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

D3.3 - Report on the work of the Open Call Projects

WP3 – Open Calls and Demonstrators
Lead Beneficiary: EMBL-HD and CSIC
WP leaders: Arina Rybina and Pauline Audergon (former WP leaders: Natalie Haley, Johanna Bischof, Frauke Leitner, Jose-Maria Carazo)
Contributing partner(s): listed below in detail

Authors of this deliverable: **Arina Rybina, Pauline Audergon, Claudia Pfander**

Contractual delivery date: **28 February 2023**
Actual delivery date: **18 August 2023**
H2020-INFRAEOSC-2018-2

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

Co-authors and contributors: Bérénice Batut, Martin Beracochea, Johanna Bischof, Michael Bottländer, Ivo Buchhalter, Jean-Marie Burel, Ibon Cancio, Jose- Maria Carazo, Nayari Castillo-Rutz, Maria Luisa Chiusano, Justin Clark-Casey, Yann Cointepas, Marco Comerci, Frederik Coppens, Erwan Corre, Cymon J. Cox, Sara Crockett, Romain David, Gianluca De-Moro, Ludovic de Rochefort, Ignacio Eguinoa, Katrina Exter, Jonathan Ewbank, Robert Finn, Juan José García-Granero, Carole Goble, Philipp Gormanns, Tobias Grantner, Philip Gribbon, Björn Grüning, Miguel Guevara, Kim Gurwitz, Natalie Haley, Henriette Harmse, Maarten Hekkelman, Anna Henger, Jean-Karim Heriche, Petr Holub, Martin Houde, Robbie P. Joosten, Nick Juty, Geoffrey Karnbach, Emma Karoune, Philip R. Kensche, Antje Keppler, Celine Kerfant, Martin Krajiczek, Carla Lancelotti, Frauke Leitner, Simone Leo, Dario Longo, Rebecca Ludwig, Marco Madella, Pedro Madrigal, Mario Magliulo, Jean-Francois Mangin, Jonathan Manning, Jose A. Marquez, Vera Matser, Rafaele Matteoni, Michaela Mayrhofer, Tomasz Miksa, Pablo Moreno, Engy Nasr, Steven Newhouse, Stelios Ninidakis, Nicolas Pade, Evangelos Pafilis, Maria Panagiotopoulou, Irene Papatheodorou, Helen Parkinson, Christina Pavloudi, Anastassis Perrakis, Isabelle Perseil, Luca Pireddu, Niki Popper, Antonis Potirakis, Marcus Povey, Andreas Rauber, Paolo Romano, Antonio Rosato, Javier Ruiz-Pérez, Valentin Schneider-Lunitz, Stian Soiland-Reyes, Ekaterina Sakharova, Ralf Seyfarth, Stian Soiland-Reyes, Andrey Solovyev, Elisabetta Spinazzola, Philipp Strubel, Jason Swedlow, Ellie Taverner, Sven Twardziok, Sameer Velankar, Alexandre Vignaud, Harald Wagener, Martin Weise, Hans Wienk, Dietmar Winkler, Haris Zafeiropoulos, Andrea Zaliani, Sara Zullino, Friederike Schmidt-Tremmel, Niklas Blomberg

WP3 leads, Open and Internal Call leads:
Arina Rybina; Euro-BioImaging ERIC Bio-Hub, EMBL-HD, DE
Pauline Audergon; Instruct-ERIC Hub, Oxford, UK
Claudia Pfander; Euro-BioImaging ERIC Bio-Hub, EMBL-HD, DE
Natalie Haley; Instruct-ERIC Hub, Oxford, UK
Johanna Bischof; Euro-BioImaging ERIC Bio-Hub, EMBL-HD, DE
Frauke Leitner; Euro-BioImaging ERIC Bio-Hub, EMBL-HD, DE
Nayari Castillo-Rutz; BBMRI-ERIC, AT
Jose-Maria Carazo; Instruct-ERIC, CSIC, ES

EOSC-Life Open and Internal Call project teams, EOSC-Life WPs experts and community:
Elisabetta Spinazzola, Sara Zullino, Mario Magliulo, Marco Comerci, Dario Longo; Institute of Biostructures and Bioimaging, CNR, Torino, IT (Euro-BioImaging)
Rafaele Matteoni; Institute of Biochemistry and Cell Biology, CNR, Monterotondo, INFRAFRONTIER, IT
Jean-Marie Burel, Jason Swedlow; University of Dundee (Euro-BioImaging)
Philipp Gormanns; INFRAFRONTIER GmbH, Neuherberg, DE
Andrea Zaliani; ITMP Fraunhofer, Hamburg, DE
Evangelos Pafilis, Antonis Potirakis, Stelios Ninidakis, Christina Pavloudi; Hellenic Centre for Marine Research
Haris Zafeiropoulos; Rega Institute, KU Leuven, BE
Gianluca De Moro, Cymon J. Cox; CCMAR, PT
Martin Beracochea, Ekaterina Sakharova, Robert Finn; EMBL-EBI, UK
Ibon Cancio; Universidad-del-Pais-Vasco, ES
Maria Luisa Chiusano; University of Naples Federico II, IT
Erwan Corre; SBR, FR
Katrina Exter; VLIZ, Flanders Marine Institute, BE

Nicolas Pade; EMBRC, FR

Robbie Joosten, Maarten Hekkelman, Hans Wienk, Anastassis Perrakis; NKI, Instruct-NL

Pedro Madrigal, Pablo Moreno, Andrey Solovyev, Jonathan Manning, Irene Papatheodorou: EMBL-EBI, UK

Emma Karoune: Historic England, UK

Celine Kerfant, Carla Lancelotti, Marco Madella: Universitat Pompeu Fabra, Barcelona, SP

Juan José García-Granero; HUMANE—Human Ecology and Archaeology, IMF, CSIC, Barcelona, SP

Javier Ruiz-Pérez; Texas A&M University, USA

Ignacio Eguinoa, Frederik Coppens; Flemish Institute for Life Sciences, BE

Björn Grüning; University of Freiburg, DE

Martin Weise, Tobias Grantner, Geoffrey Karnbach, Martin Krajiczek, Dietmar Winkler, Tomasz Miksa, Niki Popper, Andreas Rauber; TU Wien, AT

Jose A. Marquez; EMBL Grenoble, FR

Sameer Velankar; EMBL-EBI, UK

Sven Twardziok, Valentin Schneider-Lunitz, Philipp Strubel; Charité, Berlin, DE

Philip R. Kensche, Ivo Buchhalter; DKFZ, DE

Miguel Guevara, Alexandre Vignaud, Jean-Francois Mangin; CNRS, FR

Yann Cointepas; Neurospin, CEA, FR

Ludovic de Rochefort; Ventio

Michael Bottländer; Paris Diderot University

Engy Nasr, Bérénice Batut; University of Freiburg, DE

Ralf Seyfarth, Anna Henger; Biolytix AG, CH

Philip Gribbon; Fraunhofer IME, DE

Helen Parkinson, Justin Clark-Casey, Henriette Harmse, Vera Matser, Kim Gurwitz, Steven Newhouse, Henriette Harmse; EMBL-EBI, UK

Niklas Blomberg, Friederike Schmidt-Tremmel, Ellie Taverner; ELIXIR HUB

Antje Keppler; Euro-BoImaging ERIC Bio-Hub, EMBL-HD, DE

Antonio Rosato; CIRMMP, IT

Jean-Karim Heriche; EMBL-HD, DE

Carole Goble, Stian Soiland-Reyes, Nick Juty; University of Manchester, UK

Simone Leo; CRS4, IT

Harald Wagener; Charité, DE

Maria Panagiotopoulou; ECRIN, FR

Michaela Th. Mayrhofer; Petr-Holub; Sara Crockett; BBMRI, AT

Isabelle Perseil; INSERM, FR

Marcus Povey; Instruct-ERIC, UK

Rebecca Ludwig; EATRIS

Martin Houde, Romain David, Jonathan Ewbank; ERINHA, FR

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

Table of Contents

Executive Summary 5
Project Objectives 5
Detailed Report on the Deliverable 6
Description of work 6
Delivery and schedule 59
Adjustments 59

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 3.3 is a report on the Digital Life Sciences Open Call and two Internal Calls organised by EOSC-Life WP3[1]. The organisation of these Calls followed the successful integration and support of 8 Demonstrator projects[2] which provided the first concrete use cases in the initial phase of EOSC-Life. The three Calls overall supported 11 scientific user projects, selected to facilitate integration of concrete use-cases across Life Sciences domains into the European Open Science Cloud (EOSC)[3] framework.

Through the Calls, the practical goal was to facilitate co-creation of an open, digital collaborative space for life science research by developing FAIR[4] tools, workflows, resources, infrastructures, and guidelines together with the EOSC-Life RIs experts and communities.

We report in this Deliverable the following achievements:

1) Organisation of the EOSC-Life Open and Internal Calls;
2) Integrating and training the EOSC-Life WP3 Open Call[5] and Internal Call[6] project teams in EOSC-Life;
3) Activities for connecting project teams with EOSC-Life and LS-RI communities and dissemination of projects outcomes to broader communities;
4) Work done in the individual projects, their results, and impact of developed resources;
5) Recommendations from the EOSC-Life WP3 project teams and the EOSC-Life community for future Open Calls.

Project Objectives

With this deliverable, the project has contributed to the following objectives:

a. Organisation of three WP3 Calls for projects aiming at developing solutions for FAIR sharing of data, tools and workflows in the Life Science domain, with a specific focus on sensitive data projects and academia-industry collaborations in the second and third Call, respectively;
b. Selection of 11 ambitious pilot user projects, training and other initiatives set up and carried out together with EOSC-Life experts across WPs to facilitate onboarding these project teams in EOSC-Life, towards the EOSC framework and the LS-RI landscape;
c. Integration of WP3 Open & Internal Call project teams into the EOSC-Life community and LS-RI communities, and dissemination of solutions generated in these projects to broader communities.

