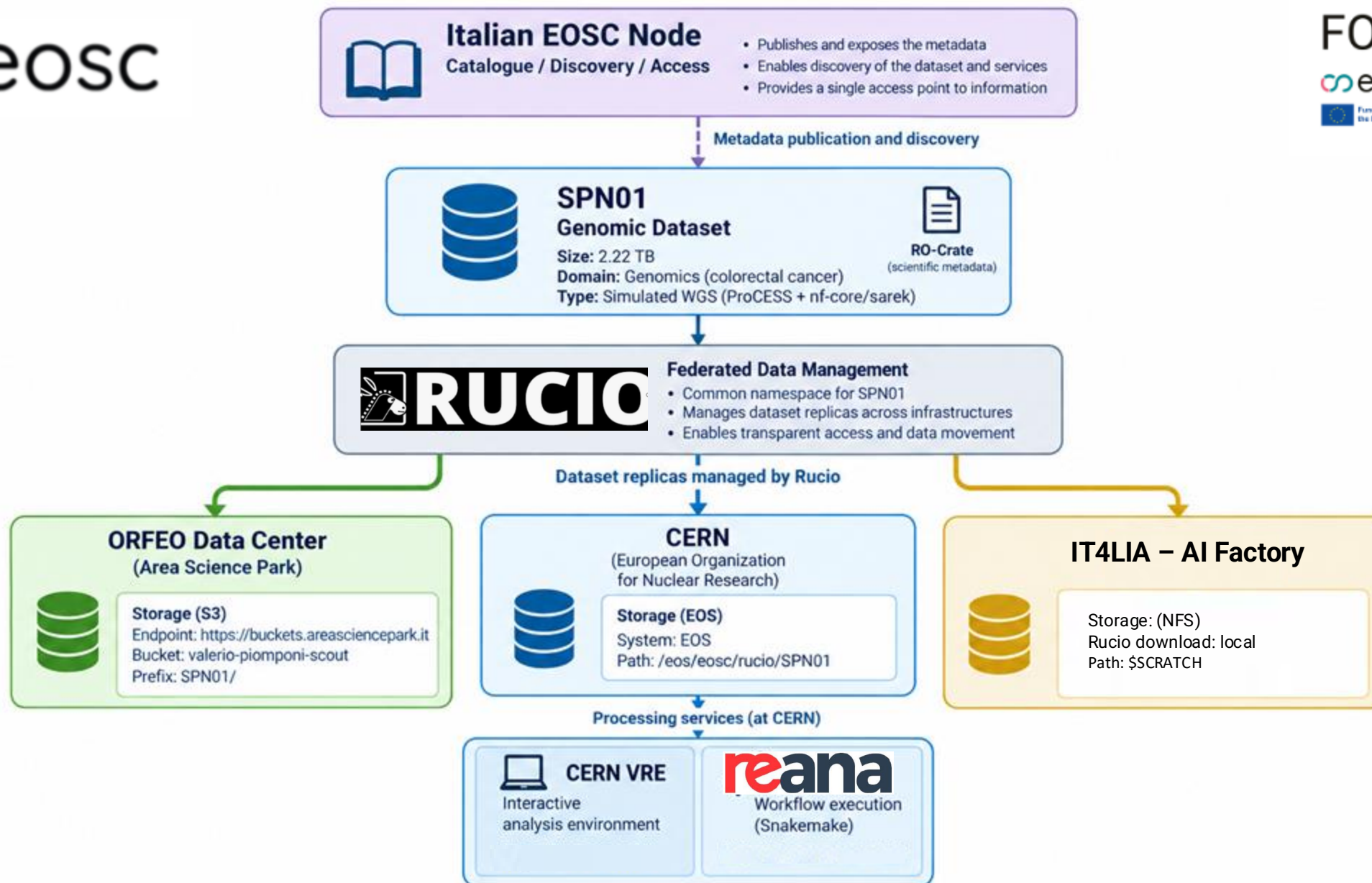
- Common namespace for SPN01
- Manages dataset replicas across infrastructures
- Enables transparent access and data movement

