EOSC-Life

EOSC-LIFE: BUILDING A DIGITAL SPACE FOR THE LIFE SCIENCES

PROJECT BRIEF

Delivery date: **13 December 2021**
Authors: **Niklas Blomberg, Work Package Co-Leads and Consortium members**

Grant agreement no. 824087
Horizon 2020
H2020-INFRAEOSC-2018-2
Type of action: **RIA**

Introduction 3

1. Integration with the EOSC infrastructure 4

Recommendations on follow-up priorities linked to the further development of EOSC 5

2. FAIR principles implementation and repositories 6

Recommendations on follow-up priorities linked to the further development of EOSC 6

3. Technical, semantic, legal and organisational interoperability 7

Recommendations on follow-up priorities linked to the further development of EOSC 7

4. Stewardship of data 7

Recommendations on follow-up priorities linked to the further development of EOSC 8

5. Cross-cluster collaboration activities and achievements 8

Recommendations on follow-up priorities linked to the further development of EOSC 8

6. Appendix 1: EOSC adoption in science - the EOSC-Life open calls 9

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

PROJECT BRIEF

Introduction

EOSC-Life (eosc-life.eu) brings together the 13 Life Science Research Infrastructures (LS RIs) within ESFRI in a project that will create within EOSC an *open collaborative space for digital life science*. The project co-creates and integrates the EOSC federated core, while simultaneously conceptualizing, adapting and adopting the services and policies for Open Science that help researchers in the life sciences to manage, publish, analyse and reuse data.

The LS RIs provide access to advanced instruments and research facilities - helping researchers to describe biology from single molecules to ecosystems and long-term population cohorts. As distributed organisations, the LS RIs each bring national facilities and centres into a connected European entity with harmonised access, quality and data management. EOSC-Life builds on the outcomes from previous cluster projects: in CORBEL the LS RIs established a foundation of collaborative scientific services that support users throughout the execution of a scientific project: from planning and grant applications through to the long-term sustainable management and exploitation of research data. EOSC-Life takes the next step and further harmonises user access, unifies data management for biomedical research in Europe and will embed the combined infrastructure capabilities into the scientific workflow of advanced users.

Service development in EOSC-Life is user-driven. EOSC-Life is supporting 27 innovative science projects making life science data, tools and workflows available in the EOSC from a series of open and internal calls for user projects. This allows the wider life science community to engage within EOSC to make data FAIR, share it in the cloud and support the implementation of large-scale data analysis workflows in the cloud. During the pandemic EOSC-Life has refocused activities in response to the *ERAvsCORONA action plan* to support open science research with workflows, training, cloud access and additional funding for two new work packages to develop the European COVID-19 Data Platform and a repository for COVID-19 Clinical Trial Data.

The training work package concluded a gap analysis of existing and needed training for users of EOSC-Life resources and services. Training materials and events aimed at filling the gap are created both by EOSC-Life and by members of the community through the Training Open Call. A total of eight Training Open Call projects have been granted with over 70 organisations involved in the submitted proposals. In addition, EOSC-Life supported large-scale training for COVID-19 data analysis on the Galaxy platform in a world-wide event that had more than 1,000 attendees globally.

The following are products and initiatives developed through EOSC-Life addressing key areas of concern for the EOSC, as well as recommendations on follow-up priorities linked to its further development.

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

1. Integration with the EOSC infrastructure

Packaging tools and workflows for use in the cloud

EOSC-Life is developing expertise in the cloud deployment of software and workflows across all domains of the life sciences. This will help the LS RIs to develop their computational infrastructure and make it FAIR.

WorkflowHub: This registry of scientific workflows makes them findable for collaboration and reuse. The federated registry supports a common API to simplify access for tool developers. An open community of developers and users has been formed and is growing and WorkflowHub currently hosts 172 workflows from 9 different systems, of which 162 are publicly available. There have been 3867 downloads; 216 registered users from 69 organizations across 17 countries have organised themselves into 76 workflow teams so far. Users outside EOSC-Life partners (LSRI) include the Australian BioCommons, the EU Bioexcel HPC CoE, and the ESFRIs IBISBA and DISSCo. WorkflowHub implements international standards that enable it to couple with workflow systems for automated execution (e.g. CWL) and with EOSC-Life services for workflow testing and workflow preservation. The standards support the emerging paradigm of FAIR Principles for Computational Workflows, in which EOSC-Life has a leadership role.