An overview of achieved WP3 milestones and deliverables is summarised in Table 1.

---

[1] https://www.eosc-life.eu/about/work-packages/
[2] https://www.eosc-life.eu/services/demonstrators/
[3] https://eosc-portal.eu/about/eosc
[4] https://www.go-fair.org/fair-principles/
[5] https://www.eosc-life.eu/opencall/
[6] https://www.eosc-life.eu/industrycall/, https://www.eosc-life.eu/sensitivedatacall/

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

WP3 Milestones and Deliverables	Achievement month
D3.1 *Publication of generic guidelines for the organisation of topic-specific Open Calls*[7]	*September 2020*
MS8 *Launch of the first Open Call*[8]	*October 2020*
D3.2 *Demonstrator Report*[9]	*May 2021*
MS40 *Launch of the Internal Calls*[10]	*July 2021*

Table 1: WP3 deliverables and milestones achieved

Detailed Report on the Deliverable

Description of work

1. Organisation of three WP3 Calls for pilot projects

EOSC-Life WP3 organised three Digital Life Science Calls for pilot projects providing solutions for the creation of a collaborative digital space and FAIR sharing of data resources, tools, and infrastructures in the Life Science domain.

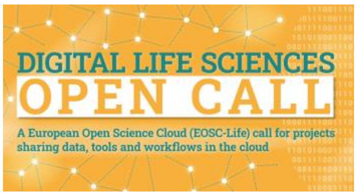
Following an internal scoping for project ideas, with "market research" that helped determine which topics should be addressed in the WP3 Calls, the following Calls were run and accomplished:

1. General Open Call for Life scientists[11]: to engage scientists internal or external to EOSC-Life
2. An Internal Call on Sensitive Data[12]: to support initiatives which aim at developing solutions for management of sensitive data and/or sharing workflows and tools specifically designed to deal with sensitive data
3. An Internal Call for Academia - Industry Collaborations[13]: to engage industries and SMEs in the EOSC initiative

---

[7] https://zenodo.org/record/4048442
[8] https://www.eosc-life.eu/opencall/
[9] https://zenodo.org/record/4817723
[10] https://www.eosc-life.eu/opencall/
[11] https://www.eosc-life.eu/opencall/
[12] https://www.eosc-life.eu/sensitivedatacall/
[13] https://www.eosc-life.eu/industrycall/

These successfully accomplished three Calls have been of high importance and relevance to the EOSC-life consortium as they led to implementation of 11 ambitious projects. This activity has been key for the onboarding of new scientific communities in the project and for increasing the visibility of EOSC-Life and EOSC across the Life Science domain.



1.1 Digital Life Science Open Call organisation

Following the successful implementation of Demonstrator projects[14,15], which provided concrete examples of cross-institutional work done in EOSC-Life for building an open digital and collaborative space for *Biological* and *Medical* research on the European level, the first WP3 Digital Life Sciences Open Call was launched on 15th September 2020.

The Call offered funding for the project team equivalent to roughly one full-time salary for one year (in the range of 70,000-85,000 EUR per project, subject to evaluation), as well as training and expert advice from within the EOSC-Life consortium, to enable the delivery of complex cloud-based life science research projects.

In preparation of the first Open Call, the following guideline documents were prepared:

- Applicant guidelines[16]
- Scientific reviewer guidelines[17]
- Technical evaluation guidelines[18]
- Moderator guidelines[19]

An overview of the guidelines developed, the outline of the WP3 call process, and background on topic selection are given in Deliverable 3.1[20].

1.2 Digital Life Science Internal Calls organisation

Following the Digital Life Sciences Open Call in 2020, two Internal Calls were launched in June 2021: “Internal Call for Sensitive Data[21]” and “Internal Call for Academia-Industry Collaborations[22]”. Application procedures for these Calls were set up re-using guidelines for the Open Call, adapted to Internal Call needs and with added topic-specific information and requirements.

---

[14] https://www.eosc-life.eu/services/demonstrators/
[15] https://zenodo.org/record/4817723
[16] https://instruct-eric.eu/network/eosc-life/eosclife-digital-life-sciences-open-call-applicant-guidelines.
[17] https://instruct-eric.eu/network/eosc-life/eosclife-digital-life-sciences-open-call-reviewer-guidelines.
[18] https://instruct-eric.eu/network/eosc-life/eosclife-digital-life-sciences-open-call-technical-evaluation-guidelines
[19] https://instruct-eric.eu/network/eosc-life/eosclife-digital-life-sciences-open-call-moderator-guidelines?t=eosclife
[20] https://zenodo.org/record/4048442
[21] https://www.eosc-life.eu/sensitivedatacall/
[22] https://www.eosc-life.eu/industrycall/

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

Digital Life Sciences Call for Sensitive Data

Due to the special requirements to handle and share sensitive data in the cloud, a specific Call was launched to support projects dealing with sensitive data aspects, such as management of sensitive data and/or sharing workflows and tools to operate with sensitive data. Sensitivity of data may arise from its nature (personal data in the sense of the GDPR[23], e.g., health data, biological samples and associated personal data, genetic data, individual research data, etc.), but can also be caused by biohazard concerns (e.g., Dual Use Research of Concern) or application of the Nagoya protocol[24]. Considering the challenging aspects and specific angles of sensitive data projects, preparation of this Call and guidance of incoming applications took place in close consultation and technical evaluation by the experts from EOSC-Life WP4 (“Policies, specifications and tools for secure management of sensitive data for research purposes”).

Digital Life Sciences Call for Academia - Industry Collaborations

The goal of this Call was to support projects that demonstrate how companies can participate in and benefit from Open Science. It therefore aimed at EOSC-Life consortium members collaborating with the private sector to jointly implement a project. Adaptations of application guidelines for this Call were implemented with the help of and based on recommendations from the Euro-BioImaging Industry Board[25].

In order to comply with the EOSC-Life and EC guidelines, funding in the Academia-Industry Call has been only available to applicants from among the existing EOSC-Life consortium members, while the industry partners were expected to contribute to the project *in kind*.

These two Digital Life Sciences Internal Calls offered up to 60,000-70,000 EUR per project for personnel costs, travel, and consumables to carry out the project, as well as training and expert advice from within the EOSC-Life consortium, to enable the delivery of their project.

1.3 Open and Internal Call Organisation - Procedures

Details of the Open Call and Internal Call organisation and procedures are summarised in the following.

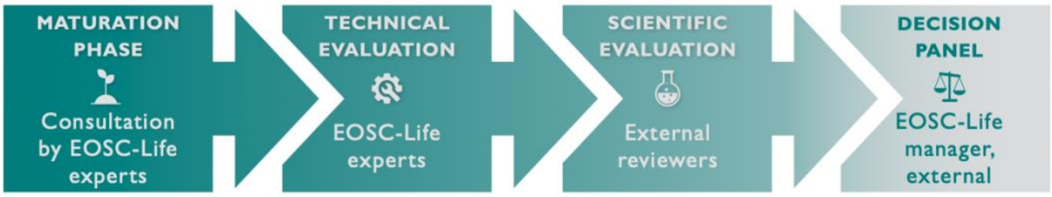


Figure 1: EOSC-Life WP3 Open and Internal Calls evaluation procedures (Source: Euro-BioImaging Annual report 2021[26])

---

[23] https://gdpr-info.eu/

[24] https://treaties.un.org/Pages/ViewDetails.aspx?src=IND&mtdsg_no=XXVII-8-b&chapter=27&clang=_en

[25] https://www.eurobioimaging.eu/about-us/industry-board

[26] https://www.eurobioimaging.eu/upload/document_gallery/Euro-BioImaging%20Annual_Report_2021.pdf

Maturation phase

During the application process, all potential applicants were requested to contact EOSC-Life experts from WP1 (*Data experts*), WP2 (*Tools and Workflow experts*), WP4 (*Sensitive Data experts*) and/or WP7 (*Cloud experts*) to discuss their project, ensure that the project was aligning with the EOSC-Life Calls, get feedback about their project proposal, and, if interested, to discuss the possibility of receiving support from the EOSC-Life experts during project implementation. These interactions were supported and mediated by the WP3 open call managers, which was key for the success of the Calls. In total, 144 applicants made use of the consultation in this so-called "maturation phase" of at least one EOSC-Life expert (122 applicants in the first Open Call and 22 applicants in the two Internal Calls).

The maturation phase has been an important and valuable step for increasing the quality of incoming proposals and performing a timely eligibility and feasibility check on project plans, tools, software and ideas also in regard to their alignment with the scope of EOSC-Life Calls and expertise in the consortium. While advancing promising applications on the one hand, the maturation phase naturally allowed for EOSC-Life to have an impact beyond the successful incoming projects as many scientists received valuable feedback and advice, even if their project did not receive funding via the Call.

The important process of cross-disciplinary consultation by EOSC-Life experts across relevant WPs during the application phase in the Open and Internal Calls, however, also added significant workloads on the WP3 open call managers as well as the EOSC-Life experts involved - which should be acknowledged and taken into consideration in the future.

Technical evaluation

From the 144 potential applicants that consulted with the EOSC-Life experts, 91 applications were received through the three EOSC-Life WP3 Calls. All complete applications first underwent a technical evaluation by EOSC-Life experts on data (WP1), tools and workflow (WP2), sensitive data (WP4), and/or cloud storage (WP7), depending on the project focus and alignment with the respective areas.

The technical experts assessed individual proposals on the following criteria:

- Technical feasibility
- Impact
- Maturity of proposal

Applications that clearly passed the technical evaluation were straight away directed to the scientific evaluation. In the Open Call, 61 out of 70 applications sent for review passed the technical evaluation and were redirected to the scientific evaluation. In the Calls for Sensitive Data projects and Academia-Industry collaborations, no hard rejection of applications could be done at the technical evaluation stage by EOSC-Life experts due to the consortium internal nature of these Calls and the risk of any bias. Consequently, all 19 applications which were sent to the technical evaluation in these Calls were also sent to the scientific review.