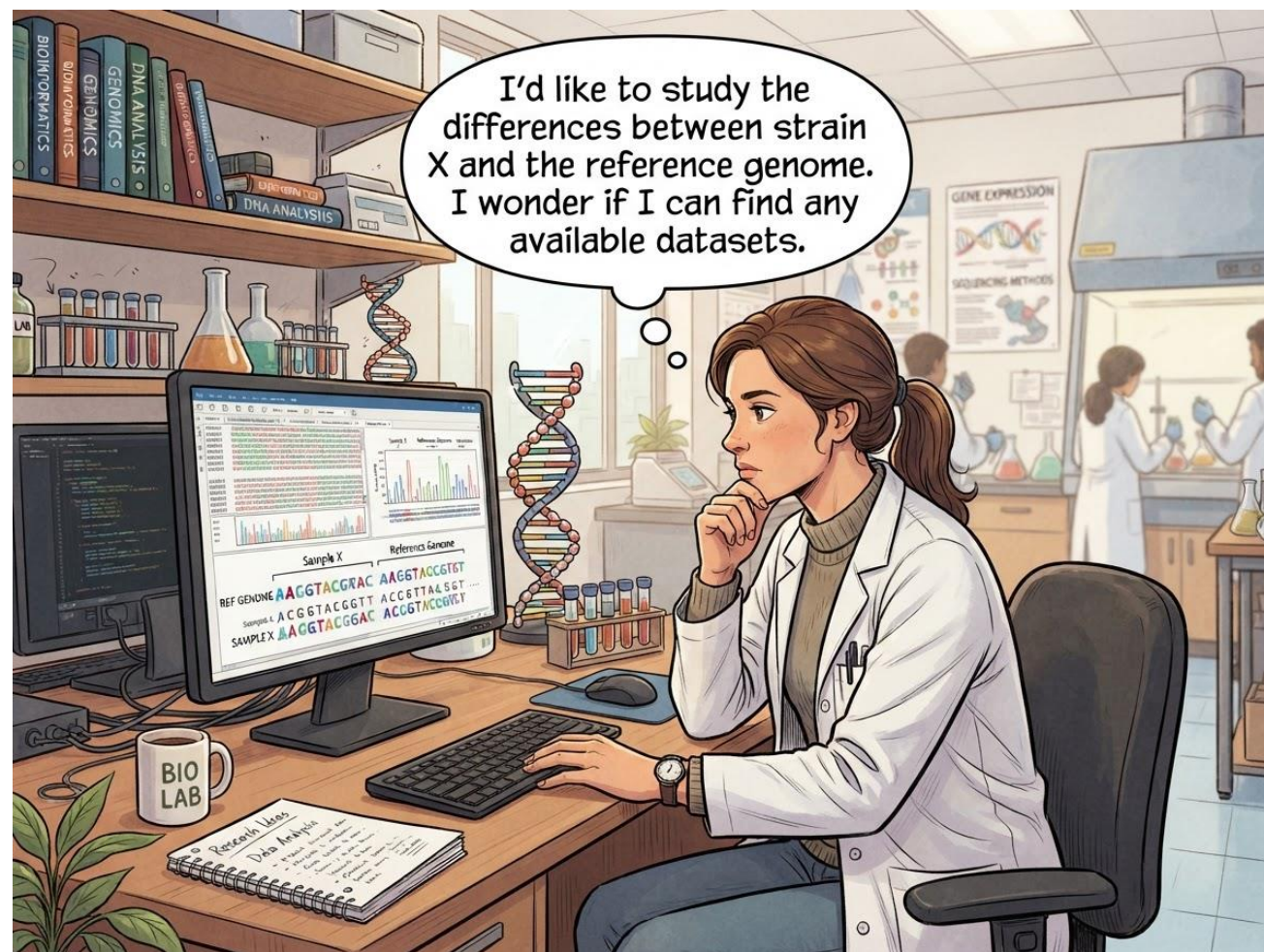


CERN VRE
Interactive analysis environment



reana
Workflow execution (Snakemake)





AI-generated image created by the authors.



EOSC Node | Italy

Welcome to EOSCNodeIT!

The EOSCNodeIT is the central hub for discovering of federated resources contributing to the EOSC Node Italy.

[See All Items](#)

[See All Tags](#)

 Service 37

Organizations

EOSC-FORGE 1 +

Groups

There are no Groups that match this search

Tags

Genomics 1 +

Metadata 1 +

Sequencing 1 +

Formats

BAM (Binary Alignment Map) 1 +

BIN 1 +

colorectal cancer

1 dataset found for " genomic data

Order by: Relevance

Genomics Dataset Repository: SPN01

EOSC-FORGE

SPN01 is a synthetic matched tumour-normal whole-genome sequencing dataset for colorectal cancer benchmarking, generated using ProCESS and analysed using nf-core/sarek. The...

BAM (Binary Alignment Map)

RDS (R Data Serialization)

FASTQ GZipped

TXT

BIN

You can also access this registry using the A

Followers 1

Unfollow

FORGE

EOSC | GRAVITY

EOSC-FORGE

Distributed data management for Genomics in the EOSC Federation This project deploys and validates a cross-node genomics scientific use case within the EOSC Federation,...

read more

Dataset Groups

Genomics Dataset Repository: SPN01

SPN01 is a synthetic matched tumour-normal whole-genome sequencing dataset for colorectal cancer benchmarking, generated using ProCESS and analysed using nf-core/sarek. The crate documents scientific provenance, data inventory, federated storage locations and available or planned research services.

Data and Resources

n0.bam

Aligned sequencing reads, produced by the nf-core/sarek pipeline from...

Explore

n1.bam

Aligned sequencing reads, produced by the nf-core/sarek pipeline from...

Explore

n2.bam

Aligned sequencing reads, produced by the nf-core/sarek pipeline from...

Explore

n3.bam

Aligned sequencing reads, produced by the nf-core/sarek pipeline from...

Explore

n4.bam

Aligned sequencing reads, produced by the nf-core/sarek pipeline from...

Explore

n5.bam

Aligned sequencing reads, produced by the nf-core/sarek pipeline from...

Explore

seq_results_muts_merged_coverage_30x.rds

R data (.rds) file with merged, cohort-level somatic mutation calls,...

Explore

More information

Go to resource

Edit resource

Views

The FORGE dataset is published in an Italian research data catalog to be onboarded into the EOSC-IT Node.

