# EOSC-Life D3.2 Report on the work of the initial demonstrators

Frauke Leitner, Maria Carazo Jose, Johanna Bischof, Natalie Haley, Pauline Audergon, Carlos-Oscar Sorzano, Laura Del-Cano, Pablo Conesa, Cymon-J. Cox, Gianluca De-Moro, et al.



EOSC-Life

# EOSC-Life: Building a digital space for the life sciences

## D3.2 — Report on the work of the initial demonstrators

WP3 – Demonstrators and Open Calls
Lead Beneficiary: Instruct-ERIC, CSIC
WP leader: Natalie Haley, Pauline Audergon, Frauke Leitner, Jose Maria Carazo and Johanna Bischof
Contributing partner(s): EMBL-HD, CSIC, Instruct-ERIC, CCMAR, EMBRC ERIC, ALUFR, UNIVDUN, UNITO, CNR, FZJ, ERINHA, UNIMAN, NIB, UNI Bielefeld, Fraunhofer, Diamond, IMG, EATRIS, BSC, CRG, INRA, CNRS

Authors of this deliverable: **Frauke Leitner**, **Jose Maria Carazo**, **Johanna Bischof**, **Natalie Haley**, **Pauline Audergon**, **Carlos Oscar Sorzano**, **Laura del Cano**, **Pablo Conesa**, **Cymon J. Cox**, **Gianluca De Moro**, **Corre Erwan**, **Katrina Exter**, **Gildas Le Corguillé**, **Romain Dallet**, **Lorraine Gueguen**, **Jean-Karim Heriche**, **Beatriz Serrano**, **Yi Sun**, **Jean-Marie Burel**, **Sara Zullino**, **Dario Livio Longo**, **Cyril Pommier**, **Asis Hallab**, **Constantin Eiteneuer**, **Romain David**, **Bjorn Usadel**, **Stuart Owen**, **Kristina Gruden**, **Roland Pieruschka**, **Alexander Sczyrba**, **Alfred Puhler**, **Martin Beracochea**, **Robert Finn**, **Philip Gribbon**, **Andrea Zaliani**, **Rachael Skyner**, **Frank von Delft**, **Ctibor Skuta**, **Andrew Leach**, **Jing Tang**, **Salvador Capella**, **José M. Fernández**, **Jordi Rambla**, **Sergi Beltran**

Contractual delivery date: **30 April 2021**
Actual delivery date: **27 May 2021**
H2020-INFRAEOSC-2018-2

Grant agreement no. 824087
Horizon 2020
Type of action: RIA

# Table of Contents



This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

# Executive Summary

This deliverable 3.2 is a report on the demonstrator projects, the eight scientific and technical pilot projects that were selected to provide concrete scientific use-cases and guide and structure the work done in EOSC-Life to build an open digital and collaborative space for biological and medical research.

We report in this deliverable

1) the process of integration of the demonstrators within EOSC-Life,
2) the achievement of the demonstrators who developed and made available to the scientific community several valuable resources (databases, workflows, web platform...),
3) the actions undertaken within EOSC-Life to disseminate the demonstrator achievement and finally
4) the results of the demonstrator survey to learn from the demonstrator experience and improve the integration of the new pilot project within EOSC-Life.

# Project Objectives

With this deliverable, the project has reached the following objectives:

a. Support an initial set of selected demonstrators to start their work with the start of the EOSC project.

# Detailed Report on the Deliverable

## 1. Description of Work

### 1.1. Selection of the demonstrator projects and preparation for demonstrator projects launch

#### 1.1.1. Selection of 8 demonstrator projects

The objective of EOSC-Life is the development of an open collaborative digital space for biological and medical sciences in Europe. EOSC-Life brings together 13 Life Sciences Research Infrastructures (LS RIs). Within EOSC-Life work package 3 (WP3), eight scientific and technological pilot projects, the “demonstrators"[1], representing a broad scope of different life science domains, were selected prior to the start of the project. The demonstrators provide concrete scientific use cases to both inform and test the outputs of other work packages in EOSC-Life and to provide

[1] http://www.eosc-life.eu/services/demonstrators/

This project has received funding from the *European Union’s Horizon 2020 research and innovation programme* under grant agreement No 824087.

FAIR data, tools and workflows to the wider life science community. To achieve this, demonstrators received funding for personnel (15 PM) along with support from the data and tools experts from the EOSC-Life consortium (WP1 and WP2 respectively) and access to EOSC-Life resources and training.

The demonstrator projects were selected by a light-weight process established during the grant writing phase with the overarching aim of identifying pilot projects that would allow the EOSC-Life consortium to start work immediately on concrete pre-identified challenges. The demonstrators were selected to target defined user communities in the life sciences and to benefit the scientific community as a whole through the deployment of a wide range of data types and services coming from the different LS RIs. Selection criteria for the identification of suitable demonstrator projects included expected impact, diversity of communities addressed, availability of data set or tools to be deployed as well as representation of different RIs.

### 1.1.2. Demonstrator kick-off meeting

Although the demonstrator projects already had scientific and technical capabilities within their own team, EOSC-Life WP3 provided added value to its demonstrator projects through additional support and direct contact with the technical experts from other work packages within EOSC-Life, which was provided throughout the lifetime of their projects. WP3 performed the match-making between demonstrators and technical experts from WP1 and WP2. WP3 also ensured connections between the demonstrators and the wider EOSC-Life project and integration with the life science research infrastructure community. WP3 invited the demonstrator participants, of which some were otherwise not affiliated with the consortium, to attend the EOSC-Life kick-off meeting. At the kick-off meeting, WP3 organized dedicated sessions for the demonstrators to provide them with a forum to meet EOSC-Life partners and exchange with other demonstrator participants, as well as to highlight their relevance for the overall project and to emphasize that initial work in EOSC-Life should be driven by actual scientific use cases and their real-life requirements.

To familiarize everyone with the concept and idea behind each demonstrator, demonstrator partners were invited to present their projects. They also had the opportunity to meet partners in the work package for data experts (WP1) and tools and workflow experts (WP2), initiating the process of identifying relevant experts that could support each demonstrator to achieve its objectives. During the kick-off meeting, the first valuable connections were established that led to the formation of dedicated project teams working on the demonstrators.

