

Scientific Collaboration Agreement between EOSC-Pillar and EOSC-Life

I. EOSC-Life and EOSC-Pillar

EOSC-Life brings together the 13 Life Science 'ESFRI' Research Infrastructures (RIs) to create an open, digital, and collaborative space for biological and medical research. The project will publish 'FAIR' data and a catalogue of services provided by participating RIs for the management, storage, and reuse of data in the European Open Science Cloud.

In particular, the EOSC-Life Work Package 2 (*Tools Collaboratory*) plans to achieve the following goals:

- Make tools and workflows interoperable and reusable in the EOSC across RIs.
 - Containerization technology.
 - Workflow development.
 - Supporting the adoption and development of Galaxy at all RIs.
- Tool findability and registries.
- Support and address demonstrator projects.
- The Tools Collaboratory.
 - Community building: workshops, hackathons and outreach.
 - Making tools "cloud-native".
 - Documentation, training material.
- Sustainability.

The roadmap of the work package includes the software stack on which the services are being built, as well as the interaction with the underlying infrastructure (WP7) and the different user profiles (Fig. 1).

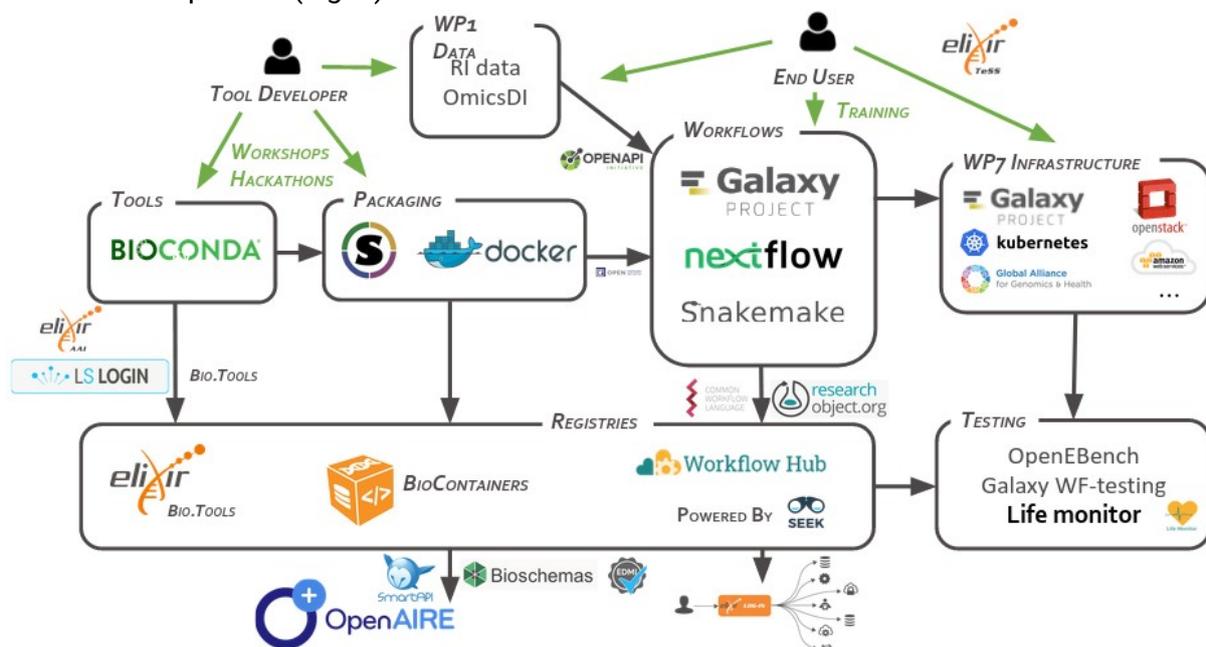


Figure 1. EOSC-Life Tools Collaboratory roadmap.

Similarly, [EOSC-Pillar](#) coordinates national EOSC initiatives in different countries: Austria, Belgium, France, Germany, and Italy. The project follows a bottom-up approach to fulfil the needs of scientific communities and a top-down one to harmonize the different national strategies.