COVID workflows in the WorkflowHub: Due to the COVID-19 pandemic, we shifted our focus towards providing tools and workflows to tackle analyses of COVID-related data in an open and reproducible way. The WorkflowHub was accelerated by 6 months and released early to provide a registry for COVID-19 workflows. In collaboration with the global Galaxy community, workflows have been made available on public Galaxy instances worldwide. 40 entries in WorkflowHub are COVID-19 related and a freely available, curated Galaxy SARS-CoV2 surveillance workflow collection is maintained. The most popular workflow in the Hub is for COVID variant calling with 5 versions, 2114 views, 57 downloads and launched daily on usegalaxy.eu.

LifeMonitor: Periodic testing is important to prevent workflows from collapsing over time, allowing problems to be exposed when they arise and providing a machine-actionable way to verify changes to the workflow structure. LifeMonitor aims to facilitate the maintenance of computational workflows by supporting the creation, application and monitoring of workflow tests and the adoption of workflow best practices.

The Workflow services above use a **metadata framework** developed by EOSC-Life based on Bioschemas (an implementation of schema.org metadata markup) and RO-Crate (an implementation of FAIR Digital Objects, as envisioned by the FDO Forum and the EOSC Interoperability Framework). Bioschemas metadata cataloguing and exchange has been piloted for openAIRE in EOSC Enhance, to be continued in EOSC Future. RO-Crates are used in a number of EOSC projects, including Reliance and CS3MESH4EOSC. EOSC-Life provides **access and support to cloud resources** for the life science community and offers deployable platforms for workflows. In addition, we establish capabilities and identify suitable cloud providers for secure hosting of sensitive data.

EOSC-related trainings: Trainings have been held on the following topics: Climate Science Workflows and Training on EOSC-Life Infrastructure (cross-cluster collaboration), Galaxy Admin Training on EOSC-Life infrastructure, and the expansion of the Training Infrastructure as a Service (TIaaS).

Providing and improving user access

We are creating an access and user management system to enable multi-RI applications and workflows that build on existing approaches and support access to sensitive data with their specific requirements.

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

LS Login: The Life Science Login enables researchers to use their home organisation credentials or community/ commercial identities (e.g. ORCID, LinkedIn) to sign in and access data and services they need across multiple platforms. It also allows service providers (both in academia and industry) to control and manage access rights of their users and create different access levels for research groups or international projects. The EOSC-Life project also actively contributes to EOSC AAI work (EOSC association, EOSC Future). By the end of the project, Life Science AAI will have entered production and has a post-project sustainability plan. Life Science AAI is integrated to EOSC AAI, enabling life science researchers to log in and access the EOSC services.

ARIA: ARIA is currently used by 5 research infrastructures, as well as 4 national facilities. ARIA was presented cross cluster to both ACTRIS (Atmospheric sciences) and CESSDA (social sciences) for use within their infrastructures, the latter with a view for the expansion of ARIA as part of EOSC Future. By the end of EOSC-Life, ARIA will be implemented with BBMRI Negotiator to help ARIA users better select services. We will expand the support for APIs in ARIA, allowing for data ingress and egress, and provide a platform for other RIs to integrate with.

Toolbox for sharing of sensitive data: A toolbox is currently under development, providing pooled information on recommendations, best practices, software tools, etc. to researchers who wish to share and/or use sensitive data in a cloud environment in general, and the EOSC in particular.

Improve data discoverability in the EOSC ecosystem

FAIRsharing and OpenAIRE connection: FAIRsharing is the catalogue of resources. It is improving its discoverability and connection in the EOSC ecosystem via the FAIRsharing-openAIRE collaboration to assist with the selection, documentation and visualisation of standards; contribute the data discovery, profiling the databases' access methodologies; and connect to EOSC metadata aggregators. EOSC-Life RIs register their resources (services, databases, etc.) in FAIRsharing to promote discoverability via FAIRsharing's EOSC-Life service and standards collection. To date >100 services are registered across 13 Life Sciences RIs and metadata standards for FAIRsharing are compatible with other EOSC metadata catalogs promoting cross-domain interoperability.