Scientific evaluation

The scientific evaluation was done by independent scientific reviewers external to the EOSC-Life consortium (a minimum of two, and maximum of four, experts per application). The independent scientific reviewers were invited by WP3 to evaluate the proposals that passed technical evaluation according to three main criteria:

- Scientific Excellence
- Impact
- Sustainability

More details can be found in the Open Calls Reviewer guidelines[27].

Panel decision

A decision panel composed of EOSC-Life managers, coordination and external scientific panel members was constituted to make the final selection of projects for funding. The panel was presented with proposals shortlisted according to scientific and technical excellence, and asked to make the final selection based on the following criteria:

- Evaluated factors comprising: Scientific Excellence, Impact, Sustainability
- Proposal maturity and realistic timeline: start within 2 months, duration ~1 year
- Portfolio balance:
  - Diversity of scientific domains covered
  - Gender and geographic diversity of project teams
- Portfolio funding constraints

Management of the Calls and the applications

The management of WP3 Open and Internal Call user applications has been carried out using ARIA[28] - an application management system developed by Instruct-ERIC. The application set up for each Call was implemented with support by the EOSC-Life WP5 ARIA team (*User management and access services*). The EOSC-Life WP3 collaborated closely with the respective software developers so that specific needs of individual Calls were considered, and the platform customised accordingly.

The access management system allowed for application submission, moderation of applications by WP3 leads and Call managers, independent scientific peer-review by external experts, and technical review by EOSC-Life experts to be conducted centrally in a unique point of access.

Taking advantage of these procedures, EOSC-Life WP1 and WP9 used ARIA for the management of their Calls, which allowed for consistency and easier acceptance and familiarity of this management system across the consortium.

While the Open Calls have been managed through adapted access management system tracks in ARIA, an additional helpdesk tool (Freshdesk) had been implemented via the Freshworks[29] platform - for communicating with the applicants and the EOSC-Life experts during the proposal maturation process phase prior to application submission. This platform also served for consultative purposes and advice as well as for establishing interactions between EOSC-Life experts from different WPs to the applying Open and Internal Call team.

---

[27] https://instruct-eric.org/network/eosc-life/eosclife-digital-life-sciences-open-call-reviewer-guidelines

[28] https://instruct-eric.org/help/about-aria

[29] https://www.freshworks.com/

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

2. Integration of Open Call project teams in EOSC-Life and LS RIs

Once the Open Call, and later on the Internal Call, projects were selected, WP3 with the support of other WPs organised a series of events to initiate onboarding and supporting the project teams and introducing the resources and expertise available in EOSC-Life that could benefit their respective projects.

2.1 Organisation of Orientation meetings

A remote one-day event “The Orientation meeting” was organised by WP3 in September 2021 at the start of the Open Call projects. The aim of this “Kick Off” event was to introduce to the project teams the EOSC-Life project and its participants and give them a better idea of the expertise and the resources available through EOSC-Life and RI communities that may be of use and benefit for their projects. Experts from WP1 (*Data*), WP2 (*Tools and Workflows*), WP6 (*Provenance*), WP7 (*Cloud Storage*), and WP9 (*Training*) were invited to introduce key resources/services related to their WP to the project teams, while the project teams were asked to introduce their project goals and needs. The presentations session was then followed by interactive sessions between the experts and the teams.

Another remote “Orientation meeting” for the projects selected during the Internal Calls was run by WP3 in February 2022 at the start of these projects. Although most of the members of the teams were already part of the consortium, this was an important step to inform the project teams about the expertise and resources developed in EOSC-Life of potential relevance for their project implementation. The meeting also aimed to introduce the external industry partners to EOSC-Life. Together with the project teams, experts from WP1 (*Data*), WP2 (*Tools and Workflows*), WP4 (*Sensitive Data*), WP7 (*Cloud Storage*), and WP9 (*Training*) participated, following a similar format as for the induction of the Open Call projects.

Materials from both Orientation meetings, e.g. presentations and recordings, were made openly available to all participants in the shared EOSC-Life Google Drive.

2.2 Individual meetings with project teams

Apart from organising joint Orientation meetings for all project teams, WP3 collaborated with WP1 task 2.2 experts (“FAIR Access to Data/Data Resources”) to organise regular individual meetings with each of the Open Call and Internal Call project teams and EOSC-life experts from WP1 (*Data*), WP2 (*Tools and Workflows*), WP7 (*Cloud storage*), WP4 (*Sensitive Data*), and WP14 (*Covid Trial Repository*) that were relevant to the projects.

The aim of these individual meetings was to assess the progress in individual projects and the arising needs of the project teams in terms of resources and support. In addition, the session facilitated establishing new contacts and strengthening exchange between the project teams and EOSC-Life members that could provide support and training to the teams and inform them of potential opportunities within the EOSC-Life and EOSC landscape as well as in LS RI communities.

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

2.3 Organisation of the EOSC-Life FAIR Hackathon

The WP3 has been working closely together with EOSC-Life experts from several WPs (WP1, WP2, WP3, WP6) and organised a FAIR Hackathon for the project teams of the WP3 Open Call[30] and Internal Call[31] project teams, WP1 Open Call[32] project teams as well as the WP3 Demonstrators[33] project teams.

The goals of the FAIR Hackathon were:

- to facilitate training of the project teams in assessing and improving FAIRness in their projects,
- to orient and introduce the teams to available resources in EOSC-Life, also based on examples and solutions provided by EOSC-Life experts,
- to address challenges and gaps, and to identify further support needs for increasing FAIRness in individual projects.

The EOSC-Life Hackathon ran as a series of two sessions - the "Background Session" in December 2021, and the "Hackathon Session" in February 2022. The FAIR Hackathon Background Session served to introduce the project teams to a spectrum of EOSC-Life related topics, available resources, and FAIRification processes.

The event was run as an interactive Webinar by the experts covering three main subject areas:

- ➢ Session 1: Data life cycle for the distributed provenance chain (WP6)
- ➢ Session 2: Resources and tools for data FAIRification (WP1)
- ➢ Session 3: Tools and workflows FAIRification (WP2)

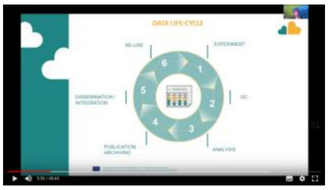
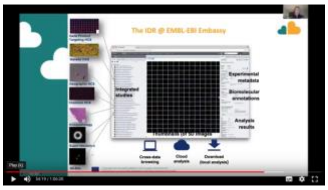
Data life cycle and the distributed provenance chain	**Resources and tools for data FAIRification**	**Tools and Workflows FAIRification**
 Topics covered: FAIR data principles, provenance / data life cycle, metadata	 Topics covered: FAIRplus, RDMkit, IDR, Ontology services	 Topics covered: WorkflowHub, RO-Crate, LifeMonitor

Figure 2: EOSC-Life FAIR Hackathon sessions and topics overview

The training materials of the EOSC-Life FAIR Hackathon (presentations and recordings) are openly available and can be accessed through the EOSC-Life Training webpage[34].

---

[30] https://www.eosc-life.eu/opencall/
[31] https://www.eosc-life.eu/industrycall/
[32] https://www.eosc-life.eu/services/ri-nominated-resources-from-wp1/
[33] https://www.eosc-life.eu/services/demonstrators/
[34] https://www.eosc-life.eu/services/training/

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

Following the introduction and the raised awareness in the FAIR Hackathon Background Session, the Hackathon Session ran in a format of joining interactive webinars, hands-on sessions and project specific discussions between the EOSC-Life experts and the project teams. After a very brief introduction by the EOSC-life experts, the project teams had the possibility to choose specific sessions of interest with an option to dial into any of the five sessions running in parallel:

- WorkflowHub (WP2)
- RO-Crate (WP2)
- Drop-in clinic for 10 Simple Rules for making a software tool workflow ready (WP2)
- Ontology: OLS and OxO API use cases (WP1)
- Provenance steps along the material/data life cycle (WP6)

Overall, this event provided a very fruitful opportunity and platform for goal-oriented hands-on teaching, introducing the EOSC-Life experts and topics in a more targeted way and establishing networking and collaborations between EOSC-Life experts and the teams across the EOSC-Life community. Overall, 27 participants joined this event.

2.4 EOSC-Life FAIR Assessment Support

Following the EOSC-Life FAIR Hackathon sessions, WP3 helped further connect the project teams with the experts in a targeted way such as through establishing links to the EOSC-Life FAIR Assessment and FAIRification group organised and run by experts from WP1, WP2 and WP6. This WP1 FAIR Assessment initiative helped several Open and Internal Call teams to dive deeper into FAIRification and improving FAIR aspects of their project tools and resources, and to receive a thorough consultation and support from the EOSC-Life experts for their specific project needs.

2.5 EOSC-Life WP9 Training Open Call

The EOSC-Life Training Open Call was organised by WP9 and allowed communities to apply for funding to run a cross-RI course or a workshop related to dissemination of subjects and topics related to EOSC-Life developments. This format offered a great funding opportunity also for the Open and Internal Call teams to team up with EOSC-Life experts to develop a proposal and upon successful outcome to set up and run together training related to their project topic (e.g. the "Open Phytoliths Community Training Workshops and E-learning Courses"[35]).

WP3 supported WP9 in running the 3rd Training Open Call[36] and selecting excellent proposals for high quality workshops in collaboration with the EOSC-Life experts and projects teams. At the same time, WP3 helped the teams to connect with suitable partners in EOSC-Life and RI communities.