### 1.1.3. Individual demonstrator meetings

To continue the match-making process between the demonstrators and WP1 and WP2 experts following the EOSC-Life kick-off meeting, WP3 collected more detailed technical information regarding each demonstrator project. To streamline this information collection, WP3 invited WP1 and WP2 to develop short questionnaires with a focus on technical project details that demonstrator partners then answered within their teams. WP3 then organized individual video conference calls with each demonstrator project team and the WP1 and WP2 leads in summer 2019 during which the content of their questionnaire was discussed in detail. These in-depth discussions helped to identify the areas where support was most needed for the project and further helped to identify suitable experts to join the demonstrator teams.

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

It has to be noted, however, that the assignment of technical experts was not as fast and straightforward as initially anticipated, as there was not sufficient pre-existing capacity within the consortium and hiring processes across institutes were much slower than expected. This situation also highlighted the urgent need for training of staff members, as the necessary skill set to contribute to the demonstrator projects was found to be limited in some RIs.

Following this set of demonstrator kick-off meetings, WP1 and WP2 performed a technical analysis of the projects. They assembled information on data types, data formats, cloud-readiness, listed software tools and registries in use and rated the overall maturity level of the projects. The analysis helped to detect potential issues or risks such as the use of commercial software or missing metadata, which could thereby be addressed early on in the project work. The outcome of this analysis was presented in a large 'Orientation TC' (04 October 2019) that was open to all interested parties across the consortium to learn more about the background of the demonstrator projects. Most importantly, it allowed the WP1 and WP2 leads to present their findings and to discuss with the demonstrator partners how to arrange the necessary support for them.

WP3 leads supported the progress of the demonstrator projects during their duration by stimulating and organising regular meetings among team members. Demonstrator team members were also included in the monthly WP3 meetings to keep them engaged and updated on the EOSC-Life project and to address any issues as they developed. To facilitate progress and offer project management support, WP3 leads stayed in close contact with the demonstrator projects, regularly collected progress updates from the demonstrators and made that information available in reporting (first periodic report, Mid Term review).

## 1.2. Individual projects description and achievements

### 1.2.1. D1 - European Open Science Cloud resources for Chemical Biology and Structure-Based Drug Discovery workflows

A huge amount of data is produced daily from drug sensitivity screens or protein/ligands 3D structural interaction screens. However, from this data produced, only a very small subset is published and made publicly available. The majority of data produced is lost due to lack of annotation/standardisation that would make the data FAIR and re-usable for all. Furthermore, structural data and bioactivity data are so far not interconnected and cannot be easily integrated.

In this demonstrator project, scientists from different research infrastructures (Instruct-ERIC, EU-OPENSCREEN, EATRIS, ELIXIR, Euro-BioImaging ERIC) collaborated to develop a set of tools to increase the FAIRness of chemical biology data and structural data and to enable the integration of structural biology data and chemical biology data to increase access and re-use of resources.

- The fragalysis platform[2] was developed, enabling rapid access to data from fragment screens in a collaborative environment. Fragalysis is a ligand-centric platform that provides information regarding ligands and their target. The development of Fragalysis within EOSC-Life also allowed the platform to have impact in COVID-19 work for instance by allowing almost real-time release of initial fragment screens via Fragalysis to the worldwide

[2] Fragalysis Platform, https://fragalysis.diamond.ac.uk/, Accessed: 14.05.2021

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

community. Virtual screening workflows for fragment-based drug discovery were developed using Fragalysis alongside Galaxy. This was for instance used to rapidly screen follow-up compounds for Mpro hits, Mpro is the main SARS-CoV-2 protease[3].

- EU-OPENSCREEN is working on the development of the European Chemical Biology Database (ECBD)[4] that will be the central repository for Chemical Biology data generated within the EU-OPENSCREEN network.
- The aim of the repository is to make these data as FAIR as possible by making them accessible to a wide community, establish standard formats and identifiers for molecules and targets, establish standards for ontologies/vocabularies for the description of the results acquisition process, and finally, make the description process user friendly for the data uploader. ECBD will be released in 2021 at a dedicated website[5].
- Drug sensitivity screens are increasingly used for pre-clinical drug discovery and clinical trial optimisation. A web portal called MICHA[6] was created and is already in use to facilitate the annotation of critical drug sensitivity assay components and thus increase the FAIRness of drug sensitivity screening data. MICHA increases the visibility of sensitivity assays by providing information regarding compounds, targets, samples reagents and protocol used in an experiment.
- The demonstrator 1 team contributed to the COVID-19 research by making SARS-CoV-2 data available in ChEMBL[7]. These data were then used to compile a list of drugs that might be of interest in the current pandemic. To this purpose, datasets from different sources were compiled in ChEMBL for deeper analysis.
- Image datasets from primary screening campaigns to identify repurposed compounds against SARS-CoV2 were deposited in the IDR[8] and these have been linked to data analysis workflows[9] in further collaborative work with WP1 and WP2.

This demonstrator benefited from the direct contribution of EOSC-Life WP1 and WP2 experts in the development of their project which was highly beneficial for their work according to their report.

This demonstrator project was presented internally to the EOSC-Life community at the AGM in 2020 by Andrea Zaliani and Andrea Giachetti and at the final Demonstrator Webinar in January 2021 in which several demonstrator 1 partners presented a series of tools developed to make chemical biology data and structural data more accessible and re-usable by the scientific community. Demonstrator 1 work was also presented at the OECD/Science Europe workshop "Research Infrastructures mobilisation in response to COVID-19: lessons learned" satellite event to the International Conference on Research Infrastructures (ICRI 2021)[10].