Recommendations on follow-up priorities linked to the further development of EOSC

1. Ensure that all EOSC actors have skills and competencies needed to interact with EOSC and increase uptake of EOSC resources and Open Science. This should be a coordinated European effort to ensure that less technical communities or individual countries aren't left behind. Develop long-term training and support models for the EOSC services developed in this project.
2. Many services developed here have become mission critical for several ESFRI RIs, and so the ongoing sustainability of these platforms must be addressed in terms of operating costs and development resources. This includes clarity on long-term business and funding models for cost-recovery (e.g. from large-scale use within HE projects). Align cost-recovery models for cloud-based storage and compute with trans-national facilities access.
3. Data discovery is fundamental to EOSC – deepen links such as FAIRsharing-openAIRE to facilitate light-weight, cross-domain discoverability.
4. Widen and deepen the EOSC user base through open data use cases and science-led projects, bringing together services across research infrastructures of direct relevance to the other HE missions, related European Partnerships (IMI, Rare Disease, OneHealth) and global, environmental, social and economic challenges. These projects will embed relevant EOSC services for the management of research digital outputs (e.g. data, models, software, preprints and workflows) and cover the full research life-cycle. Particular

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

emphasis should be on use cases and proof of concepts to establish best practices for the integration of heterogeneous datasets for advanced AI, aligned with global environmental, social and economic challenges.

2. FAIR principles implementation and repositories

The LS RI data experts supported via EOSC-Life have driven cataloging and development of FAIR workflows for capturing data from FAIRsharing (see section 1). In addition EOSC-Life has established important components for making LS data FAIR:

Provenance standard for life science data: A provenance model is being developed and standardised under ISO 23494 to describe the history of data in life sciences in distributed environments, in order to assess reusability of data for further research and to improve reproducibility of research results. The standard supports compliance with the Nagoya Protocol.

FAIRassist: The FAIRassist wizard, part of the FAIRsharing resource, is under development to offer personalised guidance to discover resources, such as data and metadata standards and databases, that should be used to make data FAIR.

Terms4FAIRskills: Terms4FAIRskills is a terminology for the skills, competencies and knowledge necessary to make data FAIR and to keep it FAIR. The terminology can be used to assist with the creation and assessment of stewardship curricula, support development of training materials and to enable the formalisation of job descriptions and CVs with recognised, structured competencies.

Landscape mapping on sharing and re-use of health data: EOSC-Life is currently mapping the national landscapes on sharing and re-using health data to understand the implications for Life Science RIs and their services, especially with regards to data protection, appropriate safeguards (e.g. de-identification techniques), risk-based approaches, data ownership, and conditions for sharing and re-use of health data. The results will allow us to assess EU countries' preparedness for EOSC.

Training Open Calls: Our Training Open Calls offer funding to support training activities and provide expertise and guidance. Examples of funded projects are training and mentoring for *Open Life Science* Ambassadors, COVID-19 modelling, and cross-cluster training on modelling workflows with FATES for improving climate models.

Training the EOSC-Life community: Training—from staff to end users—on FAIR principles, using sensitive data in research, and the use of our data resources, tools and workflows, including ResOps, cloud native tools and technology for researchers. In addition, a FAIR hackathon is planned for early 2022 to train the open call project teams on assessing and improving FAIRness in their projects.

Recommendations on follow-up priorities linked to the further development of EOSC

1. Continuously populate EOSC with FAIR data. This can build on 100+ high-value research data sources already catalogued within EOSC-Life, the RO-Crate community effort for packaging research data with their metadata as well as common metadata work in EOSC Future building on EOSC Enhance.
2. Address provenance, reproducibility, traceability and secure management of sensitive data and advance common packaging tools and workflows for use in the cloud.

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

3. Technical, semantic, legal and organisational interoperability

Metadata specifications for tools and workflows: Community definitions of the metadata required for describing, documenting and registering a Computational Workflow and Computational Tool were developed and harmonised. The schema has been presented using Schema.org, as part of the Bioschemas community. *Bioschemas* markup enables discovery by search engines and other aggregators such as Google and OpenAIRE.

RO-Crate: (See section 1)

Semantic standards for life sciences are developed across the community in the form of ontologies (structured vocabularies are) made available from the Ontology Lookup Service. Cross-mapping between ontologies enables data interoperability across life sciences and provides domain neutral infrastructure and provenance for semantic mappings within EOSC. FAIRsharing lists a number of other standards, such as formats, models and schemas.

Toolbox for sharing of sensitive data: (See section 1)

Recommendations on follow-up priorities linked to the further development of EOSC

1. Create a common user access infrastructure for managing proposal submission and access to the infrastructure across all domains of life science and beyond.
2. A key priority going forward should be to address and harmonise the various data sovereignty issues around the storage and processing of data produced in various national infrastructures by transnational researchers.