2.6 Dissemination supported through EOSC-Life

Along the continuous integration of the project teams in EOSC-Life, WP10 supported WP3 by highlighting the progress of the Open and Internal Call projects, as well as the teams' participation in workshops, conferences and webinars, through the regular consortium-wide EOSC-Life

---

[35] https://open-phytoliths.netlify.app/course/open-res-workshop/

[36] https://www.eosc-life.eu/services/open-call-training/

newsletter and on the EOSC-Life website[37]. This support helped increase the visibility of the ongoing work in individual projects across LS-RIs and partners.

"EOSC-Life AGM 2022"

During the 3rd EOSC-Life Annual General Meeting, which took place in a hybrid format in March 2022 in Brussels, several Open and Internal Call teams presented the progress of their projects to the EOSC-Life community, the invited stakeholders from the EC and the public. During the internal part of the meeting the teams also had the opportunity to contribute to interactive sessions to discuss consortium strategies for dissemination and sustainability of EOSC-Life expertise, tools and solutions.

"BioImaging and the EOSC workshop"

On 19-20th April 2023, Euro-BioImaging, its Industry Board, WP3 and WP9 jointly organised a 1.5-day workshop on "BioImaging and the EOSC"[38] with more than 85 participants. The aim of the workshop was to bring together the EOSC-Life community, Euro-BioImaging facility staff working on image data management, analysis and workflow development, and representatives from leading imaging companies. Several teams from the Open and Internal Calls as well as experts from WP1 and WP2 were invited to participate in this workshop and present their EOSC-Life projects or activities.

The meeting helped increase visibility of various EOSC-Life initiatives and outputs, especially those that can be beneficial for the imaging community and encouraged exchange between different stakeholders. It provided participants with an insight into the wider EOSC landscape and the opportunities it offers, while enabling a fruitful exchange on long-term sustainability with the participating panellists from academia, research infrastructures, publishers and policy makers.

"EOSC-Life AGM 2023"

The final big EOSC-Life dissemination highlights for the Open Call and Internal Call teams took place at the 4th and final EOSC-Life Annual General Meeting (AGM)[39] which ran face to face on 27-28 June, 2023 in Brussels. This remarkably successful concluding consortium event brought together the EOSC-Life RIs communities, experts and project teams, as well as representatives from other EC projects and science clusters, the EOSC Association and the EC - to present and discuss the achievements, impact and the role of EOSC-Life, to strengthen the LS RI community and partners, and to discuss sustainability aspects for the solutions and expertise generated and gained in EOSC-Life, and possible future perspectives and directions.

WP3 in collaboration with WP1, WP9 and the EOSC-Life coordination (WP11) organised several sessions during the event, where the Open and Internal Call teams were invited to present their projects outcomes and future plans either during the public or the consortium internal session, and join interactive round-table sustainability discussions with all participants including a session on data re-use organised by WP8 (*International Impact, Innovation and Sustainability*).

The interactive sessions also provided a very fruitful platform for establishing more connections in the LS RI community and discussing together the following topics of acute relevance:

➢ Sustainability of Open Calls outputs

---

[37] https://www.eosc-life.eu/news-events/
[38] https://www.eurobioimaging-industryboard.com/workshop_report_2023/
[39] https://www.eosc-life.eu/news/final-eosc-life-annual-general-meeting/

- Sustainability of experts’ groups and community
- Ideal Open Calls framework in future initiatives
- EOSC-Life user projects vision for EOSC
- How recommendations (sustainability paper) can turn into practical solutions
- Importance of a LS RI cluster project

3. Lessons learned - Recommendation for future Open Calls

WP3 in collaboration with WP1, experts from other WPs, and the project teams, collected feedback and recommendations on the aspects considered important for future Open Calls organisation and implementation from the EOSC-Life community perspective. This collective input, gathered throughout the EOSC-Life project, including at the 3th and 4th EOSC-Life AGM’s, is consolidated in *Figure 3* and in *Table 2*.

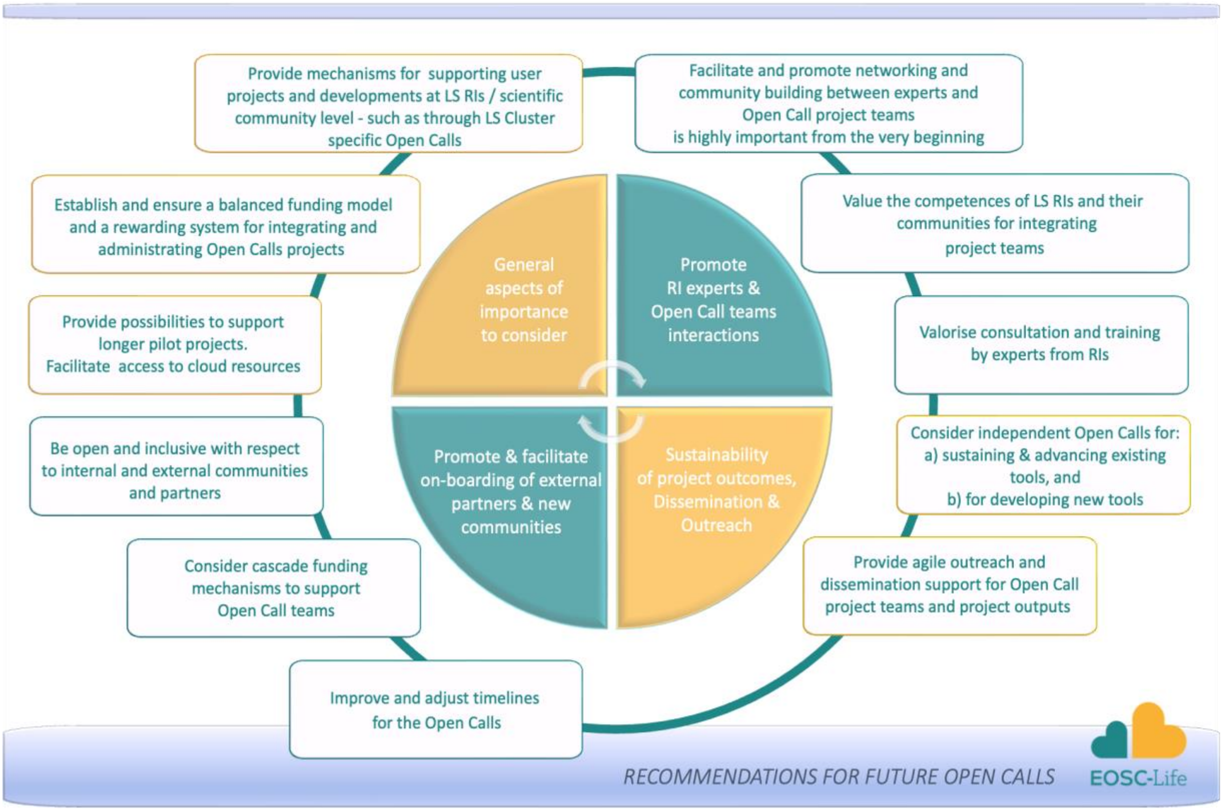


Figure 3: EOSC-Life community recommendations for future Open Calls

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

RECOMMENDATIONS	COMMENTS & IMPLEMENTATION STRATEGIES
	I. General aspects of importance to consider
Provide mechanisms for supporting user projects and developments at LS RIs / scientific community level - such as through LS Cluster specific Open Calls	*We need to have funding mechanisms in place to continue supporting ambitious and excellent pilot projects at the LS RI and community level and not only on cross-Cluster level. Life science communities need resources to develop and improve their own tools and services for FAIR data and workflow management that fit their specific needs. The EOSC-Life project has provided a set of FAIR resources and procedures - adapting, maintaining and expanding existing tools and expertise can provide an efficient pathway for onboarding new solutions and communities.*
Provide possibilities to support longer pilot projects	*Supporting longer projects would allow for more ambitious and challenging projects to be conducted and for establishing closer interactions and collaborations with other teams, experts and RIs communities which is difficult to achieve in 1 year.*
Facilitate access to cloud resources	*Although several project teams expressed an interest in accessing cloud and computing resources with EOSC-Life support, few partners actually benefited from it due to 1) lack of maturity/expertise of project teams, 2) the complexity of the process to access cloud resources via EOSC-Life, 3) The lack of sustainable cloud resources support that could be provided for the Open and Internal Calls project teams. This is an important aspect to be addressed in future EOSC Calls.*
Achieving high goals requires appropriate investment in human resources: **Establish and ensure a balanced funding model and a rewarding system for integrating and administrating Open Calls projects**	*Organising, administering and managing Open Calls, consulting and guiding researchers, performing technical and scientific evaluation of incoming project proposals, and providing continuous support and training to the teams during project implementation, are processes that demand a lot of human resources (with the respective technical, scientific and management competences). For successful implementation of the Open Calls within a consortium, it is therefore crucial to secure and assign sufficient financial resources to the consortium partners such as e.g., RIs experts and managers for their support activities and also reward contributions by external partners (e.g., reviewers). A realistic prognose of commitments versus provided resources allocations also ensures minimising project dependence on in kind requests and contributions.*
	II. Promote RI experts and Open Call teams interactions
Facilitate and promote networking and community building between experts and Open Call project teams is highly important from the very beginning	*Early and efficient on-boarding of Open Call teams within a consortium and interactions with the consortium experts is crucial to foster good collaboration and efficient support to the teams. The organisation of general Orientation meetings, early individual project team meetings with relevant experts for the teams, and the cross-WP and cross-RI FAIR Hackathon sessions provided efficient platforms for a fruitful match making between the Open Call project teams and EOSC-Life experts.*