Apart from the presentations, four publications were submitted by demonstrator 1 partners, with two accepted and in print. These publications describe the MICHA platform, structure-based screening efforts and work related to cloudification of COVID-19 related bioactivity data that was

[3] https://covid19.galaxyproject.org/cheminformatics
[4] European Chemical Biology Database, https://www.eu-openscreen-data.eu, Accessed: 14.05.2021
[5] https://ecbd.eu/
[6] MICHA, https://micha.fimm.fi/about/, Accessed: 14.05.2021
[7] ChEMBL, https://www.ebi.ac.uk/chembl, Accessed: 14.05.2021
[8] IDR, https://idr.openmicroscopy.org, Accessed: 14.05.2021
[9] https://github.com/IDR/idr0094-ellinger-sarscov2, Accessed: 14.05.2021
[10] https://www.oecd.org/sti/inno/Research-Infrastructures-mobilisation.htm

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

performed by demonstrator 1:

- Tanoli et al[11]
- Ellinger et al (2021)[12]
- Kuzikov et al (2021)[13]
- Achdout et al[14]

### 1.2.2. D2 - Increasing the FAIRness of data and image processing workflows in Cryo Electron Microscopy

Cryo-electron microscopy (Cryo-EM) has evolved extremely rapidly in recent years with great advances in the quality of the instrumentation and the methods used for data analysis and has become the technique of choice to determine the structure of macromolecules. The aim of this demonstrator project is to increase the application of FAIR principles in Cryo-EM microscopy from the moment of data acquisition. For doing so, this demonstrator proposes a new method to handle Cryo-EM data that includes deposition of data and workflows from the facility to a centralised public repository (such as EMPIAR[15]) with the possibility of updating data and workflows throughout the data analysis process before submitting the final results to the EMDB[16] database.

The transfer of the data and associated workflow to EMPIAR is a streaming transfer. Together with the transfer of the resources from the facility to the public repository, there is a transfer of ownership of the data from the facility to the lab user that is handled by the workflow engine. To achieve ownership transfer, the facility requests the ownership change at EMPIAR, EMPIAR sends an email to the user and once that user accepts the ownership, the Cryo-EM facility loses access, and the entry belongs to the user who can then make updates.

As part of this project, a viewer was integrated in EMPIAR that allows users to view the Scipion workflow used for the analysis of the data. A data viewer was also created and should be functional very soon that allows users to view how the data looks at different steps of the analysis and provides some evaluation regarding the quality of the data.

This WP3 demonstrator was granted subsequent funding from the EOSC-Life WP1 open call to continue work on this project and thus increase the FAIRness and the functionality of the deposition system. The main goal of this new follow-up project is to get a CWL output, to describe the Cryo-EM data and workflow through an ontology, to deposit the workflows in the workflow hub and to make the submission automatic.

This demonstrator benefited from the direct contribution of a EOSC-Life WP1 expert in the development of their project which was highly beneficial for their work, according to them.

This demonstrator project was presented internally to the EOSC-Life community at the AGM in 2020 and at the final Demonstrator Webinar in January 2021 as a concrete example of the projects and collaborations that EOSC-Life is aiming to support. A poster presenting this

[11] https://www.biorxiv.org/content/biorxiv/early/2020/12/04/2020.12.03.409409.full.pdf
[12] https://doi.org/10.1038/s41597-021-00848-4
[13] https://doi.org/10.1021/acsptsci.0c00216
[14] https://www.biorxiv.org/content/10.1101/2020.10.29.339317v1.full
[15] EMPIAR, https://www.ebi.ac.uk/pdbe/emdb/empiar/, Accessed: 14.05.2021
[16] Electron Microscopy Data Bank (EMDB), https://www.ebi.ac.uk/pdbe/emdb/, Accessed: 14.05.2021

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

demonstrator and the Digital Life Sciences Open Call was also presented at the ELIXIR 3D-Bioinformatics 2020 Annual Workshop to help advertise EOSC-Life to the broader community of scientists.

Apart from this poster presentation, 2 publications have been submitted by this demonstrator supported by the ESOC-Life grant[17,18].

### 1.2.3. D3 - Rapid, scalable and reproducible deployment of (meta-)genomics assembly and analysis pipelines tailored to the biome of interest.

This demonstrator project aims to democratise the analysis of microbiomes by providing metagenomics researchers with a toolkit to produce and reuse a range of analysis pipelines that can be efficiently deployed on cloud computing infrastructures. Metagenomics studies routinely perform deep sequencing of the entire microbial DNA recovered from any biome by using various sequencing technologies (short- and long-reads). This can result in the need to scale data analysis over potentially TBs of data. A lack of standardization hinders the interoperability of tools within workflows that would facilitate data analysis. Developing the best analysis workflow that is also scalable is currently the biggest challenge for life scientists using metagenomics.

To tackle this issue, this project worked to standardize and facilitate the analyses of complex biomes by using standardized interfaces/containers and integrate containerized workflow components (depending on the needs) within workflow descriptions (e.g. CWL[19]). Specifically, the team combined EMGB (Elastic MetaGenome Browser)[20] pipelines with CAMI's benchmarking datasets and containers[21] to produce biome and sequencing technology-specific workflows, based on the best combination of tools. This process of combining pre-existing workflow components within a final workflow allows the rapid development of new pipelines and extension or modification of existing ones, depending on the needs and the microbiome complexity.

The CWL-based EMGB pipeline developed at Bielefeld University has been successfully deployed on two different de.NBI Cloud sites (Bielefeld and Gießen). To automate the deployment process, BiBiGrid[22] was used in combination with Ansible scripts. BiBiGrid is a tool for easy cluster setup within a cloud environment, based on a general cloud provider API. Currently, the implementation can configure deployments on OpenStack, Google Compute Engine, Amazon AWS, and Microsoft Azure. BiBiGrid offers an easy configuration and maintenance of a started cluster via command-line and uses Ansible to configure standard cloud images. Depending on the configuration, BiBiGrid can set up an HPC cluster for grid computing (Slurm Workload Manager[23]), including a shared file system. During resource instantiation BiBiGrid configures the network, local and network volumes, (network) file systems and also the software via Ansible for an immediate usage of the started cluster. The EMGB metagenome workflow was deployed this way, which again utilizes Docker images for the specific analysis tools part of the workflow.