4. Stewardship of data

EOSC-Life populates EOSC with data: the project helps each of the 13 LS RI develop and connect the repository infrastructure to EOSC. For example, if a particular dataset is being deployed into the cloud and connected to workflows, the project enhances the accessibility of mouse model data for researching biology and understanding diseases. EOSC-Life is developing data expertise within the LS RIs, helping to build a community of experts.

Extension of the COVID-19 Data Portal: The COVID-19 Data Portal was launched in April 2020 to bring together relevant datasets for sharing and analysis in an effort to accelerate coronavirus research. Within EOSC-Life the COVID-19 Data Portal is extended to mobilise open biomolecular data (500,000 records from the biomolecular and literature domains are available openly to users), to mobilise new SARS-CoV-2 data (currently >160,000 viral isolates with raw sequence data; 75% of the world's data flows through the Data Hubs) and to connect to clinical and epidemiological data.

FAIRsharing EOSC-Life Collection: (See section 1)

Cloud Data Deployment: Launched in January 2021 after a call from WP1 across 13 RIs, 8 projects will make their data resource available in an EOSC cloud instance, implementing FAIR metrics. The year-long projects have now been completed and publications and resources are becoming available, including training material and documentation to facilitate uptake in the community.

Clinical Research Metadata Repository: The MDR allows researchers to access clinical studies and related data objects, including COVID-19 data. These include, for example, protocols, information sheets and consent forms,

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

data management plans, statistical analysis plans, case report forms, results, publications, descriptive metadata, etc. MDR contributes to making clinical research data from all disease areas FAIR by increasing data Findability.

FAIRsharing, FAIRassist, Terms4FAIRskills and FAIRhackathon: See Section 2

EOSC Association Data Stewardship Curricula and Career Paths task force: EOSC-Life personnel are well represented in the task force; we aim to integrate an EOSC-Life use case into the task force activities, potentially in collaboration with EOSC Future.

Recommendations on follow-up priorities linked to the further development of EOSC

1. We identified the need for projects focusing on the development of data stewardship activities, competences and best practices to find adequate support and funding, to stimulate and foster cross-country, cross-discipline collaboration and alignment of practices. Recognising these data stewardship roles and professionalising career paths is also key to incentivise further development.

5. Cross-cluster collaboration activities and achievements

The Life Science RIs are the foundation for sustainability of EOSC-Life services. Building on long-standing coordination between the LS RI coordinators and on previous EC funded cluster project BioMedBridges (2012-2015) and CORBEL (2015-2020) the LS RI cluster has established tight cooperation based on a) strategic coordination within the LS RI Strategy Board established in 2015, and b) joint service provision pipelines supported by bilateral Memoranda of Understanding (MoU).

The experience from the rapid response to support COVID research and manage cross-disciplinary data flows during the pandemic shows the long-term value of maintaining active interfaces between the RIs. Societal challenges such as pandemics, climate change and food security, but also rapid technical advancement in the life sciences demonstrate the need for integrating across scientific disciplines and ongoing investments to ascertain integrated and cross-disciplinary RI services that can support advanced user projects.

All five EOSC Science Cluster projects (SCLs) unanimously express that the “cluster action” is a powerful enabler of European research excellence and competitiveness, and the cluster action is one of the EC's most effective and innovative network tools within the EU frameworks:

- The SCLs occupy a unique position in the R&I landscape and serve as engines for development of user-focussed interdisciplinary services
- The SCLs are uniquely sensitive to the challenges of Open Research data, data federation and preparation, because they are data producers and data consumers simultaneously. In this way they act as important bridges between the scientific communities, their infrastructures and the EOSC. The SCLs also provide vital links to the community and community governance.
- The SCLs build and maintain key community-centred initiatives. Their collective track-record of integrating infrastructure and managing large-scale user access is a significant asset for the EOSC.

Recommendations on follow-up priorities linked to the further development of EOSC

1. Strengthen inter-cluster collaboration via science-led projects that bring together services across research infrastructures to support projects and missions. By delivering solutions for major societal challenges such projects will drive the embedding of the EOSC infrastructure services for management of research digital outputs (e.g. data, models, software, preprints and workflows) to cover the full research life-cycle such that FAIR digital assets are reused and help accelerate the transition towards a socially inclusive green and digital

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

future, effective and responsive health system as well as being widely available to support evidence based policy-making.

