

EOSC-Pillar

Coordination and Harmonisation of National & Thematic Initiatives to support EOSC

Development of a use case in bioinformatics within the EOSC-Pillar project

Marwa BELHAJ-SALEM, Gilles MATHIEU

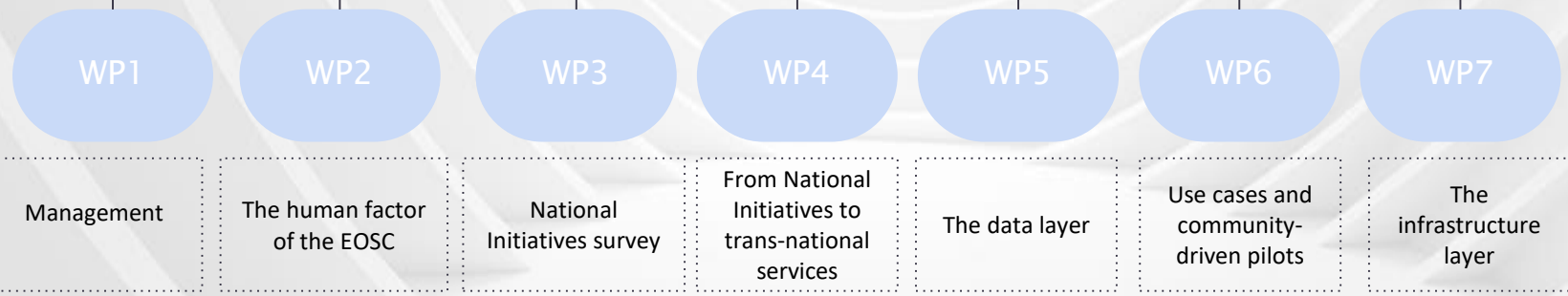
INSERM DSI – SSDUN/Domis
Dec. 2022



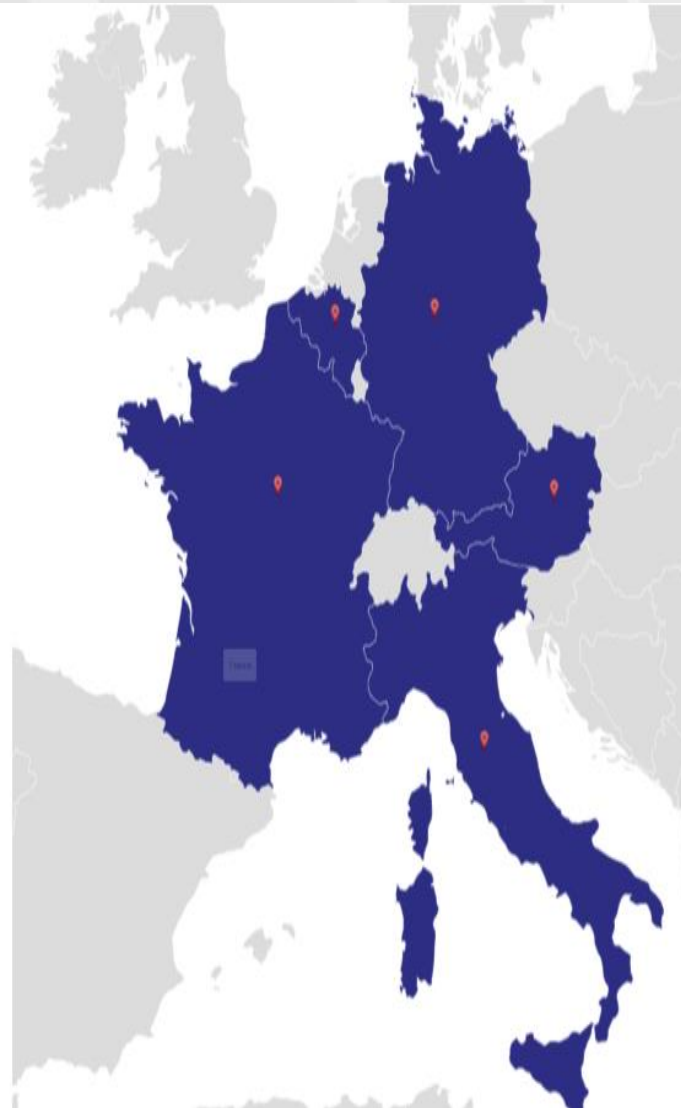
Position of EOOSC-Pillar in EOOSC



INFRAEOOSC-05b call projects
"Supporting the Overall Governance of EOOSC"