More specifically, Work Package 6 (*EOSC in action: Use cases and community-driven pilots*) addresses the needs of different use cases. Of interest in this collaboration agreement is the use case in task 6.6, "[Exploring reference data through existing computing services for the bioinformatics](#)", which explores the interactions between already available Galaxy computing services and data repositories. New features of Laniakea, a software framework that facilitates the provisioning of on-demand Galaxy instances as a cloud service over e-infrastructures, is developed in this use case. It leverages the open-source software platform developed by the INDIGO-DataCloud H2020 project, which aimed to make cloud e-infrastructures more accessible by scientific communities.

Both projects are collaborating to provide on-demand Galaxy instances, and beyond, in isolated cloud environments, accessible only through virtual VPN by users. This, together with the on-demand Storage Encryption already in place, would provide the possibility to deploy secure instances suitable for sensitive data analysis. Being Laniakea among the EOSC-Life WP7 managed services, these improvements will be available to both communities.

Work Package 7 (*The infrastructure layer: delivering horizontal data storage and computing services, from national to transnational*), as in EOSC-Life, provides the infrastructure to run the services. Laniakea@RECAS is a "Galaxy as a service" platform provided in task 7.4 (*Services ready to use*), which covers the use case in Task 6.6, and the [Galaxy-E](#) initiative dedicated to biodiversity communities.

As can be derived from the overlap of the two projects described above, similar technologies are sustaining the services provided. Moreover, EOSC-Pillar uses Bioconda, BioContainers and plans to use services delivered by EOSC-Life, such as the [WorkflowHub](#) registry.

While EOSC-Life is working on workflow FAIRification with the WorkflowHub, EOSC-Pillar is trying to improve data FAIRness, allowing to find, access, aggregate and reuse the vast amount of existing data, developing the [Federated FAIR Data Space](#) (F2DS). F2DS will provide a unified search and data-retrieval solution over multiple data sets from heterogeneous and distributed community-specific repositories.

II. Ongoing collaboration

The [Galaxy-E](#) initiative is relying on the services offered by the European Galaxy Project. Thus, for development purposes we use a GitHub repository to both share ongoing development, centralise reviews and contributions and automatically deploy Galaxy tools on the main Galaxy App Store. The Galaxy-E subdomain (<https://ecology.usegalaxy.eu>) of the European Galaxy server is now used at the French level for ecology-related data analysis and internationally for the creation [GEO BON](#) Biodiversity indicators. Combined with the work done on data and metadata management through the use of the Ecological Metadata

Language (EML) standard, Galaxy-E development guidelines pave the way to fundamental operational best practices for FAIRification of Ecological research objects thanks to the BiodiFAIRse GO FAIR Implementation Network initiative.