2. Increase uptake of FAIR data management practices, including harmonised machine-actionable Data Management Plans by provisioning of supporting infrastructure and training for European Missions and Partnerships. Particular emphasis should be on use cases and proof of concepts to establish best practices to be adopted in the context of integration of heterogeneous datasets. Close alignment between EOSC and RI – in particular distributed RI – has the potential to contribute to local and regional socio-economic development. There is an opportunity to develop strategies for ESFRI / EOSC cooperation in the creation and strengthening of local / regional knowledge innovation hubs by aligning advanced facilities, data capabilities and upskilling of highly qualified staff.

6. Appendix 1: EOSC adoption in science - the EOSC-Life open calls

One of the goals within EOSC-Life with the support of demonstrators and user projects from the open and internal calls is to populate the EOSC with FAIR data resources, tools and workflows that will be open and accessible for re-use by scientific communities across RIs and beyond. These projects provide concrete solutions and resources ensuring FAIR data handling in the cloud for a spectrum of use cases. Here you can find details about the demonstrator projects, the WP1 open call projects, and the WP3 open call projects.

- In open call project APPID 1244 - **PombeMine**, the team have successfully built a new PombeMine database containing many of the major datasets from PomBase. This includes sequence features and curated data including the Gene Ontology, phenotypes, protein domains, physical and genetic interactions, disease associations and orthologue data. A set of predefined “template” searches with editable default parameters have been created to allow researchers to easily access these data. The team will continue to refine the representation of PomBase data in PombeMine, with additional data sources added and an increased range of template searches. Part of this will involve workflows linking PombeMine and HumanMine and PombeMine and the Image Data Resource.
- The open call project APPID 1211 - **OmicsDI Cloud** has reviewed and simplified the algorithm for calculation of OmicsDI impact metrics to provide these metrics for open datasets to third party websites. This API is now used in production mode by two third party sites, BioModels (example https://www.ebi.ac.uk/biomodels/MODEL4780784080) and PRIDE (example https://www.ebi.ac.uk/pride/archive?keyword=PXD005011). OmicsDI has been extended to support deployment of a local satellite instance, and a pilot site for systems biology is established as a demonstrator (http://wwwdev.ebi.ac.uk/Tools/omicsdi/ .
- Open call project APPID 1231 - **Macromolecular crystallography** has established the IceBear data management software in the cPouta environment maintained by the CSC IT Center for Science (ELIXIR Finland). The production server for the Biocenter Oulu has been moved to the cloud with automated uploading of images and metadata. Communication and exchange of metadata has been implemented for the ExiMC ISPyB of MAXIV synchrotron and connections to ESRF ISPyB are established. The cloud version of IceBear, IceBox, will be established in Finland early 2022 followed by the Weizmann Institute of Science. The team will explore possibilities to develop ML based scoring of crystallization drops based on the annotations from multiple instances hosted in the same cloud and at the high performance computing facility of CSC.

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

- Open Call project APPID 1234 **PDB-REDO Cloud** sees that FAIR protein structures have enabled the conversion of all PDB-REDO data converted to FAIR data formats including mmCIF, PDB, MTZ and JSON. Provenance tracking of PDB-REDO entries has been completed with versions of input data (PDB models and diffraction data) and all software used in calculations now recorded. The team have greatly enhanced documentation of PDB-REDO data by adding JSON schemas to all stored metadata. Persistent storage of PDB-REDO entries with full version stacks will be established. This will involve a new frontend for dataset creation based on PDB-REDO metadata queries and implementing a data structure describing sets of PDB(-REDO) entries that can be used for documenting research data in publications and other scientific output.
- Open Call project APPID 1220 - **Cloudification of the IMPC datasets** has modified and extended the existing IMPC ETL process to restructure data for the EOSC-Life API and load the data directly into a distributed database that is easily scalable, and provides high availability. An initial implementation of the API application has been created to provide a REST service supporting hypermedia aware browsers. This application has been deployed into a Kubernetes cluster and is publicly available.
- Open Call project APPID 1237 - **Cloudification of CryoEM data and metadata** has developed workflows for real-time submission of raw (and processed) data from electron microscopes to the cloud. This includes development of the web browser based frontend application to facilitate submission of the data to the OneData cloud solution. A new backend fs2od application running on the cloud provider manages real-time updates and data management (removal, archival, publication after the embargo period). Development of further tools for automated submission of electron microscopy data to public repositories (EMPIAR). Implementation of the established workflow at additional sites and finalization of the associated data catalogues.

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