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

Value the competences of LS RIs and their communities for integrating project teams	*The interdisciplinary training and on-boarding of Open Call project teams by experts from different RIs has been very important for certain incoming communities which were not already a part of a specific RI community to get trained or familiarised with tools and expertise an RI might have available and that would be beneficial for the project.*
Valorise consultation and training by experts from RIs	*Based on the experience in EOSC-Life, high commitment and availability of RI experts is needed to enable targeted training, consultation and directing incoming projects onto efficient tracks towards a successful project implementation. Along this way, regular exchange between the experts is also important for continuously expanding and improving competences. Therefore, results in Open Calls projects highly benefit from allocating more financial resources (PMs) to the RIs experts from their training and support activities to project teams.*
III. Promote and facilitate on-boarding of external partners and new communities	
Be open and inclusive with respect to internal and external communities and partners	*One of the main objectives of the EOSC initiatives is to develop high-quality interoperable tools and standards beneficial for a broad scientific community and integrate contributions and applications from researchers across disciplines and sectors. Progressive advancement of technical solutions and resources, and further development of available approaches can only be achieved by supporting ambitious, innovative projects and in addition to engaging with RI-internal communities by allowing integration of external communities and partners (including industries and SMEs). This not only allows for the broadest and most widely applicable set of tools and solutions to be developed, but also increases the visibility of RIs and the EOSC with a wider group of stakeholders.*
Consider cascade funding mechanisms to support Open Call teams	*The required onboarding of new Open Call teams in the EOSC-Life consortium has been a heavy administrative burden for the consortium and the teams and is in some cases even disproportionate - in terms of administrative costs to the funding distributed. The predefined bureaucracy framework also prevented WP3 from on-boarding teams external to the consortium in the second and third Call (e.g., industry & SME partners or new communities not represented in EOSC-Life that would gain and bring benefits from integration into the EOSC-Life project). The use of cascade funding to support integration of external partners should facilitate this process, particularly as the required sum per project is well aligned with usual cascade funding limits.*
Improve and adjust timelines for the Open Calls	*Raising awareness among new communities and conceptualization of projects take time. In particular R&D timelines in the private sector are significantly longer than in academia, while for academia-industry collaborations contractual issues might have to be resolved prior to proposal submission. Therefore, future projects that target SME and industry participation should have realistic timelines for open calls (minimal 6 months to submission), a predefined, minimal framework, preferably employing cascade funding, and sufficient resources for outreach to provide better opportunity for new project teams to participate.*

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

IV. Sustainability of project outcomes	
Consider independent Open Calls for a) sustaining & advancing existing tools and b) for developing new tools	*Sustainability of Open Call projects outcomes can be very challenging especially in absence of sources of subsequent funding (such as Open Calls) that would allow continuation of work, developments and maintenance. At the moment this follow-up relies on resources and possibilities through hosting institutions, RIs or applying for funding through other projects.* *This challenge is particularly the case for new services developed, as the project teams often need time to reach their new user base and onboard their community with their initiative; which is key for driving sustainability. For several project teams, the EOSC-Life Calls were actually used to further develop and sustain existing tools, services and progressing developments.* *This lack of dedicated funding sources for sustaining tools could be addressed through an Open Call track targeted at sustaining and advancing newly generated or existing tools which are useful and not redundant for the community. (For more recommendations on sustainability aspects from the EOSC-Life community, please see the preprint: "Be Sustainable Recommendations" for FAIR Resources in Life Sciences research: EOSC-Life's Lessons[40])*
V. Dissemination / Outreach	
Provide agile outreach and dissemination support for Open Call project teams and project outputs	*Several project teams have been struggling to reach users that are outside of their known user base to properly disseminate their tools and get broad uptake by the community. Continuous activation and use of consortium and partner networks to disseminate new initiatives and results that can be relevant to a broad community is essential. Also, engagement of communication and community building managers are highly valuable and important, together with availability of resources for generating attractive outreach materials.*

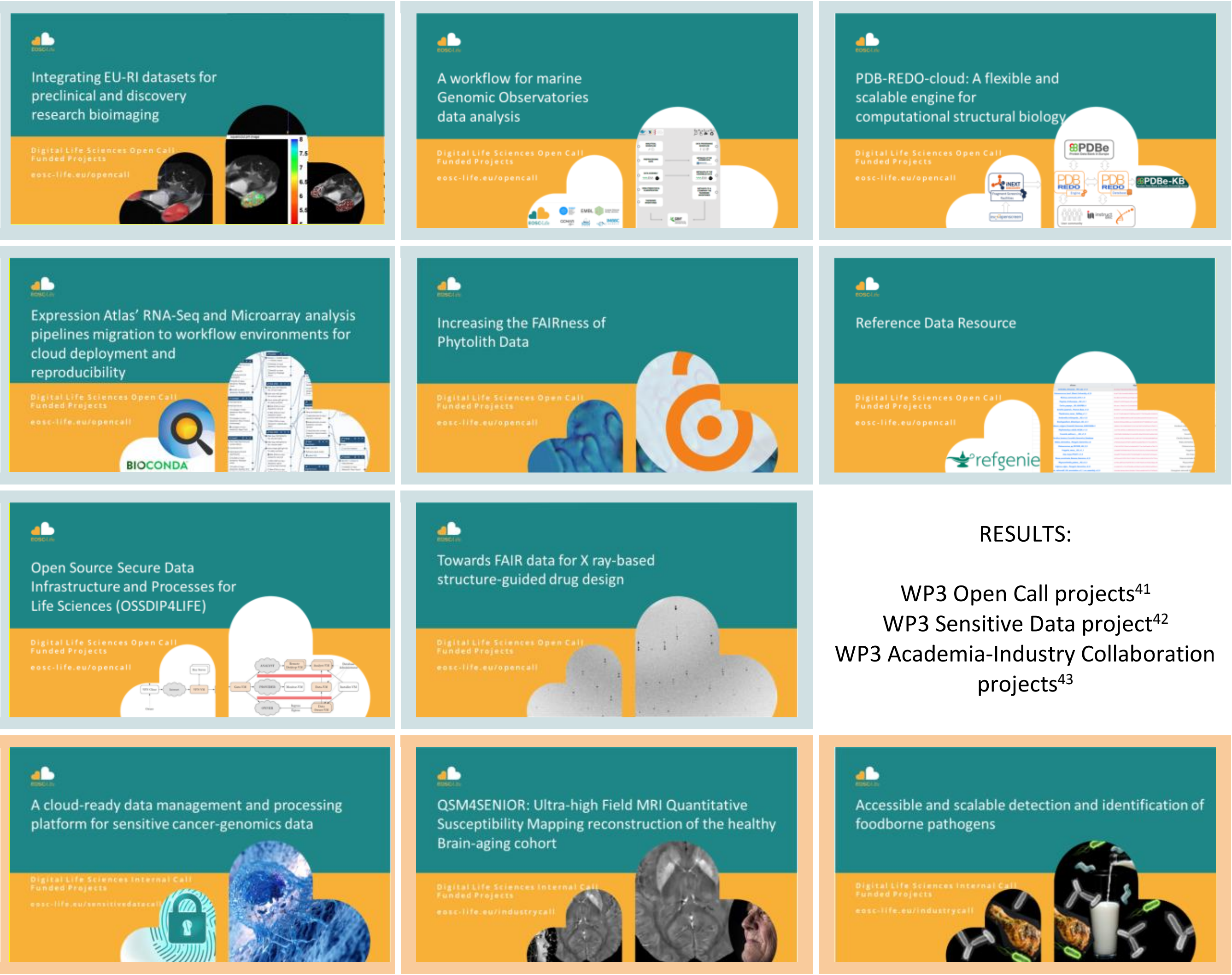
Table 2: EOSC-Life community recommendations for future Open Calls

4. Description of Work done in the Open Call and Internal Call projects

Through the three WP3 EOSC-Life Digital Life Science Calls, 11 pilot projects have been selected for funding and to promote building an open, digital collaborative space across Life Science domains for biological and medical research. An overview of these projects and the outcomes is illustrated in *Figure 4* and summarised in *Table 3*. The accomplished work and achievements of each project in the Open and Internal Calls is described in detail in the following sections.

---

[40] https://zenodo.org/record/8059830



Integrating EU-RI datasets for preclinical and discovery research bioimaging

Digital Life Sciences Open Call Funded Projects

eosc-life.eu/opencall

A workflow for marine Genomic Observatories data analysis

Digital Life Sciences Open Call Funded Projects

eosc-life.eu/opencall

PDB-REDO-cloud: A flexible and scalable engine for computational structural biology

Digital Life Sciences Open Call Funded Projects

eosc-life.eu/opencall

Expression Atlas' RNA-Seq and Microarray analysis pipelines migration to workflow environments for cloud deployment and reproducibility

Digital Life Sciences Open Call Funded Projects

eosc-life.eu/opencall

Increasing the FAIRness of Phytolith Data

Digital Life Sciences Open Call Funded Projects

eosc-life.eu/opencall

Reference Data Resource

Digital Life Sciences Open Call Funded Projects

eosc-life.eu/opencall

Open Source Secure Data Infrastructure and Processes for Life Sciences (OSSDIP4LIFE)

Digital Life Sciences Open Call Funded Projects

eosc-life.eu/opencall

Towards FAIR data for X ray-based structure-guided drug design

Digital Life Sciences Open Call Funded Projects

eosc-life.eu/opencall

RESULTS:

WP3 Open Call projects[41]
WP3 Sensitive Data project[42]
WP3 Academia-Industry Collaboration projects[43]

A cloud-ready data management and processing platform for sensitive cancer-genomics data

Digital Life Sciences Internal Call Funded Projects

eosc-life.eu/sensitivedatacall

QSM4SENIOR: Ultra-high Field MRI Quantitative Susceptibility Mapping reconstruction of the healthy Brain-aging cohort

Digital Life Sciences Internal Call Funded Projects

eosc-life.eu/industrycall

Accessible and scalable detection and identification of foodborne pathogens

Digital Life Sciences Internal Call Funded Projects

eosc-life.eu/industrycall

Figure 4: Illustration of all WP3 Open Call and Internal Call projects

---

[41] https://www.eosc-life.eu/opencall/
[42] https://www.eosc-life.eu/sensitivedatacall/
[43] https://www.eosc-life.eu/industrycall/