Enabling Access through Metadata



  [Download this dataset: SPN01 access point at ORFEO Data Center](#)
Genomics Dataset Repository: SPN01

@id	./
name [?]	Genomics Dataset Repository: SPN01
@type	Dataset
description [?]	SPN01 is a synthetic matched tumour-normal whole-genome sequencing dataset for colorectal cancer benchmarking, generated using ProCESS and analysed using nf-core/sarek. The crate documents scientific provenance, data inventory, federated storage locations and available or planned research services.
datePublished [?]	2026-10-09
citation [?]	Tumour evolution as ground truth for cancer whole-genome sequencing
distribution [?]	SPN01 access point at ORFEO Data Center SPN01 access point at CERN
hasPart [?]	SPN01
humanReadableSize	2.22 TB
identifier [?]	SPN01
keywords [?]	- Genomics - Metadata - Sequencing
publisher [?]	Cancer Data Science Laboratory (Caravagna Lab)
isAccessibleForFree [?]	true
subjectOf [?]	FORGE SPN01 access and processing guide



MDMC
Master in Data Management
and Curation

The dataset metadata are automatically imported from a **RO-Crate package**, preserving rich descriptive information provided by the data producer.

Additional access metadata has been included to describe the **Rucio-based access mechanism**.

Enabling Access through Metadata



  [Download this dataset: SPN01 access point at ORFEO Data Center](#)
Genomics Dataset Repository: SPN01

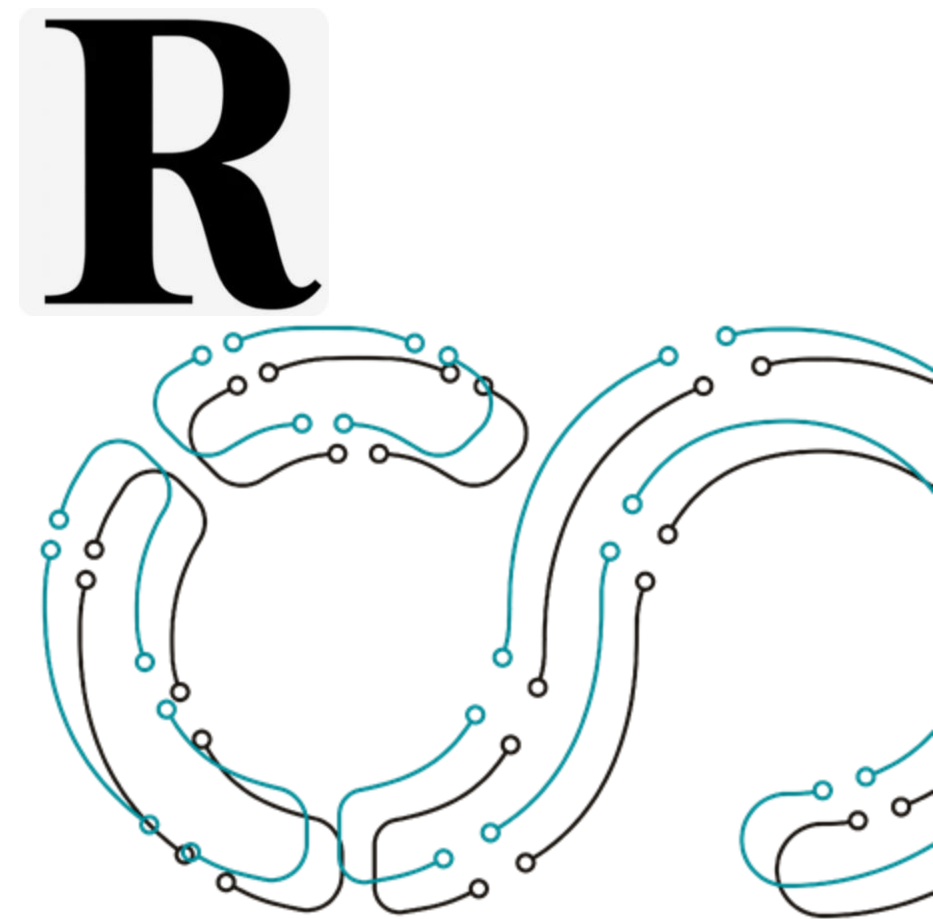
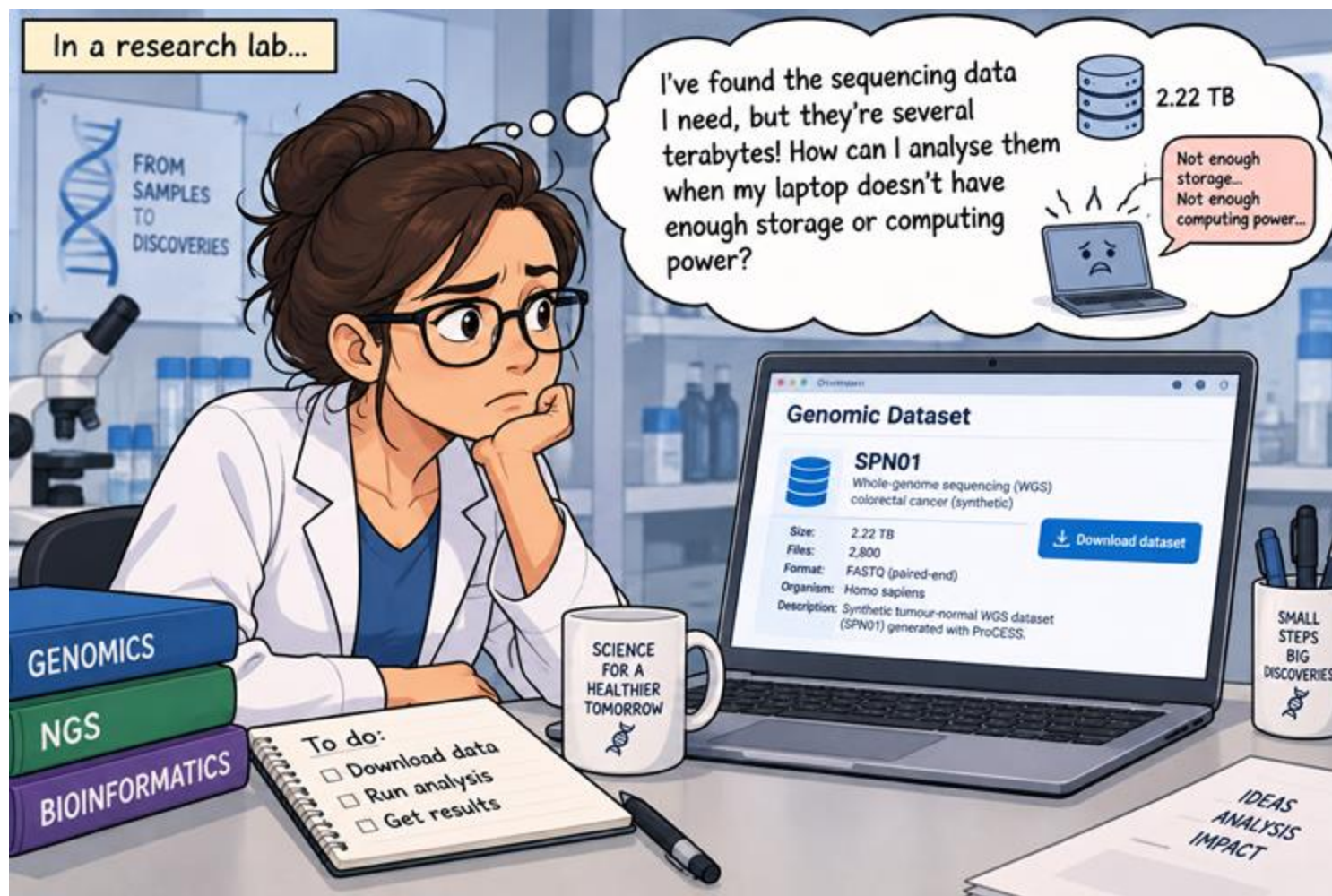
@id	./
SPN01 access point at ORFEO Data Center	
@id	#orfeo-access
name [?]	SPN01 access point at ORFEO Data Center
@type	DataDownload
description [?]	S3 object prefix; use the separate S3-compatible endpoint to access it. Access may require credentials.
additionalProperty [?]	- Protocol: S3 - Bucket: valerio-piomponi-scout - S3 object prefix: SPN01/ - S3-compatible API endpoint: https://buckets.areasciencepark.it
contentUrl [?]	s3://valerio-piomponi-scout/SPN01/
provider [?]	ORFEO Data Center
Items that reference this one	
distribution [?]	Genomics Dataset Repository: SPN01
isAccessibleForFree [?]	true
subjectOf [?]	FORGE SPN01 access and processing guide