[17] https://www.biorxiv.org/content/10.1101/2020.05.12.069831v1.full
[18] https://www.biorxiv.org/content/10.1101/2020.05.22.110445v1.full
[19] "Common Workflow Language." https://www.commonwl.org/. Accessed 12 Oct. 2020.
[20] "Bioinformatics for NGS-based metagenomics and the ...." 10 Nov. 2017, https://www.sciencedirect.com/science/article/pii/S0168165617315985. Accessed 12 Oct. 2020.
[21] "Critical Assessment of Metagenome Interpretation-a ... - PubMed." 2 Oct. 2017, https://pubmed.ncbi.nlm.nih.gov/28967888/. Accessed 12 Oct. 2020.
[22] BiBiGrid, https://github.com/BiBiServ/bibigrid, Accessed: 14.05.2021
[23] Slurm Workload Manager, https://slurm.schedmd.com/overview.html, Accessed: 14.05.2021

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

In parallel, EMBL-EBI has been working on the deployment of the MGnify[24] pipelines (described in CWL)[25] on Embassy Cloud (OpenStack), Oracle Cloud, and Google Cloud platforms. Two different classes of pipelines have been deployed, which have very different computational requirements. The assembly pipeline has a high memory requirement (200MB-8TB), but requires relatively few cores (<16), whereas the assembly analysis pipeline has the opposite characteristic of a relatively small footprint (~32MB), but scaled across many cores, depending on the analysis tool (~100). EMBL-EBI has primarily focused on issues surrounding the assembly pipeline as this represents a major bottleneck. Deployment on the Oracle Cloud revealed incompatibilities between the HPC job scheduler (Slurm) and the underlying operating system kernel, thus highlighting one of the disadvantages of cloud computing. Nevertheless, EMBL-EBI has successfully deployed the assembly pipeline on both the Embassy Cloud and Google Cloud platforms and have processed >100 datasets, as well as combining samples (co-assembly). In collaboration with WP7, a key area of ongoing development focuses on how to utilize a heterogenous framework within the MGnify production system, which involves issues with data transfer and communication with the underlying tracking infrastructure (database). This new decentralized architecture will be responsible for the orchestration of jobs across cloud tenancies using message queues, allowing an arbitrary number of workers, and is not confined to a cloud provider (a project to develop this further has been proposed as a Google Summer of Code project). For the analysis pipeline, all the workflow tools have been containerized[26] and successfully deployed across all the aforementioned cloud platforms. The greater complexity of this pipeline also required a significant investment in improving the compatibility of Toil[27,28] (the execution engine) with the job schedulers (Slurm and IBM LSF) and the CWL specification.

Within this demonstrator project, the EMBL-EBI team has been collaborating with EOSC-Life WP7 to develop and deploy a new version of the MGnify analysis pipeline that incorporates the VIRify viral analysis workflow[29]. The ability to smoothly combine different workflows into more complex workflows demonstrates the benefit of standardising workflow descriptions, not just for deployment in remote settings but also for streamlined management within a project. The VIRify viral pipeline descriptions developed in CWL and Nextflow have already been deposited as examples in the WorkflowHub[30], and the manuscript describing VIRify is in preparation. The knowledge gained during this demonstrator project has enabled the EMBL-EBI team to generate additional workflows (beyond the scope of this work), which have been cited in publications and improve the reproducibility of science[31].

Alexander Sczyrba presented this demonstrator project, specifically the advances in the deployment of the EMGB metagenomics analysis pipeline internally to the EOSC-Life community at the 2020 AGM. At the final Demonstrator Webinar in January 2021, Martin Beracochea presented the outcomes of the work done by the EMBL-EBI team in the development of the new version of MGnify and its deployment into the cloud.

[24] MGnify, https://www.ebi.ac.uk/metagenomics/, Accessed: 14.05.2021
[25] https://github.com/EBI-Metagenomics/pipeline-v5/tree/master/workflows
[26] https://hub.docker.com/search?q=microbiomeinformatics&type=image
[27] https://github.com/DataBiosphere/toil/pull/3445
[28] https://github.com/DataBiosphere/toil/pull/3229
[29] https://github.com/EBI-Metagenomics/emg-viral-pipeline
[30] https://workflowhub.eu/search?utf8=%E2%9C%93&q=Virify
[31] Almeida, A., Nayfach, S., Boland, M. *et al.* A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat Biotechnol* 39, 105–114 (2021). https://doi.org/10.1038/s41587-020-0603-3

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

### 1.2.4. D4 - Marine Eukaryote Genomics Portal – access to tools and data-flows for marine genome annotation

This EOSC-Life demonstrator project developed an online marine genomic resource to aid community-driven annotation of marine eukaryotes and help to provide a focus for post-assembly genomic workflows and data access of specific (closely related) groups of marine organisms. The more closely two taxa are evolutionarily related, the more they are expected to share genomic structure and synteny, and it is therefore of benefit to compare and contrast genome annotations between closely related organisms. This is especially important for communities that work on specific genomes and provide manual annotations via community platforms (e.g. ORCAE[32]). This demonstrator project addresses the lack of tools to compare and transfer annotations and features between the sequenced genomes of closely related species.

A software tool for comparison of genome annotations (GFFalign[33]) has been designed, implemented in Python, and integrated into the Galaxy platform. In addition a Snakemake workflow has been designed and implemented for cloud deployment and the tool source code and Docker module will be made available from a Github repository[34]. The tool is used in a soon-to-be launched public-facing portal for the community annotation of species belonging to the pelagic herring fish family, Clupideae. The genomes of eight Clupideae will be available for comparison initially, but the tool could also be used to compare any two genome annotation libraries. It is also intended that the tool will be able to automate the updating of annotations to the community annotation platform, ORCAE, through the platform API. All FAIR data and privacy issues surrounding data usage have been addressed. The website was developed in Django v3 using the database PostgreSQL v12. Each task of the Snakemake workflow is submitted to a batch system, SLURM. From the website each submission task is managed by a queue created using Celery. The Life Science AAI, the development of which was driven by WP5 of EOSC-Life, will be integrated in the next release, and will be the only login method to access the website.