Digital Life Sciences Open Call[44] (OC) and Internal Call[45] (IC) projects

PID	Project title	p.	References to sustainable project outcomes
14176 (OC)	*Integrating several EU-RI datasets with focus on preclinical and discovery research bioimaging*	21	https://workflowhub.eu/workflows/516?version=1
14325 (OC)	*A workflow for marine Genomic Observatories data analysis*	24	Homepage: https://data.emobon.embrc.eu/MetaGOflow/, Github: https://github.com/emo-bon/MetaGOflow Use case: https://zenodo.org/record/7771821
14308 (OC)	*PDB-REDO-cloud: A flexible and scalable engine for computational structural biology*	27	API documentation: https://pdb-redo.eu/api-doc Publication: https://doi.org/10.1107/S2059798323003595 Presentation, PDB-REDO-cloud integration test: https://www.ccp4.ac.uk/schools/APS-2023/
14157 (OC)	*Expression Atlas' RNA-Seq and Microarray analysis pipelines migration to workflow environments for cloud deployment and reproducibility*	31	Homepage: https://www.ebi.ac.uk/gxa/home
13557 (OC)	*Increasing the FAIRness of Phytolith Data*	33	Github: https://open-phytoliths.github.io/FAIR-phytoliths/ Training workshops: https://youtu.be/yNcLApDHxFY Community website: https://open-phytoliths.netlify.app/ Publication: https://www.nature.com/articles/s41597-023-02296-8
13840 (OC)	*Reference Data Resource*	36	Github: https://github.com/refgenie/plantref Pilot Server: http://plantref.databio.org/
13888 (OC)	*Open Source Secure Data Infrastructure and Processes for Life Sciences (OSSDIP4LIFE)*	38	Homepage: https://www.ifs.tuwien.ac.at/infrastructures/ossdip/ Gitlab: https://gitlab.tuwien.ac.at/martin.weise/ossdip Publication: https://datascience.codata.org/articles/10.5334/dsj-2022-004
14366 (OC)	*Towards FAIR data for X ray-based structure-guided drug design*	40	PDBx/mmCIF Dictionary Resources: https://mmcif.wwpdb.org/ PDBe homepage: https://www.ebi.ac.uk/pdbe/
17410 (IC)	*A cloud-ready data management and processing platform for sensitive cancer-genomics data*	43	Demo site: https://otp-demo.bihealth.org/ Documentation: https://gitlab.com/one-touch-pipeline/OTP-Documentation
17989 (IC)	*QSM4SENIOR: Ultra-high Field MRI Quantitative Susceptibility Mapping reconstruction of the healthy Brain-aging cohort*	46	Conference presentations: https://zenodo.org/record/7303615#.ZEquQXZBzEY https://zenodo.org/record/7871306#.ZEqDIHZBzEY
17836 (IC)	*Accessible and scalable detection and identification of foodborne pathogens*	51	Training materials: https://training.galaxyproject.org/training-material/topics/metagenomics/tutorials/pathogen-detection-from-nanopore-foodborne-data/tutorial.html

Table 3: WP3 Open Call and Internal Call projects overview with references to project outcomes and resource

---

[44] https://www.eosc-life.eu/opencall/

[45] https://www.eosc-life.eu/industrycall/

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

4.1 Digital Life Sciences Open Call project - PID 14176

Integrating EU-RI datasets for preclinical and discovery research bioimaging

Project team: Elisabetta Spinazzola, Sara Zullino, Jean-Marie Burel, Marco Comerci, Philipp Gormanns, Dario Longo, Mario Magliulo, Rafaele Matteoni, Jason Swedlow, Andrea Zaliani

PROJECT DESCRIPTION

European RIs generate an increasing amount of data across different Life Science domains, such as animal model strains, chemical entities and/or biomedical imaging. Each has developed their own platforms for the annotation and standardisation of both data and metadata produced within their respective field.

In particular, INFRAFRONTIER - European Mutant Mouse Archive[46] (EMMA), EU-OpenScreen[47], EuroBioImaging[48] IDR[49] and EuroBioImaging PIDAR[50] provide public repositories enriched with metadata descriptions. To address the challenges of integrating several EU-RI datasets focusing on preclinical and to improve the discovery in research bioimaging, we developed tools to perform cross queries across several RIs.

The novel aspects of this tool are mainly:

- User friendly Interface;
- API access i.e. the user has the flexibility to query among the datasets easily with a simple API, intuitive for researchers and biomedical users. Possibility of making a query between different platforms and repositories, using a common mechanism;
- Use of ontologies as a means for cross resources investigations.

The project follows the FAIR principles in the treatment of data and datasets.

PROJECT RESULTS AND ACHIEVEMENTS

The results achieved in the project are summarised in the following.

The EuroBioImaging - PIDAR Preclinical Image Dataset Repository[51], a public repository of metadata associated with preclinical image datasets has been curated using several ontologies to allow the relationships between multiple resources.

For drugs and contrast agents, we used the chEB[52] ontology and the references from the EU-OpenScreen database. INFRAFRONTIER mouse strain data enriched with MGI[53] disease model annotations as well as phenotypes from MGI and IMPC[54] have been provided through an API, in order to capture metadata necessary for re-use with preclinical imaging data. Disease metadata are associated with Disease Ontology (DO)[55] with a proper DO_ID code used also in the PIDAR

---

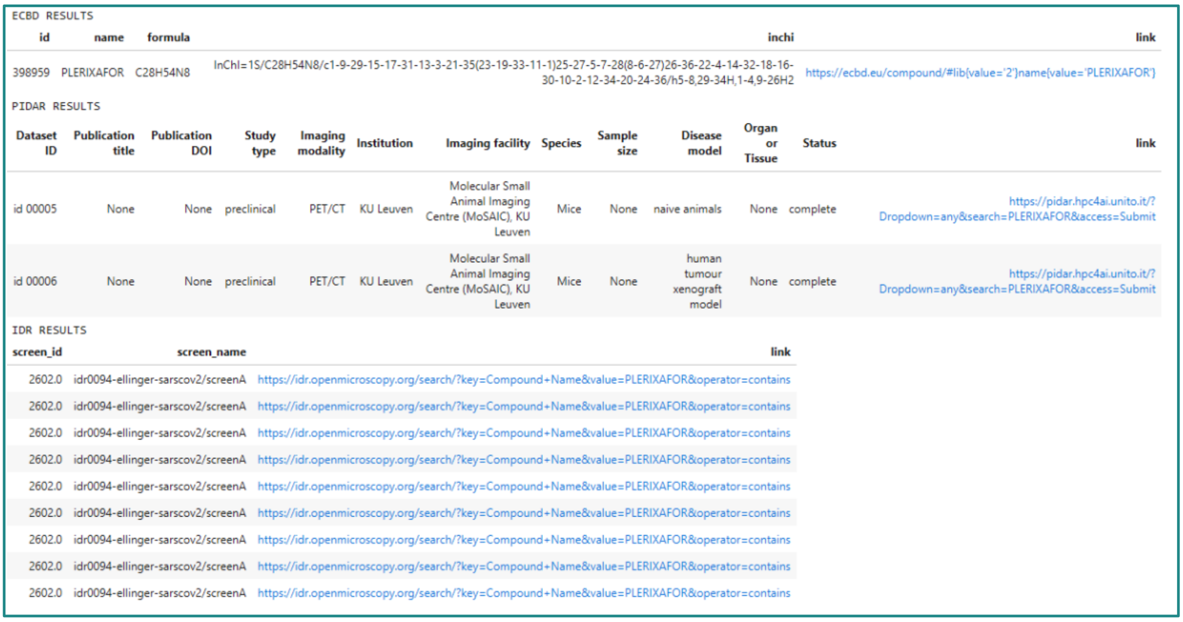
[46] https://www.infrafrontier.eu/emma/
[47] https://www.eu-openscreen.eu/
[48] https://www.eurobioimaging.eu/data/idr
[49] https://idr.openmicroscopy.org/about/
[50] https://pidar.hpc4ai.unito.it/
[51] https://pidar.hpc4ai.unito.it/
[52] https://www.ebi.ac.uk/chebi/
[53] https://www.informatics.jax.org/
[54] https://www.mousephenotype.org/
[55] https://disease-ontology.org/

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

repository, which makes the research univocal. Other main features as anatomical districts, imaging modality have been related to the NCI Thesaurus ontology[56] which contains an extensive number of terms, each one identified with a specific code.

The project team has developed a series of Jupyter notebooks demonstrating how to query and retrieve relevant data and metadata from several public repositories. The scripts are written in Python, but other programming languages could be used. The query code is also easy to download and shareable.

A visual representation of the detailed results (dataset, metadata, information, query results) of the workflow can be printed immediately after the query run. Examples are shown in *Figure 1* and *Figure 2*.