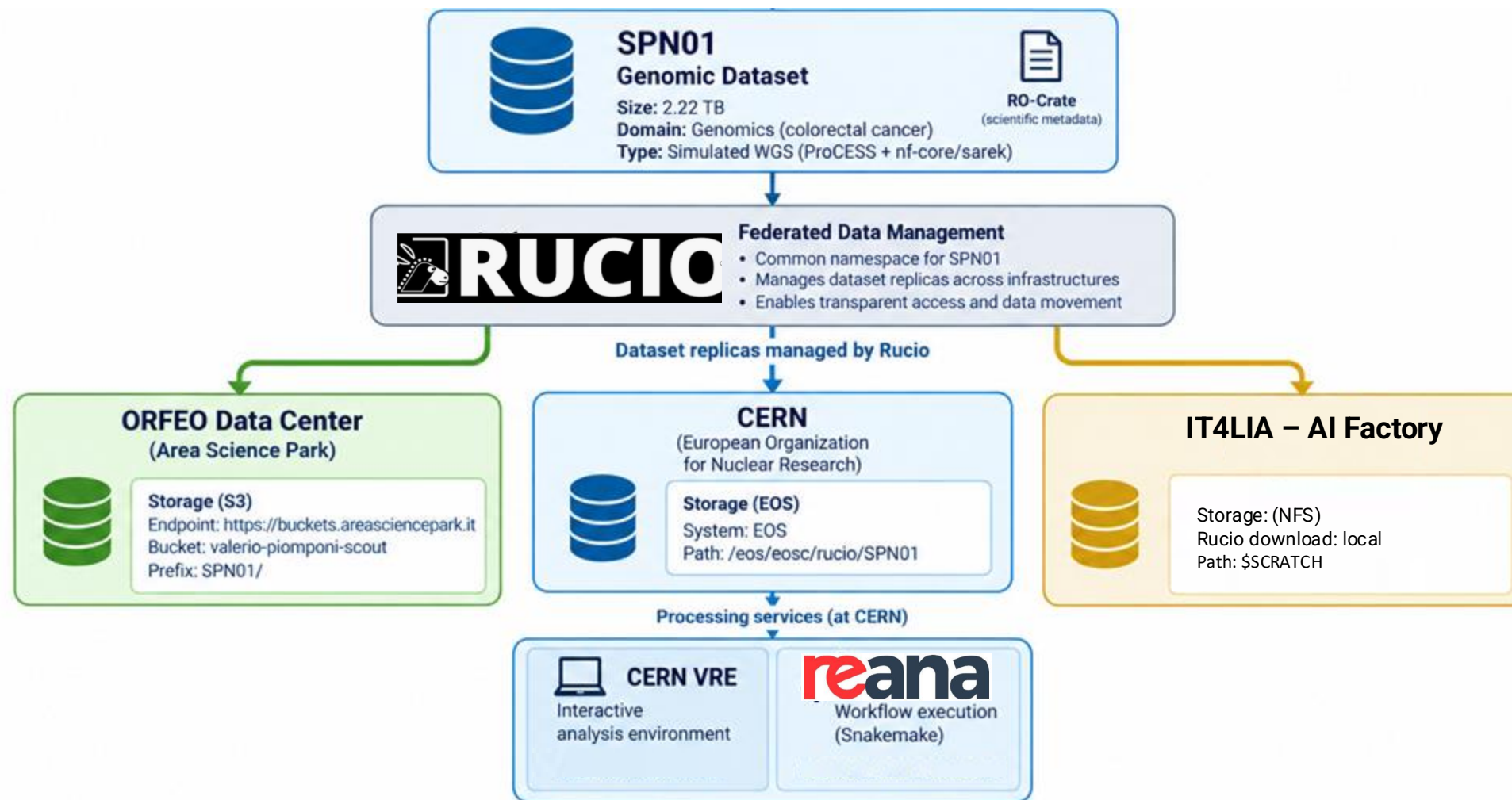


MDMC
Master in Data Management
and Curation

This allows users to
configure **Rucio**
directly from the
dataset information.

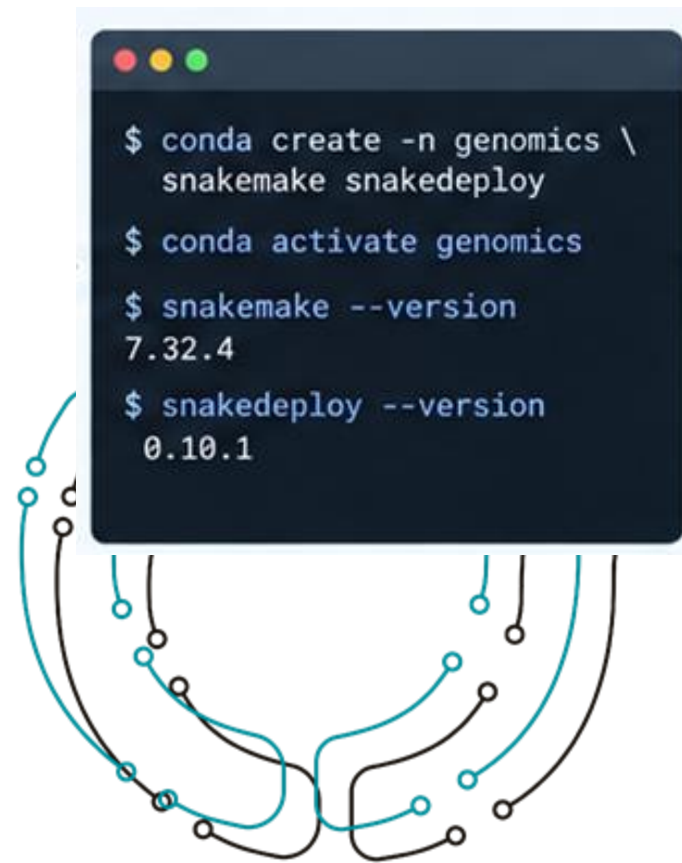
**This is the first step
toward machine
actionable data
catalogs.**





Running Genomic Analysis at CERN with Reana

- Login to CERN VRE
- Set up working environment with Snakemake and Snakedeploy
- Deploy a Workflow from the Snakemake genomic catalogue
- Edit configuration file
- Connect to Reana Server
- Create a reana.yaml file
- Run the Workflow
- Monitor the execution (check the logs)
- Download results from REANA workspace



```
$ conda create -n genomics \
  snakemake snakedeploy

$ conda activate genomics

$ snakemake --version
7.32.4

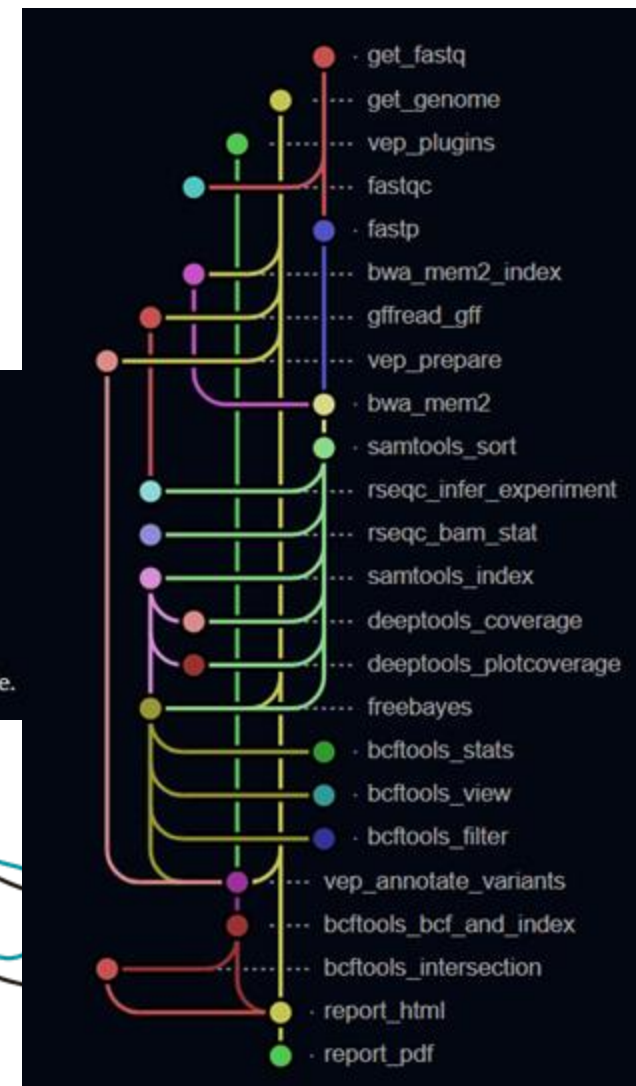
$ snakedeploy --version
0.10.1
```