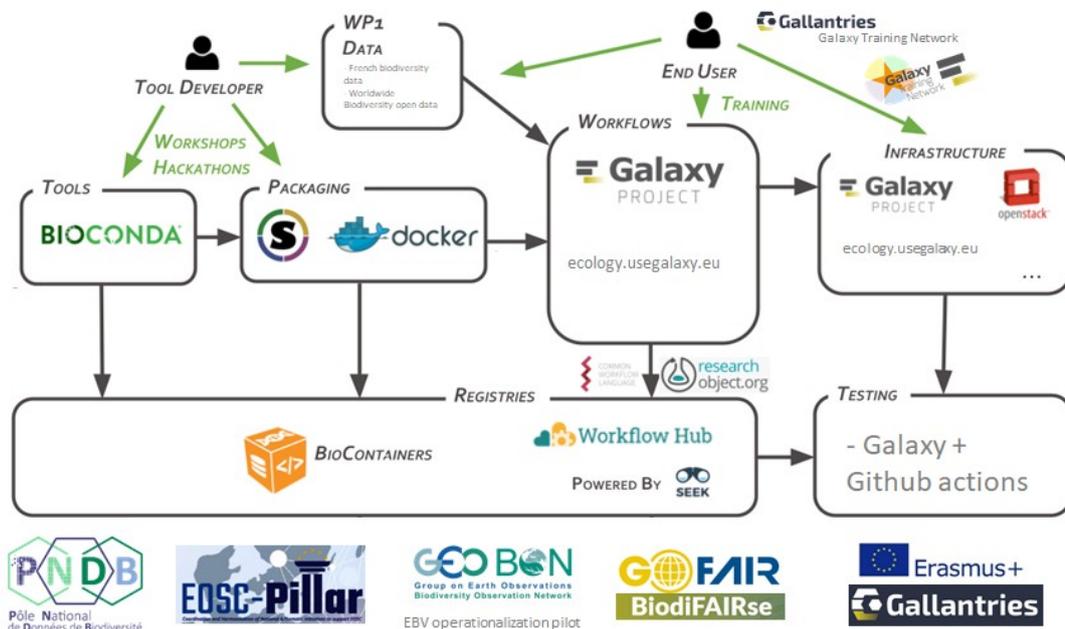


Figure 2. “Galaxy for Ecology” initiative position reusing the EOSC-Life Collaboratory roadmap elements.

Currently, both EOSC-Pillar and EOSC-Life exploit a set of common tools to support their use cases, which consequently are actively supported by both projects.

The Galaxy workflow management system plays a key role in both projects, through the European Galaxy server for EOSC-Life and the on-demand service Laniakea@ReCaS for EOSC-Pillar. Moreover, cloud deployment of Galaxy is performed using the same set of Ansible roles, provided by the Galaxy community.

To access reference data, e.g., genomic sequences, genome indexes, variant databases, and more. To avoid unnecessary and costly data replication, those reference data are usually made available, in read-only mode, to users through a Cern-VM File System (CVMFS). Indeed, both the European Galaxy server and Laniakea exploit the Galaxy project CVMFS to provide reference data and tools. In this context, the Consortium GARR, partner of the EOSC-Pillar project, will soon provide a mirror for the CVMFS, to better support the Laniakea platform.

Bioinformatics tools employed in the two projects are usually packaged using Conda and published in the Bioconda channel. This ensures package availability on different platforms and fosters analytical reproducibility, allowing the installation of different package versions and avoiding the introduction of compatibility issues during the tool's lifetime. Moreover, each Bioconda package is automatically packaged as a Docker BioContainer and uploaded to the Docker (quay.io) registry, which makes it available to container orchestrators (e.g. Mesos and Kubernetes) commonly used within the two projects.

On the infrastructure side, links already exist between EOSC-Life and EOSC-Pillar. For example, several France Grilles sites also provide services to the French Bioinformatics Institute, the French node of Elixir. They are working together on tools for authentication and

VRE deployment on Cloud environment, and are sharing best practices in regard to storage management with iRODS.

III. Further collaboration

The [Pulsar Network](#) allows the distribution of computing jobs from the European Galaxy server across several European data centers. For instance, INFN and GARR, both involved in EOSC-Pillar, are hosting two Pulsar nodes. The network has already proven to be scalable and reliable, i.e. during the COVID-19 pandemic to run specific bioinformatics workflows (<https://covid19.galaxyproject.org>).

Nevertheless, data needs to be moved from Galaxy to the computing nodes, which impacts negatively on the job's execution time in case of large amounts of data. Both projects are interested in developing a distributed caching layer, using already available solutions, like [iRODS](#), to reduce the job staging time for those nodes that already access the required data.

With this scientific agreement, we aim at strengthening the collaboration between the two European projects. This text will be presented to both the EOSC-Pillar and the EOSC-Life boards.

For the EOSC-Pillar project

Fulvio Galeazzi (EOSC-Pillar Project Manager, GARR)

fulvio.galeazzi@garr.it



13th July 2022

For the EOSC-Life project
Niklas Blomberg, ELIXIR Director, Project Coordinator
Niklas.Blomberg@elixir-europe.org



17 August 2022