What's EOSC-Pillar



W

what

EOSC-Pillar – Coordination and harmonization of national initiatives, infrastructures and data services in Central and Western Europe

W

whom

A consortium of 18 partners

W

Where

France, Belgium, Germany, Italy, Austria

W

When

July 2019 - December 2022

H

How

- ✓ Coordinates with regional EOSC projects
- ✓ Supports EOSC implementation by building on national and thematic initiatives developed by research communities.
- ✓ Coordinates data infrastructures and services by bringing together leaders of national initiatives.
- ✓ Various services for management, analysis, and storage

W

Why

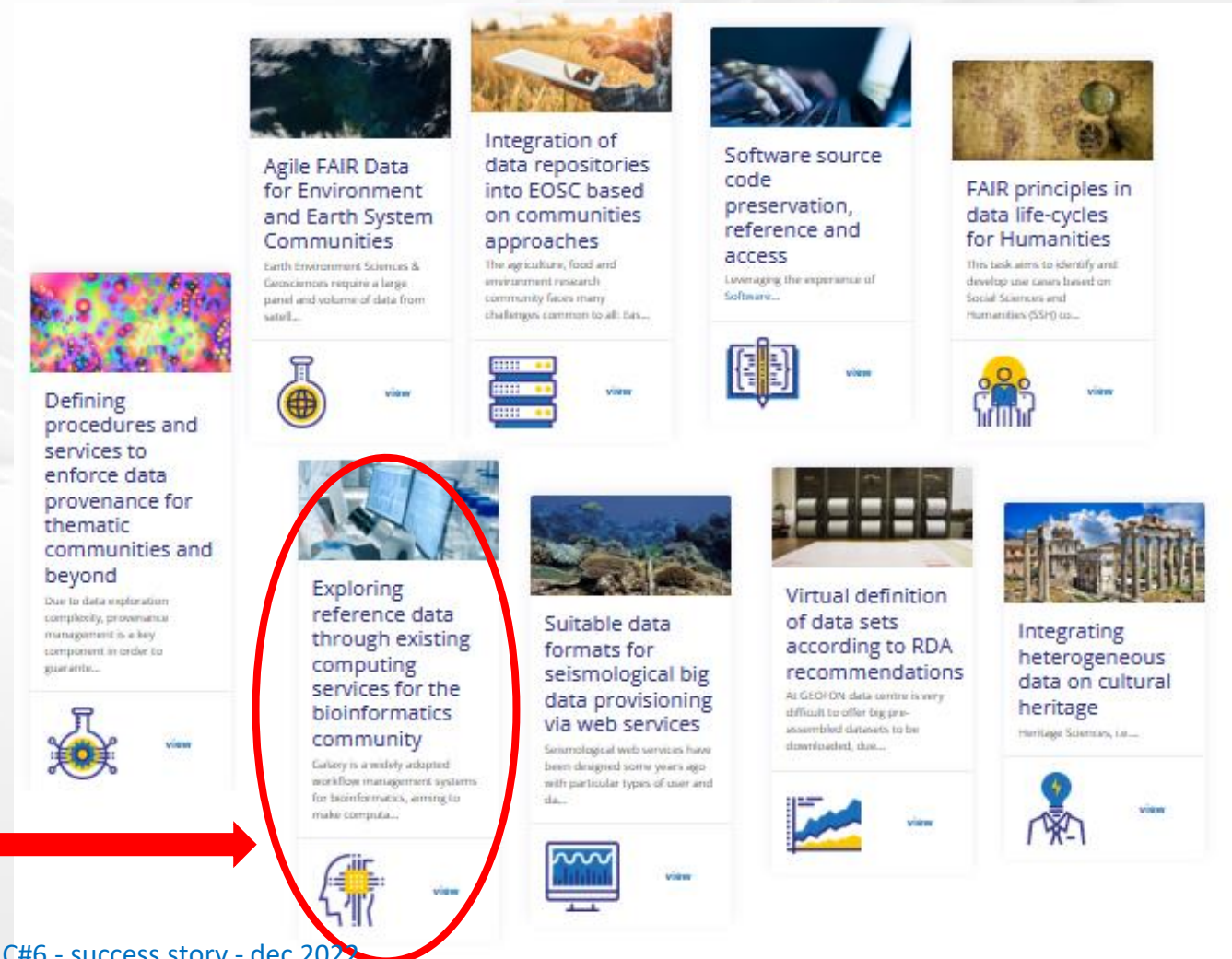
- ✓ Better European-level exchange of research data (different fields)
- ✓ Data reuse across borders and scientific disciplines through the federation of existing infrastructures and services