reana

eosc Running Genomic Analysis at CERN with Reana

- Login to CERN VRE
- Set up working environment with Snakemake and Snakedeploy
- Deploy a Workflow from the Snakemake genomic catalogue
- Edit configuration file
- Connect to Reana Server
- Create a reana.yaml file
- Run the Workflow
- Monitor the execution (check the logs)
- Download results from REANA workspace

reana



Running Genomic Analysis at CERN with Reana

- Login to CERN VRE
- Set up working environment with Snakemake and Snakedeploy
- Deploy a Workflow from the Snakemake genomic catalogue

• Edit configuration file

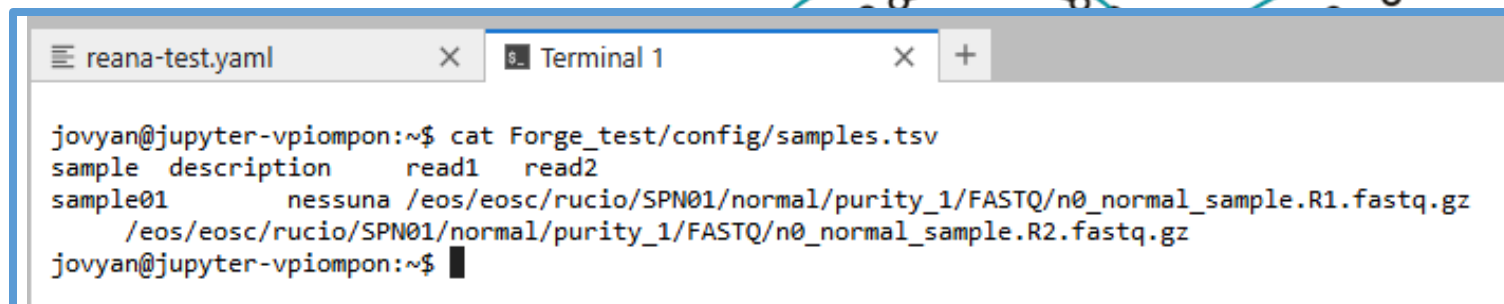
• Connect to Reana Server

• Create a reana.yaml file

• Run the Workflow

• Monitor the execution (check the logs)

• Download results from REANA workspace



```
reana-test.yaml x Terminal 1 x +
jovyan@jupyter-vpiompon:~$ cat Forge_test/config/samples.tsv
sample description read1 read2
sample01          nessuna /eos/eosc/rucio/SPN01/normal/purity_1/FASTQ/n0_normal_sample.R1.fastq.gz
                  /eos/eosc/rucio/SPN01/normal/purity_1/FASTQ/n0_normal_sample.R2.fastq.gz
jovyan@jupyter-vpiompon:~$
```