ECBD RESULTS

id	name	formula	inchi	link
398959	PLERIXAFOR	C28H54N8	InChI=1S/C28H54N8/c1-9-29-15-17-31-13-3-21-35(23-19-33-11-1)25-27-5-7-28(8-6-27)26-36-22-4-14-32-18-16-30-10-2-12-34-20-24-36/h5-8,29-34H,1-4,9-26H2	https://ecbd.eu/compound/#lib{value='2'}name{value='PLERIXAFOR'}

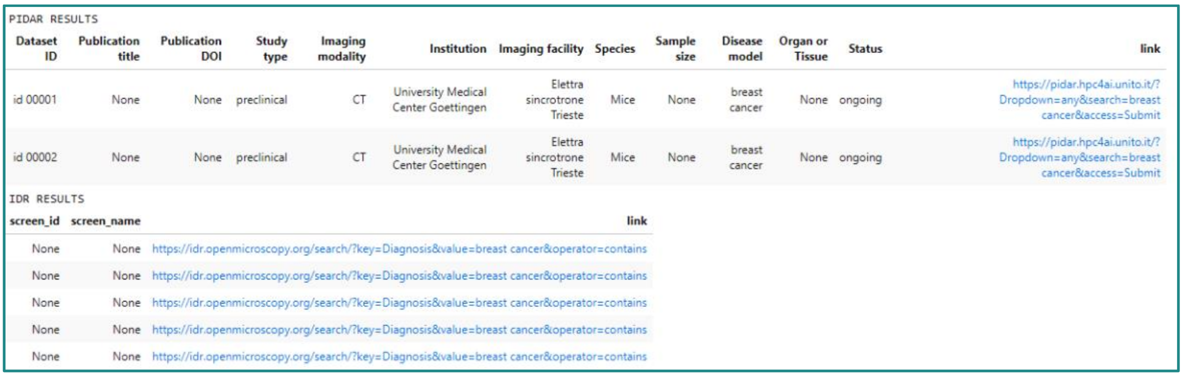
PIDAR RESULTS

Dataset ID	Publication title	Publication DOI	Study type	Imaging modality	Institution	Imaging facility	Species	Sample size	Disease model	Organ or Tissue	Status	link
id 00005	None	None	preclinical	PET/CT	KU Leuven	Molecular Small Animal Imaging Centre (MoSAIC), KU Leuven	Mice	None	naive animals	None	complete	https://pidar.hpc4ai.unito.it/?Dropdown=any&search=PLERIXAFOR&access=Submit
id 00006	None	None	preclinical	PET/CT	KU Leuven	Molecular Small Animal Imaging Centre (MoSAIC), KU Leuven	Mice	None	human tumour xenograft model	None	complete	https://pidar.hpc4ai.unito.it/?Dropdown=any&search=PLERIXAFOR&access=Submit

IDR RESULTS

screen_id	screen_name	link
2602.0	idr0094-ellinger-sarscov2/screenA	https://idr.openmicroscopy.org/search/?key=Compound+Name&value=PLERIXAFOR&operator=contains
2602.0	idr0094-ellinger-sarscov2/screenA	https://idr.openmicroscopy.org/search/?key=Compound+Name&value=PLERIXAFOR&operator=contains
2602.0	idr0094-ellinger-sarscov2/screenA	https://idr.openmicroscopy.org/search/?key=Compound+Name&value=PLERIXAFOR&operator=contains
2602.0	idr0094-ellinger-sarscov2/screenA	https://idr.openmicroscopy.org/search/?key=Compound+Name&value=PLERIXAFOR&operator=contains
2602.0	idr0094-ellinger-sarscov2/screenA	https://idr.openmicroscopy.org/search/?key=Compound+Name&value=PLERIXAFOR&operator=contains
2602.0	idr0094-ellinger-sarscov2/screenA	https://idr.openmicroscopy.org/search/?key=Compound+Name&value=PLERIXAFOR&operator=contains
2602.0	idr0094-ellinger-sarscov2/screenA	https://idr.openmicroscopy.org/search/?key=Compound+Name&value=PLERIXAFOR&operator=contains
2602.0	idr0094-ellinger-sarscov2/screenA	https://idr.openmicroscopy.org/search/?key=Compound+Name&value=PLERIXAFOR&operator=contains
2602.0	idr0094-ellinger-sarscov2/screenA	https://idr.openmicroscopy.org/search/?key=Compound+Name&value=PLERIXAFOR&operator=contains

PID 14176 - Figure 1: Example of cross queries between ECBD, PIDAR and IDR among datasets with common chemical compound PLERIXAFOR.



PIDAR RESULTS

Dataset ID	Publication title	Publication DOI	Study type	Imaging modality	Institution	Imaging facility	Species	Sample size	Disease model	Organ or Tissue	Status	link
id 00001	None	None	preclinical	CT	University Medical Center Goettingen	Elettra sincrotrone Trieste	Mice	None	breast cancer	None	ongoing	https://pidar.hpc4ai.unito.it/?Dropdown=any&search=breast cancer&access=Submit
id 00002	None	None	preclinical	CT	University Medical Center Goettingen	Elettra sincrotrone Trieste	Mice	None	breast cancer	None	ongoing	https://pidar.hpc4ai.unito.it/?Dropdown=any&search=breast cancer&access=Submit

IDR RESULTS

screen_id	screen_name	link
None	None	https://idr.openmicroscopy.org/search/?key=Diagnosis&value=breast cancer&operator=contains
None	None	https://idr.openmicroscopy.org/search/?key=Diagnosis&value=breast cancer&operator=contains
None	None	https://idr.openmicroscopy.org/search/?key=Diagnosis&value=breast cancer&operator=contains
None	None	https://idr.openmicroscopy.org/search/?key=Diagnosis&value=breast cancer&operator=contains
None	None	https://idr.openmicroscopy.org/search/?key=Diagnosis&value=breast cancer&operator=contains

PID 14176 - Figure 2: Example of cross queries between INFRAFRONTIER, PIDAR and IDR among datasets with common disease model: BREAST CANCER.

---

[56] https://ncithesaurus.nci.nih.gov/ncitbrowser/

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

The queries could be applied also to other common fields of research e.g. querying for a specific Strain, Species, Gene.

Due to time constraint, some resources exported the relevant metadata in a JSON file available for download. Other resources generate on-the-fly the JSON file containing the output of the query. An extension of this initial work will be to generate the JSON file on-the-fly for all repositories. Example notebooks have been deposited into the WorkflowHub[57] e.g. the Life Science cross-RI project[58].

SCIENTIFIC IMPACT

The developed tool allows flexible and easy-to-use cross-RIs queries that are of great interest to the whole scientific community within Life Sciences.

We expect the tool to be of great significance since it will benefit several communities associated with each RI involved: Euro-BioImaging[59], INFRAFRONTIER60, EU-OpenScreen[61]. This project fosters interoperability among these RIs to improve data curation, sharing of dataset and crosslink of curated databases. Moreover, the tool enables for the first time to search across the two resources of biological and preclinical images.

The project team has disseminated the project outcome by leaflets showing the results at the European Molecular Imaging Meeting in 2023 (the largest European conference in molecular imaging) and by organising a Data Management of Preclinical Image Datasets webinar[62] on June, 15th 2023. This webinar was addressed to the scientific community working in the imaging field, by providing both theoretical background and training activity about the tools we developed, but also regarding the principles of Open Science and FAIR, and the advantage of using ontology and standard tools to promote cross research. The webinar recordings are openly available on the Euro-BioImaging YouTube channel[63].

The project has also been presented in the EOSC-Life Final General Assembly Meeting, June 2023, among the EOSC-Life community.

ROLE AND IMPACT OF EOSC-Life

The project has been built on synergies and fruitful collaboration between the team members of EU-OpenScreen, INFRAFRONTIER, Image Data Resource (IDR), Euro-Bioimaging by taking into account the expertise within each infrastructure. In particular, within the WP3 the project exploited the expertise (WP1, WP2) on the ontology mapping needed to successfully build the cross-researching tool.

The strong links that were developed between Euro-Bioimaging and EOSC-Life pave the way for a stronger cooperation between RIs and European Research Institutes, making it possible to improve open data and open science clouds.

Regarding technical support, EOSC-Life was extremely useful in providing guidance and feedback with recognized experts during the whole development of the project. Also, attending training

---

[57] https://workflowhub.eu/

[58] https://workflowhub.eu/workflows/516?version=1

[59] https://www.eurobioimaging.eu/

[60] https://infrafrontier.eu/

[61] https://www.eu-openscreen.eu/

[62] https://www.cim.unito.it/website/events/educational/data_management.php

[63] https://www.youtube.com/watch?v=YdclTXLfzP0&list=PLW-oxncaXRqV0nsdKAiH6YL_gDHtbXJSE&index=1

events and workshops organised by the EOSC-Life project helped in the building of the needed knowledge to better develop this tool.

Taken together, the networking and cooperation among RIs and individuals have been really fruitful. Their technical support and resources, accessing EOSC-Life expertise, resources and funding have allowed the successful outcome of this project and shortened the time to carry out the projects.

4.2 Digital Life Sciences Open Call project - PID 14325

A workflow for marine Genomic Observatories data analysis

Project team: *Christina Pavloudi, Haris Zafeiropoulos, Stelios Ninidakis, Antonis Potirakis, Evangelos Pafilis, Cymon J. Cox, Gianluca De Moro, Robert Finn, Ekaterina Sakharova, Martin Beracochea, Ibon Cancio, Maria Luisa Chiusano, Erwan Corre, Katrina Exter, Nicolas Pade*

PROJECT DESCRIPTION

Genomic Observatories (GOs) are sites of long-term scientific study that undertake regular assessments of the genomic biodiversity of microbial, and often multicellular, organisms. The European Marine Omics Biodiversity Observation Network (EMO BON[64]) is a network of GOs that undertake regular environmental DNA (eDNA) sampling to generate environmental and metagenomic data of the microbial communities from designated marine stations around Europe.

To address the challenges of analysing the metagenomic data from EMO BON in a timely and standardised framework we developed metaGOflow[65]: an EMBL-EBI MGnify-derived computational workflow that implements the critical steps of a shotgun metagenomic bioinformatics analysis, and provides rich provenance metadata describing the data, data products, and workflow execution. Combining the Common Workflow Language (CWL) with the Research Object Crate (RO-Crate[66]) packaging, metaGOflow supports the fast inference of taxonomic profiles from GO-derived data based on rRNA genes and their functional annotation using its raw reads.