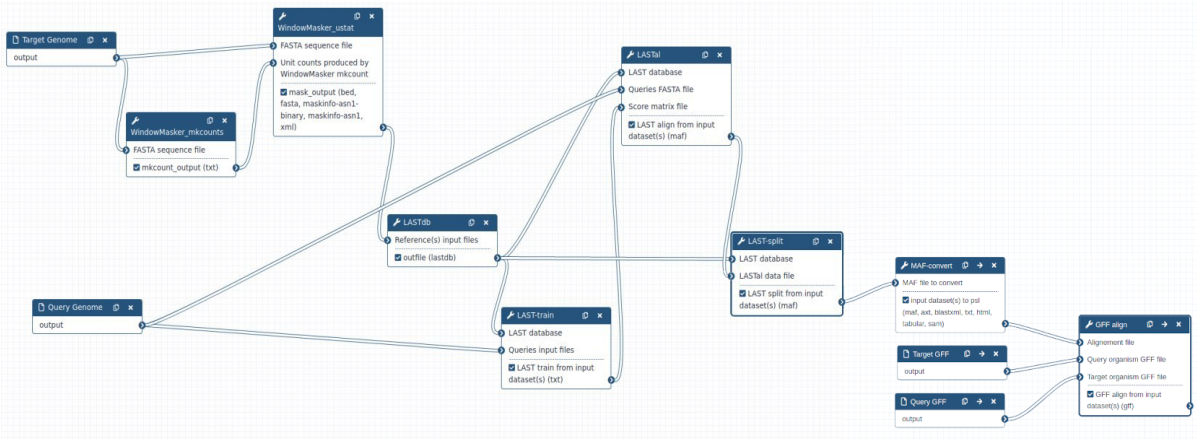


*Figure 1: Illustration of the workflow developed in Demonstrator 4.*

[32] ORCAE, https://bioinformatics.psb.ugent.be/orcae/, Accessed: 14.05.2021
[33] GFFalign, https://github.com/gdemoro/GFFalign, Accessed: 14.05.2021
[34] Marine Eukaryote Genomics Portal Github, https://github.com/eosc-life/D4_marine_eukaryote_genomics_portal, Accessed: 14.05.2021

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

This demonstrator benefited from the direct contribution of EOSC-Life WP1 and WP2 experts in the development of their project as well as from consultation with WP7 experts. The involvement of EOSC-Life experts, especially experts from WP2, was highly beneficial according to them by ensuring that the tool and data produced conform to FAIR principles.

The project team is considering making a publication of application notes in a specialist journal.

### 1.2.5. D5 - Development of novel and configurable workflow for processing preclinical images and extracting meaningful data

Preclinical research generates a huge variety of biomedical imagery with an increasing need of standard platforms for managing image data. This demonstrator project aimed at providing the scientific community with standardized tools for archiving, sharing and reusing preclinical images. Image processing tools built on top of in-house Matlab and Python scripts were developed to provide automated image analysis capabilities and the storage of the resulting image-derived data. The architecture is based on XNAT[35], a widely used open-source image informatics platform for archiving, accessing, and processing medical images. Despite its success, tools for importing multimodal, preclinical images as well as pipelines for processing large image datasets were not previously available in XNAT. To overcome these limitations, we have developed several tools in XNAT devoted to Preclinical Imaging Centers (XNAT-PIC) consisting of:

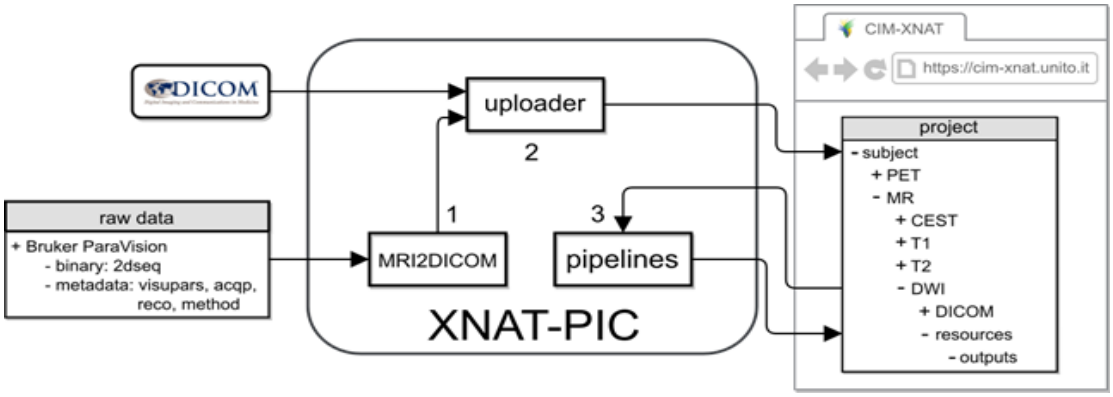


*Figure 2: XNAT-PIC pipeline and included capabilities.*

- XNAT-PIC Uploader to import large, multimodal imaging studies in DICOM standard[36] (international standard format for medical imaging) to XNAT
- XNAT-PIC MRI2DICOM, a ParaVision® (Bruker-Biospin Inc. Billerica, MA) to DICOM standard converter in Python
- XNAT-PIC pipelines for image processing at large scale based on in-house Python and MATLAB scripts.

While the default XNAT provides a system for securely hosting and processing clinical imaging studies, our demonstrator project expands its basic functionalities to support preclinical imaging facilities. The tools developed as part of this demonstrator project streamline the implementation

[35] "XNAT - Home." https://www.xnat.org/. Accessed 12 Oct. 2020.
[36] "DICOM Standard" https://www.dicomstandard.org/ Accessed 12 Oct. 2020.

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

of image dataset sharing among preclinical imaging centres, thereby facilitating the exchange and reproducibility of image-processing tools.

This demonstrator benefited from the direct contribution of an EOSC-Life WP1 expert in the development of their project and consultation with WP2 experts.

This demonstrator project was presented internally to the EOSC-Life community at the final Demonstrator Webinar in January 2021 as a concrete example of the projects and collaborations that EOSC-Life is aiming to support. This work was presented to the biological and biomedical imaging community at the Euro-BioImaging virtual pub[37]. This demonstrator work has also been presented internationally at the last European Society for Magnetic Resonance in Medicine and Biology (ESMRMB) Conference during the Software Exhibits session: S. Zullino, A. Paglialonga, W. Dastrù, S. Aime, D. L. Longo. XNAT-PIC: Expanding XNAT to Preclinical Imaging Centers. ESMRMB 2020 online, European Society for Magnetic Resonance in Medicine and Biology, Sept 30 - Oct 2, 2020.

In addition to these presentations, this work is submitted to a peer-reviewed journal and is available on arxiv[38].