- Login to CERN VRE
- Set up working environment with Snakemake and Snakedeploy
- Deploy a Workflow from the Snakemake genomic catalogue
- Edit configuration file
- Connect to Reana Server
- Create a reana.yaml file
- Run the Workflow
- Monitor the execution (check the logs)
- Download results from REANA workspace

```
Version: 0.9.0

inputs:
  directories:
    - workflow
    - config

workflow:
  type: serial

specification:
  steps:
    - name: run-snakemake
      environment: "ghcr.io/ritareasciencepark/forged-snakemake:0.1"

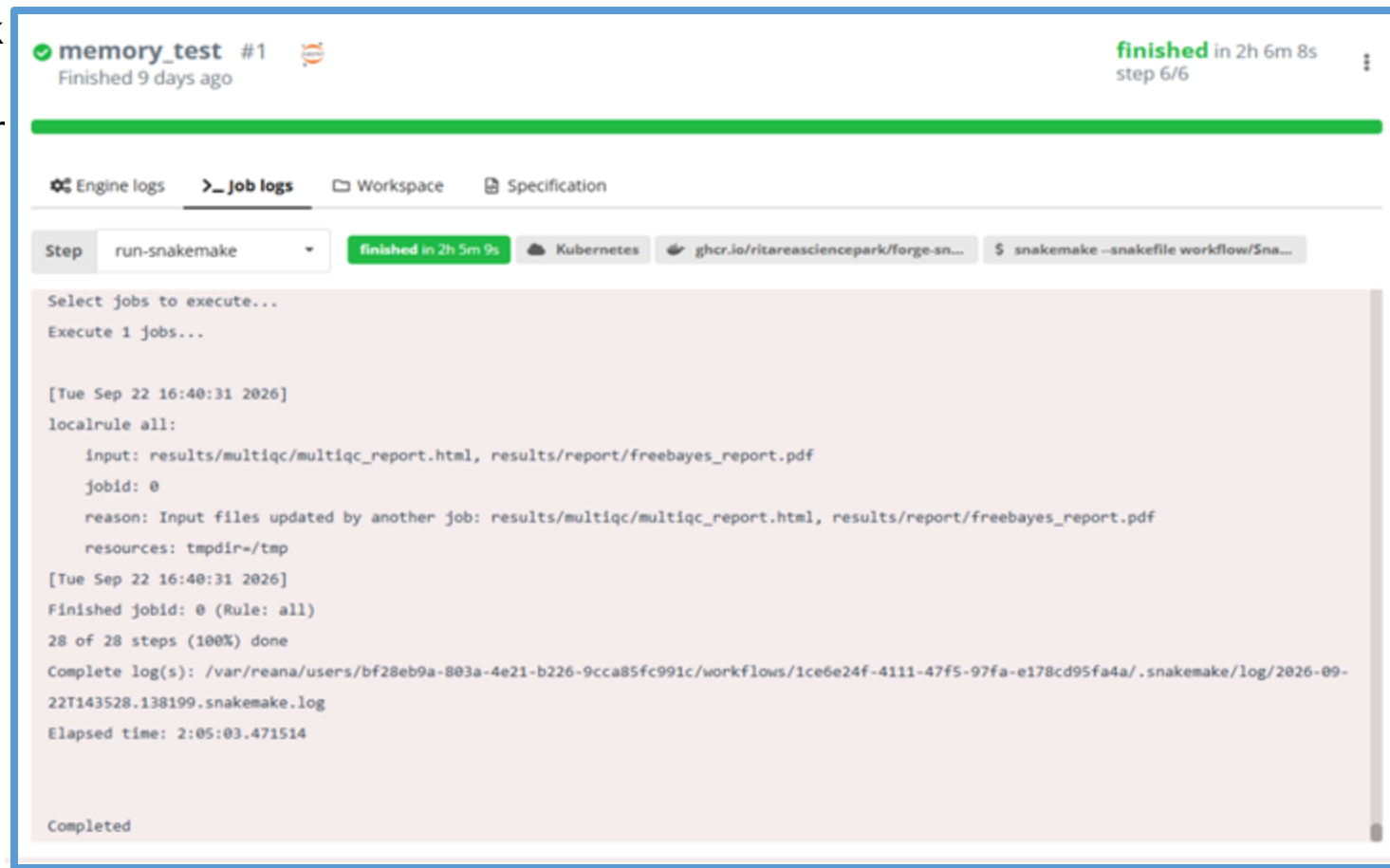
  commands:
    - python --version
    - snakemake --version
    - free -h
    - >
      echo "Cgroup memory limit:";
      cat /sys/fs/cgroup/memory.max 2>/dev/null ||
      cat /sys/fs/cgroup/memory/memory.limit_in_bytes 2>/dev/null ||
      true
    - >
      echo "Cgroup CPU:";
      cat /sys/fs/cgroup/cpu.max 2>/dev/null || true
    - >
      snakemake
      --snakefile workflow/Snakefile
      --cores 4
      --sdm conda

outputs:
  directories:
    - results
```

Running Genomic Analysis at CERN with Reana

- Login to CERN VRE
- Set up working environment with Snakemake
- Deploy a Workflow from the Snakemake gallery
- Edit configuration file
- Connect to Reana Server
- Create a reana.yaml file
- Run the Workflow
- Monitor the execution (check the logs)
- Download results from REANA workspace

reana



```
memory_test #1
Finished 9 days ago
finished in 2h 6m 8s
step 6/6

Engine logs > Job logs Workspace Specification

Step run-snakemake finished in 2h 5m 9s Kubernetes ghcr.io/ritareasciencepark/forg... $ snakemake --snakefile workflow/Sna...

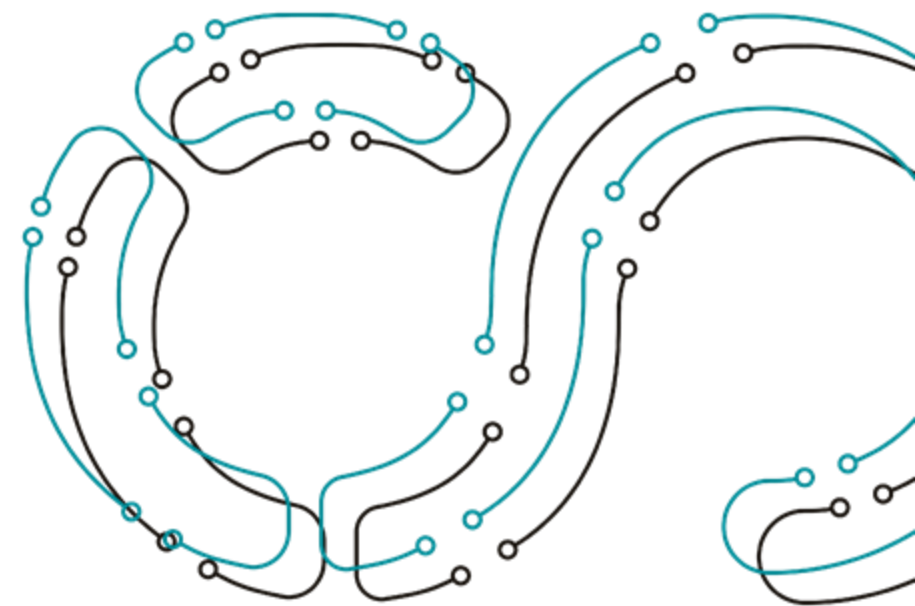
Select jobs to execute...
Execute 1 jobs...

[Tue Sep 22 16:40:31 2026]
localrule all:
  input: results/multiqc/multiqc_report.html, results/report/freebayes_report.pdf
  jobid: 0
  reason: Input files updated by another job: results/multiqc/multiqc_report.html, results/report/freebayes_report.pdf
  resources: tmpdir=/tmp
[Tue Sep 22 16:40:31 2026]
Finished jobid: 0 (Rule: all)
28 of 28 steps (100%) done
Complete log(s): /var/reana/users/bf28eb9a-803a-4e21-b226-9cca85fc991c/workflows/1ce6e24f-4111-47f5-97fa-e178cd95fa4a/.snakemake/log/2026-09-22T143528.138199.snakemake.log
Elapsed time: 2:05:03.471514

Completed
```