Biomedical Use Case within EOSC-Pillar



Work Package 6

Use-case #6 : Exploring reference data through existing computing services for the bioinformatics community



Agile FAIR Data for Environment and Earth System Communities

Integration of data repositories into EOSC based on communities approaches

Software source code preservation, reference and access

FAIR principles in data life-cycles for Humanities

Defining procedures and services to enforce data provenance for thematic communities and beyond

Exploring reference data through existing computing services for the bioinformatics community

Suitable data formats for seismological big data provisioning via web services

Virtual definition of data sets according to RDA recommendations

Integrating heterogeneous data on cultural heritage

Presentation of use case 6.6



W
What

T6.6 : Investigate existing computing services for the bioinformatics community's reference data

W
Whom

Led by Inserm
Other partners: IBIOM, INFN; IFB

W
Where

France, Italy

W
When

July 2019 - December 2021

H
How

✓ Enhances existing national services in France and Italy: Galaxy , F2DS, D4science..

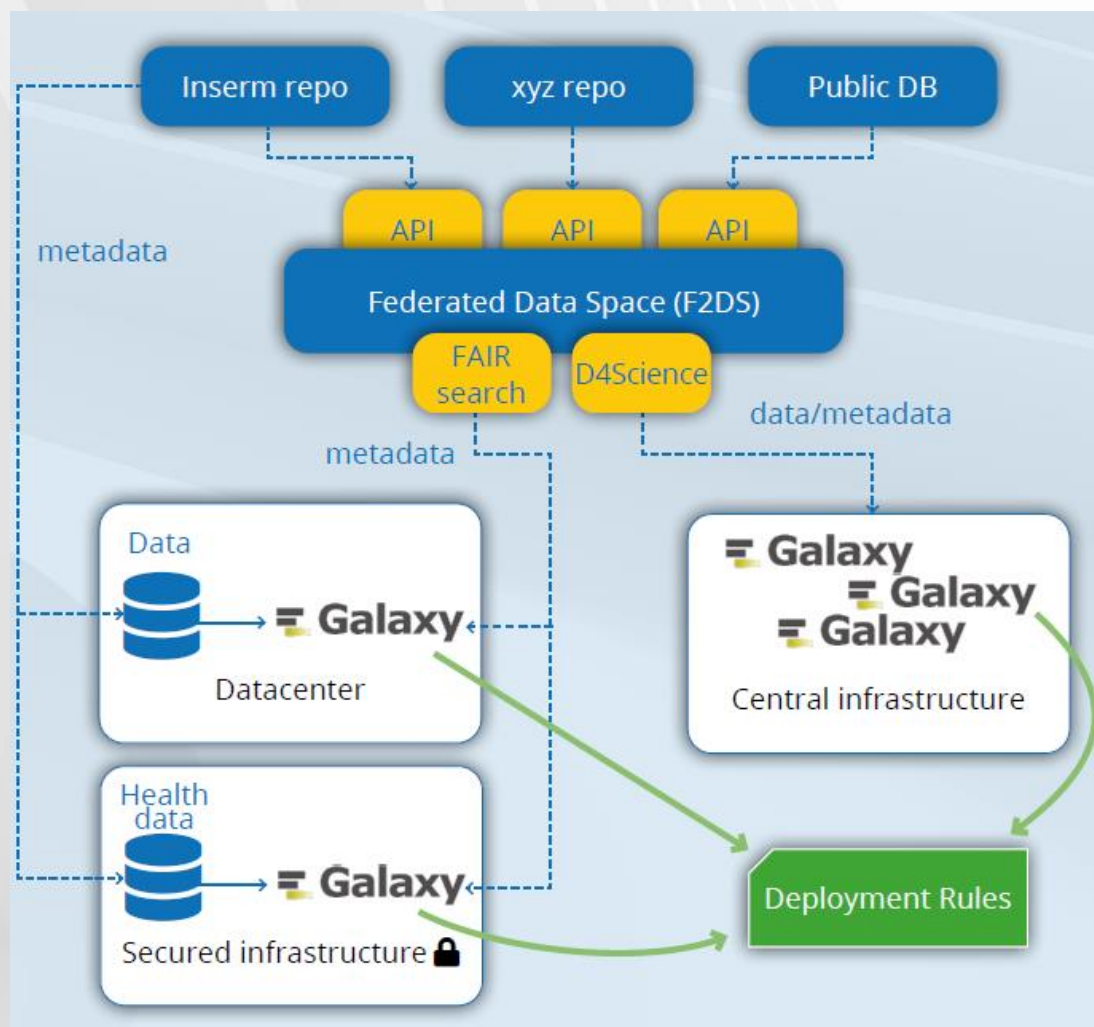
W
Why

- ✓ Different Galaxy deployments' reproducibility and consistency: ensure the same results regardless of the Galaxy instance used
- ✓ Make it simple to connect Galaxy to data sources.
- ✓ Personal health data protection: deployment in a private and secure environment

UC#6 challenges

- * Reproducibility and coherency of the various deployments
- * How to implement Galaxy in a private, secure environment with a data analysis workflow similar to that of its public equivalent
- * Integrating the service into a global authentication system
- * Make the service available to all members of the EOSC community

UC#6: what did we want to do?



4 theoretical scenarios

- 1 Use public Galaxy Instances
- 2 Deploy Galaxy locally
- 3 Deploy Galaxy on a secured infrastructure
- 4 Compare results of the 3 scenarios above

UC#6: what did we actually do?

- Implement and use better rules for Galaxy deployment
- Provide easily deployable Galaxy instances close to the data
- Publish and populate source/reference data to the F2DS
- Deploy Galaxy in a secured environment
- Provide a working demonstrator with a light workflow
- Analyse transnational health data restrictions
- Defined a full scale "real life" workflow with the hCNV community

UC#6: How did that go?



Worked well

- * Collaborating internationally
- * Publishing data to F2DS
- * Enriching metadata in F2DS
- * Conceiving a Galaxy workflow
- * Using Laniakea@ReCAS
- * Deploying Galaxy locally
- * Deploying on a secured infra