### 1.2.6. D6 - Make newly established workflows for mining large image repositories accessible, reusable and scalable in the EOSC

Genome-scale cell imaging studies generate a wealth of image data that contains information on many aspects of cellular biology beyond what was examined in the original reports of these studies. Therefore, re-use of these image datasets can substitute for performing costly new experiments. To make this easily achievable, this demonstrator built a re-usable infrastructure to run data analysis pipelines from image processing to mining of public databases in the EOSC.

In the process of this demonstrator project, the image analysis tool CellProfiler[39] was brought into Galaxy[40] and a Galaxy tool to access IDR data was created[41]. To demonstrate the application, nine data sets of genome-scale RNAi screens were analysed with a focus on selecting genes affecting nucleolus architecture, by using CellProfiler to segment and measure the nucleoli. The list of hit genes is currently being analysed for further investigation into the control of nucleolus architecture.

This demonstrator benefited from the direct contribution of an EOSC-Life WP1 expert for the creation of the link between IDR and Galaxy and the direct involvement of EOSC-Life WP2 experts in the development of the project. The involvement of both WPs was evaluated as having very high benefit to the project progression by the demonstrator.

This demonstrator project was presented internally to the EOSC-Life community at the AGM in 2020 and at the final Demonstrator Webinar in January 2021 as a concrete example of the

[37] Recording available at https://www.youtube.com/watch?v=QNiAGuFk53w&list=PLW-oxncaXRqVsVoEK5pH62_nSY2qkP0MW&index=6
[38] "Arvix." https://arxiv.org/abs/2103.02044
[39] "CellProfiler." https://cellprofiler.org/. Accessed 12 Oct. 2020.
[40] "CellProfiler now available in Galaxy - WP2 - EOSC-Life." 1 Jul. 2020, https://forum.eosc-life.eu/t/cellprofiler-now-available-in-galaxy/52. Accessed 12 Oct. 2020.
[41] "CORBEL staff visit program: Integration IDR ... - Galaxy Europe." 8 Feb. 2020, https://galaxyproject.eu/posts/2020/02/08/idr-galaxy-hackathon/. Accessed 12 Oct. 2020.

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

projects and collaborations that EOSC-Life is aiming to support.

The demonstrator team also developed related training material[42] and registered the workflow in WorkflowHub[43].

### 1.2.7. D7 - An integrative analysis pipeline of genomic and transcriptomic human data for disentangling the genetic origin of a rare-disease in the context of the European Open Science Cloud

This demonstrator project developed a pipeline that allows the interconnection of two key technological infrastructures for handling sensitive human data: the European Genome-phenome Archive (EGA)[44] and RD-Connect[45], making sure that genomic and phenotypic data, including metadata, can be used only by researchers who are allowed to do so. This project allows the integration of different omics data, Whole Genome Sequencing (WGS), and RNA-Seq and establishes a protocol to prioritize genetic variants associated with the observed phenotype. These analysis pipelines were established taking the Congenital Myasthenic Syndrome (CMS) as a use-case but are designed to be used routinely for other projects. The interconnection of high-quality analysis pipelines and metadata allows application of personalized analyses leading to the extraction of clinically relevant insights.

In order to achieve their goals, the demonstrator tested and chose the most adapted technologies to minimize possible data leakages from undisclosable sources when handling sensitive data. They started the integration of the on-the-fly decryption system from EGA into the workflow runner and implemented their workflow (an evolution of Wetlab2Variations[46]) in three different languages (nextflow, cwl and galaxy) and adapted it to the needs of their users.

Building on this demonstrator project, Salvador Capella and his team are working on follow-up activities with TransBioNet: The Spanish Translational Bioinformatics Network of units and groups at healthcare facilities. The aim of their work is to provide a reference implementation, to discuss best methods to do an analysis and make all the workflows that have been benchmarked available to the community via the EOSC-Life WorkflowHub.

Demonstrator 7 benefited from the direct contribution of EOSC-Life WP2 experts in the development of their project which was highly beneficial according to them as the EOSC-Life experts provided guidance on how to best generate workflows towards guaranteeing provenance and reproducibility. The WorkflowHub represented a great stable infrastructure to deposit their workflow.

One output of this demonstrator will be a manuscript on how to prepare a portable and fully reproducible bioinformatics workflow that is currently being written. This demonstrator project was presented internally to the EOSC-Life community at the final Demonstrator Webinar in

[42] "Nucleoli segmentation and feature extraction using CellProfiler." 30 Jun. 2020, https://training.galaxyproject.org/training-material/topics/imaging/tutorials/tutorial-CP/tutorial.html. Accessed 12 Oct. 2020.
[43] "The WorkflowHub." 24 Jun. 2020, https://workflowhub.eu/tags/88. Accessed 12 Oct. 2020.
[44] "home | European Genome-phenome Archive - EMBL-EBI." https://www.ebi.ac.uk/ega/home. Accessed 12 Oct. 2020.
[45] "RD-Connect." https://rd-connect.eu/. Accessed 12 Oct. 2020.
[46] "inab/Wetlab2Variations: Wetlab2Variations Workflow ... - GitHub." https://github.com/inab/Wetlab2Variations. Accessed 12 Oct. 2020.

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

January 2021 as a concrete example of the projects and collaborations that EOSC-Life is aiming to support.

### 1.2.8. D8 - Plant A+: Taking Plant Omics Data through Annotation, Acquisition and Analysis to Application

The aim of the Plant A+ demonstrator project was to bring together different datasets from different research infrastructures (ELIXIR, EMPHASIS, ISBE), allowing the association of genotypes and phenotypes and making data and pipelines FAIR. The project started with a detailed dataset on a tomato population comprising drought, photosynthesis parameters, enzyme activities, biochemical traits as well as a detailed characterization of genomic data as well as expression profiles, and additional data from other Solanaceous (potato and tobacco).

Plant A+ integrated plant omics data and provides tools to enable efficient phenotyping data sharing and to perform analysis and visualization of this data, including new user provided data. A critical part of this project is the work on gene expression visualization, allowing the comparison of gene expression in a single plot, as well as the comparison of different tissues or conditions in side-by-side plots[47]. The other crucial component is the ability to share phenotype data in FAIRDOM[48] and to ensure MIAPPE[49] compliance and indexing in FAIDARE[50].