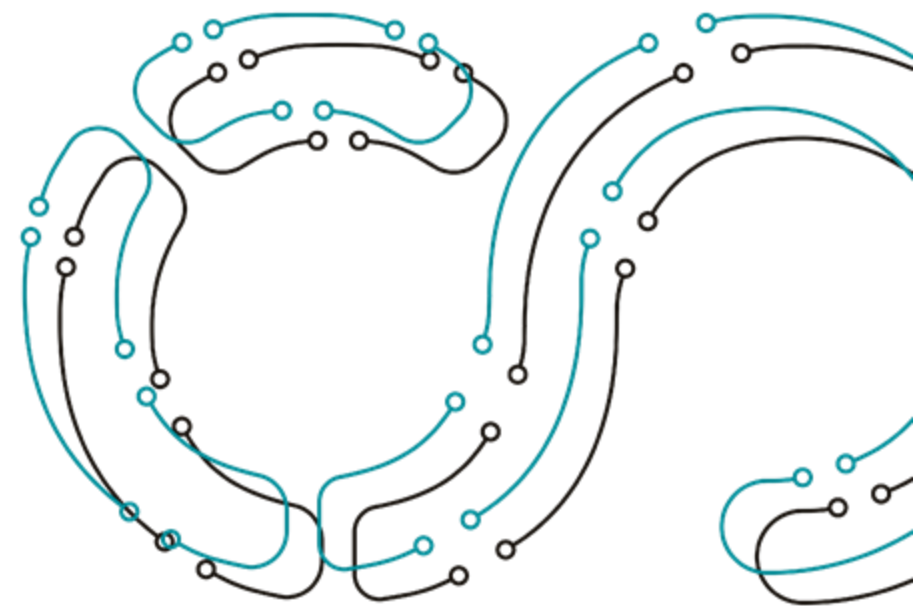
FORGE established a clear FAIR workflow enabling researchers to:

- **find data → obtain access → transfer data to available computing resources → run analysis pipelines.**
- The workflow is not yet fully automated, but its main components and interactions have been successfully validated.
- Metadata are **interoperable and machine-actionable**, using **RO-Crate** to describe data, workflows, and provenance.
- The genomics use case demonstrated that **Rucio is a viable solution** to support scientific domains beyond High-Energy Physics in the EOSC Federation.
- The approach is portable across heterogeneous storage and computing infrastructures. **Data transfer to the AI4Italy AI Factory resources has already been successfully tested.**



Future extension of FORGE and beyond:

- **Automate the end-to-end workflow**, from data discovery to analysis execution.
- Leverage the **role-based access control functionalities** being developed by CERN for the secure management of sensitive Genomic data.
- Extend the workflow to **additional EOSC Federation storage resources**, starting with the already identified **StoRM/INFN Bari solution**.





PEOPLE



MDMC

Master in Data Management
and Curation



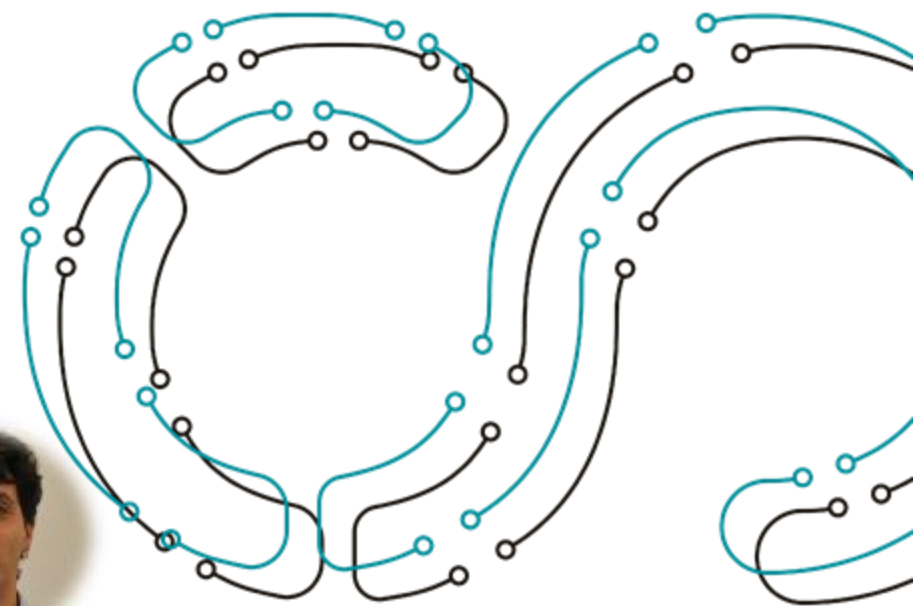
Federica Bazzocchi, Mariarita de Luca e Valerio Pionponi
Lesly Fougang e Nicola Perin (MDMC fellows)



Giovanni Guerrieri



Arnaud Ceol,
Letizia Magenta
e Alessandro Nicolini



Thank You for Your Attention!