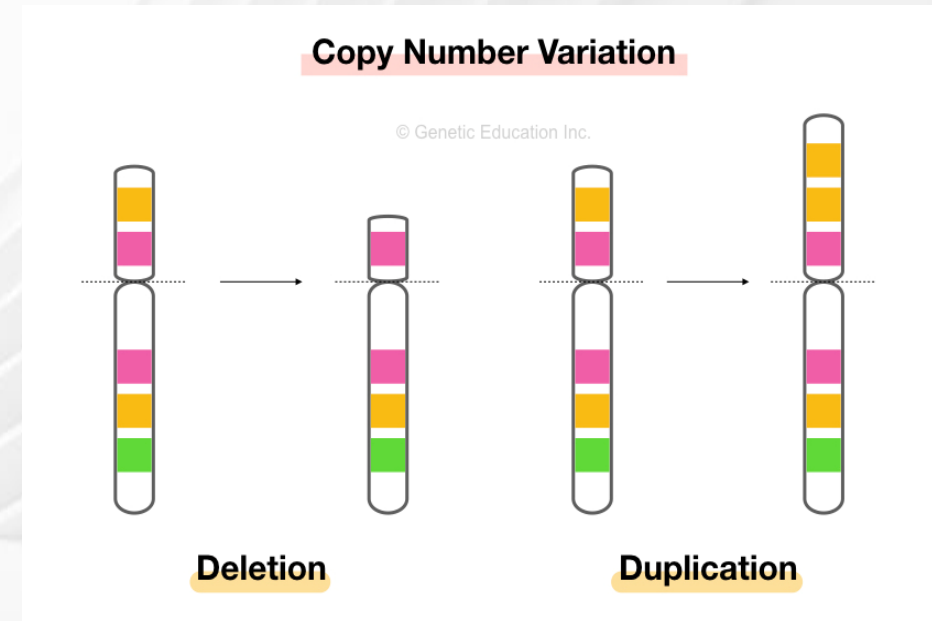


Didn't work so well

- * Running workflow on huge files (memory problem)
- * Connecting F2DS to Galaxy (AAI problem)
- * Directly using F2DS fetch method from Galaxy (function missing)

A concrete application: hCNV tool benchmarking (1)

- hCNV: Human Copy Number Variation
 - Genome modification during mitosis
 - Duplication : a gene is copied twice or more
 - Deletion: a gene is not copied
 - Play a role in some diseases
- hCNV detection
 - Represents a major challenge
 - Needs tools: “CNV callers”



A concrete application: hCNV tool benchmarking (2)

- Reference hCNV data
 - produced by NIST (National Institute of Standards and Technology, US)
 - <https://www.nist.gov/programs-projects/genome-bottle>
- Benchmarking of hCNV callers
 - Run hCNV caller on a reference sample
 - Compare results with the NIST “gold standard”
- One of the tasks of the hCNV Elixir community
 - <https://elixir-europe.org/communities/hcnv>

More info

UC#6 page on the EOSC-Pillar web site <https://eosc-pillar.eu/use-cases/exploring-reference-data-through-existing-computing-services-bioinformatics-community>

UC#6 “fact sheet”

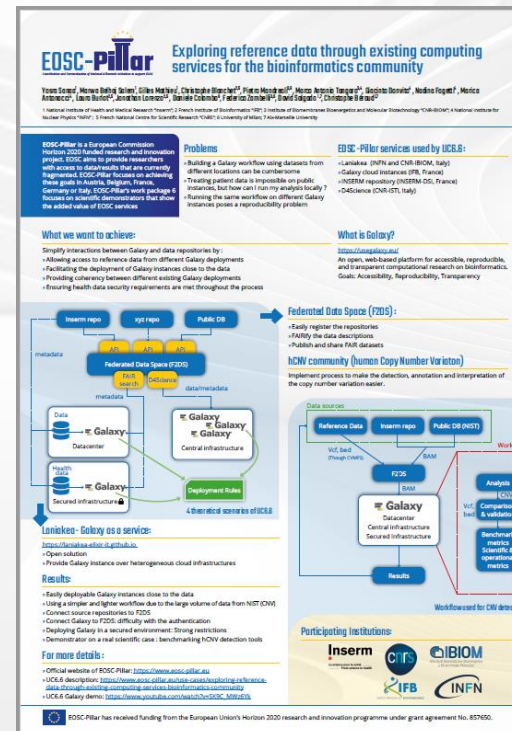
<https://doi.org/10.5281/zenodo.6726022>

UC#6 video demo

<https://youtu.be/WZey8XrCp1I>

UC#6 poster

<https://doi.org/10.5281/zenodo.7051283>



EOSC-Pillar Exploring reference data through existing computing services for the bioinformatics community

EOSC-Pillar is a European Commission Horizon 2020 funded research and innovation project. EOSC aims to provide researchers with access to data/results that are currently fragmented. EOSC-Pillar focuses on achieving three goals in Austria, Belgium, France, Germany and Italy. EOSC-Pillar work packages focuses on scientific demonstrators that show the added value of EOSC services.