Thanks to this demonstrator project, expression data can now be processed and translated into a memory database and a React.js component is currently being developed. The MIAPPE plant phenotyping data standard has been added to FAIRDOM (Biological material and traits ID) - see github seek4science issues 99[51] and 98[52] for details. The indexing of selected datasets in FAIDARE, the ELIXIR Plant data lookup service, has also been achieved.

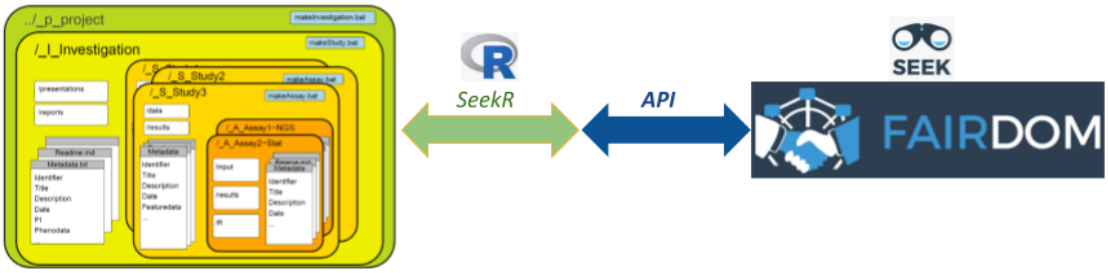


*Figure 3: Connecting the Plant A+ tools with FAIRDOM.*

This demonstrator benefited from consultation with both experts from WP1 and WP2, which were both rated as having high benefit to the project progression by the demonstrator.

This demonstrator project was presented internally to the EOSC-Life community at the final Demonstrator Webinar in January 2021 as a concrete example of the projects and collaborations that EOSC-Life is aiming to support.

[47] Gene Expression Plots: usadellab.github.io/GeneExpressionPlots
[48] "FAIRDOM." https://fair-dom.org/. Accessed 12 Oct. 2020.
[49] "About < MIAPPE < EMBL-EBI." https://www.miappe.org/. Accessed 12 Oct. 2020.
[50] "FAIDARE - URGI." https://urgi.versailles.inrae.fr/faidare/. Accessed 12 Oct. 2020.
[51] https://github.com/seek4science/seek/issues/99
[52] https://github.com/seek4science/seek/issues/98

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

Parts of the materials developed within this demonstrator are already publicly available via github and in future a publication may be submitted with the resulting works. The work started within the demonstrator has already resulted in useful and demanded tools and will be further continued. For example the developments in the demonstrator represent an integral part of an upcoming Horizon Europe proposal related to service provision for sustainable agriculture.

## 1.3. Dissemination/Outreach material

Valuable resources (databases, workflows, web portals, workflows...) have been made available through the demonstrator projects and are described in detail in the “projects description and achievements” section. Effort has been made within EOSC-Life, especially in WP3 and WP10, to disseminate the work of the demonstrators and make it more visible to the EOSC-Life community and to the scientific community as a whole.

Internally, the demonstrator projects and achievements were presented on several occasions to the consortium. Four of the demonstrators (D1, D2, D3 and D6) presented the advances on their projects at the AGM in 2020. Following completion of the demonstrator projects, in January 2021 all the demonstrators presented their work to the EOSC-Life community in a final webinar series called "Populating EOSC-Life: Success stories from the demonstrators" organised by WP3 in which the demonstrators talked about their achievements, the impact of their project and their experience with EOSC-Life and addressed questions from the community. The attendance at these webinars was high (>80 participants for each of the 2 webinar sessions), demonstrating the interest and the involvement of the EOSC-Life community in these pilot projects. The recordings of these presentations formed the basis of dissemination material on the demonstrator projects, which WP3 produced in close collaboration with WP10. This included detailed project descriptions for each demonstrator on the EOSC-Life website, their achievements and the resources made available to the community, as well as YouTube recordings of the presentations[53]. The demonstrator project results were promoted on social media through a dedicated campaign on LinkedIn and Twitter.

The outcome of the demonstrator projects represents a valuable resource to communicate about EOSC-Life and to illustrate with concrete use-cases what EOSC-Life is aiming to achieve in building an open, digital and collaborative space for biological and medical research. The work done by the demonstrators highlights how projects can be integrated and benefit from resources, training and expertise available in EOSC-Life.

The demonstrators' achievements were used to advertise the EOSC-Life Digital Life Sciences Open Call. Presenting these pilot projects illustrated what type of projects EOSC-Life is aiming to promote and showed how EOSC-Life can be beneficial to life science researchers by providing funding, training and expertise to the applicants. To that end, the demonstrator 2 project was for example presented alongside the WP3 Digital Life Science open call in a poster and a short presentation at the ELIXIR 3D-Bioinformatics 2020 Annual Workshop. The project achievements from demonstrators 2, 5 and 6 were also shown as example projects on the Digital Life Sciences Open Call website.

[53] https://www.eosc-life.eu/services/demonstrators/

This project has received funding from the *European Union’s Horizon 2020 research and innovation programme* under grant agreement No 824087.

The demonstrators themselves also benefit from EOSC-Life to disseminate their research. The project outcomes and publications are advertised via EOSC-Life thanks to the involvement of EOSC-Life WP10. Support for the demonstrators in putting together training material and placing such materials in the relevant repositories is provided by WP9 in an ongoing collaboration.

## 1.4. Demonstrator survey

At the end of the demonstrator projects, WP3 conducted an in-depth survey with the project participants to receive feedback from the demonstrators on their interactions with the different EOSC-Life WPs, the impact that EOSC-Life had on their project and potential improvements for future Open Calls.

Overall, the feedback from the demonstrators regarding their experience with EOSC-Life was highly positive. The majority of demonstrators (5/8) reported that they could not have carried out their project at all in the absence of EOSC-Life support, and two out of eight projects would have taken longer to complete without EOSC-Life support. Without EOSC-Life support, lack of funding and lack of technical expertise would have prevented the majority of projects from being completed. One demonstrator reported that EOSC-Life was crucial in allowing them to move from the prototype of an application to a well-defined set of goals. The received guidance helped them to achieve these goals in an efficient way.

Almost all demonstrators reported that the financial support provided by EOSC-Life was "about right" - at the same time four out of the eight demonstrators used additional external funding to support their project work. The general timeframe of the demonstrator projects was also positively evaluated, with almost all demonstrators reporting the timeframe was "about right" or just "slightly too short".