Problems

- Building a Galaxy workflow using datasets from different locations can be cumbersome
- Trusting patient data is responsible in public instances, but how can I run my analysis locally?
- Running the same workflow on different Galaxy instances poses a reproducibility problem.

EOSC-Pillar services used by UCB.R:

- Lanikai (JPM and CNR-ISCN, Italy)
- Galaxy cloud instances (FR, France)
- INSERM repository (INSERM DS, France)
- Chelonea (CNRS-ICT, Italy)

What we want to achieve:

- Simplify interactions between Galaxy and data repositories by:
 - allowing access to reference data from different Galaxy deployments
 - facilitating the deployment of Galaxy instances close to the data
 - providing coherency between different existing Galaxy deployments
 - ensuring health data security requirements are met throughout the process

What is Galaxy?

<https://galaxyproject.org/>

An open, web-based platform for accessible, reproducible, and transparent computational research on bioinformatics. Goals: Accessibility, Reproducibility, Transparency.

Federated Data Space (FDS):

- Easily register the repositories
- Unify the data descriptions
- Publish and share FAIR datasets

hCN community (Human Copy Number Variation)

Implement process to make the detection, annotation and interpretation of the copy number variation easier.

Lanikai - Galaxy as a service:

<https://lanikai.euro.ub.edu/>

- Open solution
- Provide Galaxy instance over heterogeneous cloud infrastructures

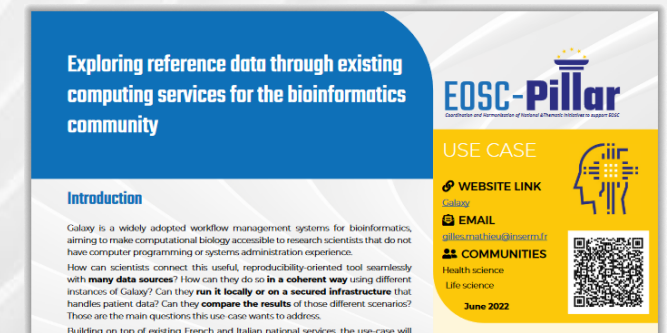
Results:

- Easily register Galaxy instances close to the data
- Using a simpler and lighter workflow due to the large volume of data from INSERM
- Connect source repositories to FDS
- Connect Galaxy to FDS efficiently with the authentication
- Deploying Galaxy in a secured environment. Strong restrictions
- Demonstrate on a real scientific case: benchmarking hCN detection tools

For more details:

- Official website of EOSC-Pillar: <https://www.eosc-pillar.eu>
- UCR6 description: <https://www.eosc-pillar.eu/use-cases/exploring-reference-data-through-existing-computing-services-bioinformatics-community>
- UCR6 Galaxy demo: https://www.youtube.com/watch?v=U5DQK_Mt9z0

EOSC-Pillar has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 852600.



Exploring reference data through existing computing services for the bioinformatics community

USE CASE

WEBSITE LINK

EMAIL

COMMUNITIES

June 2022

Introduction

Galaxy is a widely adopted workflow management systems for bioinformatics, aiming to make computational biology accessible to research scientists that do not have computer programming or systems administration experience.

How can scientists connect this useful, reproducibility-oriented tool seamlessly with **many data sources**? How can they do so in a **coherent way** using different instances of Galaxy? Can they **run it locally** or on a **secured infrastructure** that handles patient data? Can they **compare the results** of those different scenarios? Those are the main questions this use case wants to address.

Building on top of existing French and Italian national services, the use-case will



Use case 6 - Reference data through existing computing services ...

Watch Later Share

Watch on YouTube

EOSC-Pillar

Coordination and Harmonisation of National & Thematic Initiatives to support EOSC

Thank you!

Get in touch with us!



www.eosc-pillar.eu



[@EoscPillar](https://twitter.com/EoscPillar)



[/company/eosc-pillar](https://www.linkedin.com/company/eosc-pillar)