The majority of demonstrators consider that the interaction with EOSC-Life was highly beneficial thanks to the funding for the personnel provided, but was also beneficial to promote collaborations within and outside RIs and to obtain services and technical expertise from the consortium (Figure 4).

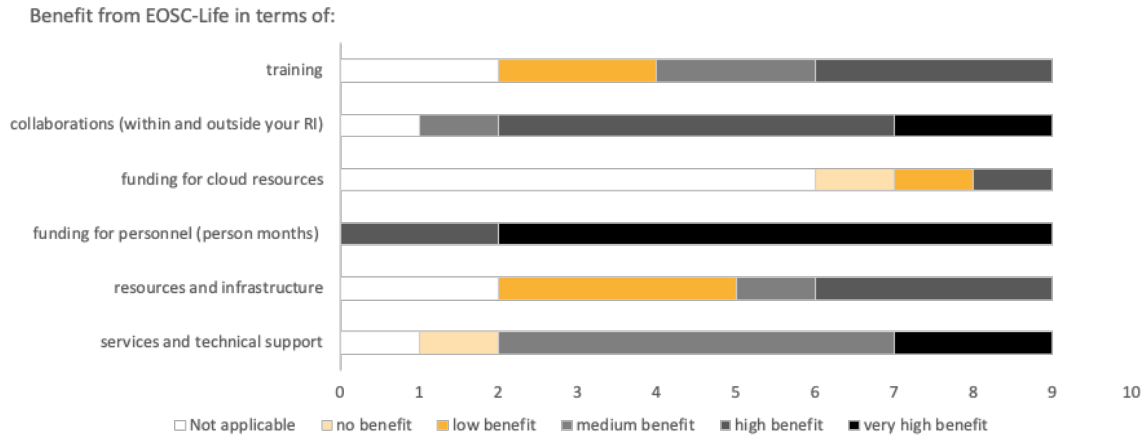


*Figure 4: Outcome of the Demonstrator survey in which participants were asked to evaluate the benefits of their participation in EOSC-Life.*

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

The demonstrators also valued the interactions with the WPs within EOSC-Life. Six out of the eight demonstrators had some level of support from WP1, with four benefiting from active participation of WP1 experts in the project teams. This active participation was valued as having medium or high benefit by all the involved demonstrators. Five demonstrator projects had support from WP2 including three with active involvement of the WP2 experts. The active participation from WP2 was evaluated as having very high or high benefit. Three projects also received support from WP7, in the form of advice on cloud resources.

The interaction with the WP3 project managers was also deemed as beneficial, with 8 respondents reporting medium, high and very high benefits from the support provided by WP3 project managers. One responded that "the interaction with the WPs was efficient and productive".

Involvement with EOSC-Life also had overall positive effects on the project teams, with 75% of respondents saying that their EOSC-Life demonstrator project had some impact on their ability to secure future funding or fellowships, with 50% even rating the involvement as having a large impact. This is also reflected in the fact that several demonstrators went on to apply for either the EOSC-Life WP1 Call or the WP3 Digital Life Sciences Open Call.

Six demonstrators would recommend that scientists take part in EOSC-Life via the Digital Life Science Open Call. Some of the listed reasonings were:

- ”EOSC-Life is developing many useful technologies that can benefit many scientists.”
- “The interaction with international colleagues is invaluable in building collaborations.”
- “EOSC-Life project offers the possibility to make expertise and resources of different European research infrastructures available to the scientific community, thus promoting the networking and cooperation among them, accelerating the awareness of the importance of Open Science.”
- “EOSC-Life represents a great opportunity to work collaboratively and ensure that any effort is aligned with the Life Sciences Community. Even if the project goes slower than an individual group, it is worth engaging to work towards the long-term sustainability of any effort.”

Consistently, the demonstrators reported that there is a real need to train and educate the scientific community in certain areas, such as cloud use and best practices, metadata and ontology, and the building of proper data management plans. Relatedly, it was also stressed that the Research Infrastructures need to develop a vision for EOSC and get more technical expertise. One demonstrator also commented that further integration of the demonstrators within EOSC-Life, in addition to the existing connections to WP1, 2, 3, and 7, would have been helpful.

The majority of demonstrators declared that they were not aware of other sources of funding for cloudification of resources outside of EOSC-Life. This response together with the high application numbers in the EOSC-Life Open Call clearly demonstrates that there is a real need in the scientific community for investment in cloud capacity, skills and training.

This project has received funding from the *European Union’s Horizon 2020 research and innovation programme* under grant agreement No 824087.

## 2. Next Steps

### 2.1. Dissemination of demonstrator project and resources to the broader scientific community

The material created for dissemination of the demonstrator projects will continue to be used for dissemination. WP3 and the demonstrator teams will continue to present results from the demonstrator projects at relevant conferences, workshops and meetings. In collaboration with WP9, WP3 is exploring the opportunity to create tutorials and training material around the demonstrator project outputs to enable and encourage uptake in the community. WP3 and WP10 will continue to promote the demonstrator project results and new publications coming from the demonstrators as they are published.

### 2.2. Use the demonstrator feedback to improve the integration of the new pilot projects from WP1 and WP3 open calls within EOSC-Life

From the project outset, the demonstrator projects were intended to inform the way EOSC-Life WP3 would select, support and integrate new projects arising from the open calls. The Digital Life Sciences Open Call closed to applications in December 2020 and project awards were made in April 2021 following an extensive review process (Deliverable 3.1). These projects are now preparing to start work and the experience from this work with the demonstrators, including the feedback from the demonstrators, provides a basis for how the open call projects will be integrated into and supported by the EOSC-Life project. Some notable improvements planned are better integration of projects teams with other work packages besides WP1 and 2 where applicable. WP4 in particular, who work on sensitive data handling will be closely involved in the work of new open call projects involving sensitive data. As many of the open call selected projects are newcomers to EOSC-Life and to the life science research infrastructures, WP3 will take care to ensure the relevant research infrastructure and work package connections are made in the working teams for the open call projects.

# Delivery and Schedule

The delivery is delayed: Yes

The delivery is delayed, as the original delivery date overlapped with the crucial evaluation and selection phase for the Digital Life Science Open Call.

# Adjustments made

None

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.