The novel aspects of this workflow are mainly:

a) the partial workflow execution, e.g. the user has the flexibility to choose whether to run the functional annotation sub-workflow or not, or running it at a later point using the data products of the previous steps;
b) the incorporation of an alternative assembler with a significantly lower computational cost and;
c) the ultimate generation and verification of a Research Object (RO) crate ensuring the workflow's FAIRness;

PROJECT RESULTS AND ACHIEVEMENTS

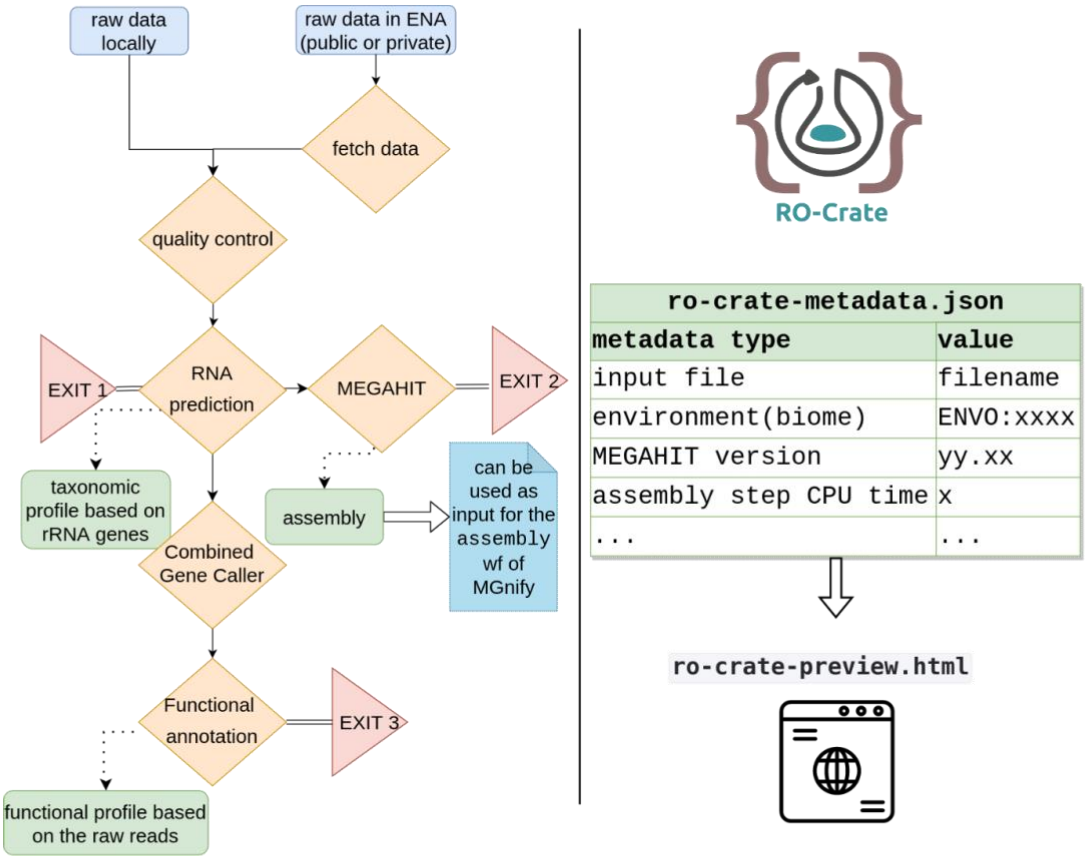
Within the project, we have managed to build metaGOflow (please see *Figure* 1), which is a user-friendly flexible workflow initially developed to address the specific needs of a GO project such as EMO BON, but which can also be used for one-sample-at-a-time analysis of shotgun

---

[64] https://www.embrc.eu/services/emo-bon
[65] https://data.emobon.embrc.eu/MetaGOflow/
[66] https://www.researchobject.org/ro-crate/

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

metagenomics data. There is a dedicated github page for metaGOflow[67] which is also registered in the WorkflowHub[68]. The workflow was built using the Common Workflow Language (CWL) and it can be implemented in Unix operating systems, using Docker or Singularity engines. A visual representation of the detailed results (quality control report, taxonomic inventories, functional annotations) of the workflow can be found in the metaGoflow github repository. An example of the complete data product of metaGOflow is available at Zenodo[69].



ro-crate-metadata.json	
metadata type	value
input file	filename
environment(biome)	ENVO:xxxx
MEGAHIT version	yy.xx
assembly step CPU time	x
...	...

PID 14325 - Figure 1: Schematic overview of metaGOflow

SCIENTIFIC IMPACT

For EMO BON and other GOs to produce applicable and fit-for-purpose data, it is of huge importance that quality-controlled and standardised data, as well as informative data products are made rapidly available. The disentanglement of the analyses from technical expertise and extensive computing infrastructures will allow the direct generation of meaningful data products even by non-experts. There is a paramount added value to the provision of preliminary results and data products (i.e., taxonomic inventories) from metagenomic GO samples as it can lead to

---

[67] https://github.com/emo-bon/MetaGOflow
[68] https://workflowhub.eu/workflows/384
[69] https://zenodo.org/record/7771821

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

the full exploitation of the data, to enhanced and timely decision-making and to successful environmental quality monitoring of the marine environment.

The workflow metaGOflow is an efficient and robust workflow that scales to the needs of projects producing big metagenomic data such as EMO BON. It highlights how containerization technologies along with modern workflow languages and metadata package approaches can support the needs of researchers when contending with ever-increasing volumes of biological data. Despite being initially oriented to address the needs of EMO BON, metaGOflow is a flexible and easy-to-use workflow that can be broadly used for one-sample-at-a-time analysis of shotgun metagenomics data.

There is also a draft publication that will be submitted shortly:
Zafeiropoulos H. et al., "metaGOflow: a workflow for the analysis of marine Genomic Observatories shotgun metagenomics data". In preparation

A workshop took place in July to highlight the technologies and the methodologies implemented in the framework of the project: Learning Data Management and Workflow Analyses: a hackathon from the EMBRC's EMO BON[70] (marine metagenomics) project.

ROLE AND IMPACT OF EOSC-Life

The strong links that were developed between EMBRC[71] and EOSC-Life paved the way for a new era of cooperation between RIs and European Research Institutes, making it possible to explore new frontiers in ecological science and support society in addressing the challenges ahead, while aiming towards open data and open science clouds. The metaGOflow project built on synergies with marine-related EU RIs (ELIXIR[72], EMBRC and LifeWatch-ERIC[73]) in order to take the next step towards the establishment of GO-related best practices, regarding analyses of metagenomic data, and create the long-term context for capacity building.

Among the main benefits of joining the EOSC-life project were, of course, the funding that gave the ability to implement the project and develop metaGOflow. Regarding technical support and networking, EOSC-Life succeeded in bringing together all the open call projects and there were numerous occasions through online meetings, where it was possible to interact both with the research team of other open call projects as well as with the EOSC-Life experts from different work packages, to receive guidance, feedback and consultation.

In addition, the data provenance sub-workflow of metaGOflow benefited greatly from being able to adopt RO-Crates and work with the RO-Crate community, which may not have been possible without EOSC Life.

---

[70] https://github.com/emo-bon/hackathon2022
[71] https://www.embrc.eu/
[72] https://elixir-europe.org/
[73] https://www.lifewatch.eu/

4.3 Digital Life Sciences Open Call project - PID 14308

PDB-REDO-cloud: A flexible and scalable engine for computational structural biology

Project team: *Robbie P. Joosten, Maarten L. Hekkelman, Hans Wienk, Anastassis Perrakis*

PROJECT DESCRIPTION

Biological and biomedical research depend strongly on detailed insight in the interactions of proteins with other proteins, nucleic acids and small molecules such as cofactors and ligands. Mechanistic knowledge of protein function and interactions commonly requires atomic detail in 3 dimensions that is provided by molecular structure models. These models are derived through experimental means in which X-ray crystallography is currently the largest source of experimental data. Particularly in fragment-based drug discovery, X-ray crystallography is now performed in high-throughput leading to hundreds or even over a thousand datasets to be analysed in a single fragment screen. This requires a high level of automation in turning this X-ray data into ready-to-use structure models.

The appearance of powerful methods (e.g. AlphaFold[74]) to predict protein structures has extended the use of 3D structure models beyond the traditional structural biology community and has increased the demand for experimental verification of predicted structure models in an audience that has relatively little experience in dealing with crystallographic data. This provides both opportunities and challenges to method developers in structural biology.

The PDB-REDO[75] software pipeline for crystallographic structure optimisation addresses both issues above by being a fully automated expert system that tries to emulate what an experienced crystallographer does without the need for manual intervention. This works well for non-expert users as well as in settings where many datasets need to be processed.

This has been shown on the one hand by the structural biologists using PDB-REDO and, on the other hand, by the PDB-REDO developers who keep a data bank of over 160 thousand ‘redone’ Protein Data Bank entries.

At the start of the project, the PDB-REDO redo pipeline was available in two ways: as a webserver and as installable software. Unfortunately, this setup had serious practical limitations. The webserver is purposely designed to be as simple as possible which means that advanced options for the pipeline are not exposed to users which limits users in setting up more tailored calculations including performing experiments beyond merely optimising structure models. Local installation of the PDB-REDO pipeline is a solution to this limitation, however local deployment of the software has proven quite challenging and, if successful, users may have limited computational resources to run high throughput experiments. Additionally, PDB-REDO is in continuous development which means that the software would need to be updated frequently which is an additional challenge in a local computing environment.

This Open Call project aimed at creating PDB-REDO-cloud, a high throughput and flexible interface to the PDB-REDO pipeline that can be run from the EOSC-Life environment and elsewhere. Additionally, proof-of-concept implementations in third party workflow are provided and a refresh of the PDB-REDO data bank has been initiated.

---

[74] https://alphafold.ebi.ac.uk/

[75] https://pdb-redo.eu/

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

PROJECT RESULTS AND ACHIEVEMENTS

Over the course of the project many advances were made to ultimately provide the new PDB-REDO-cloud interface.

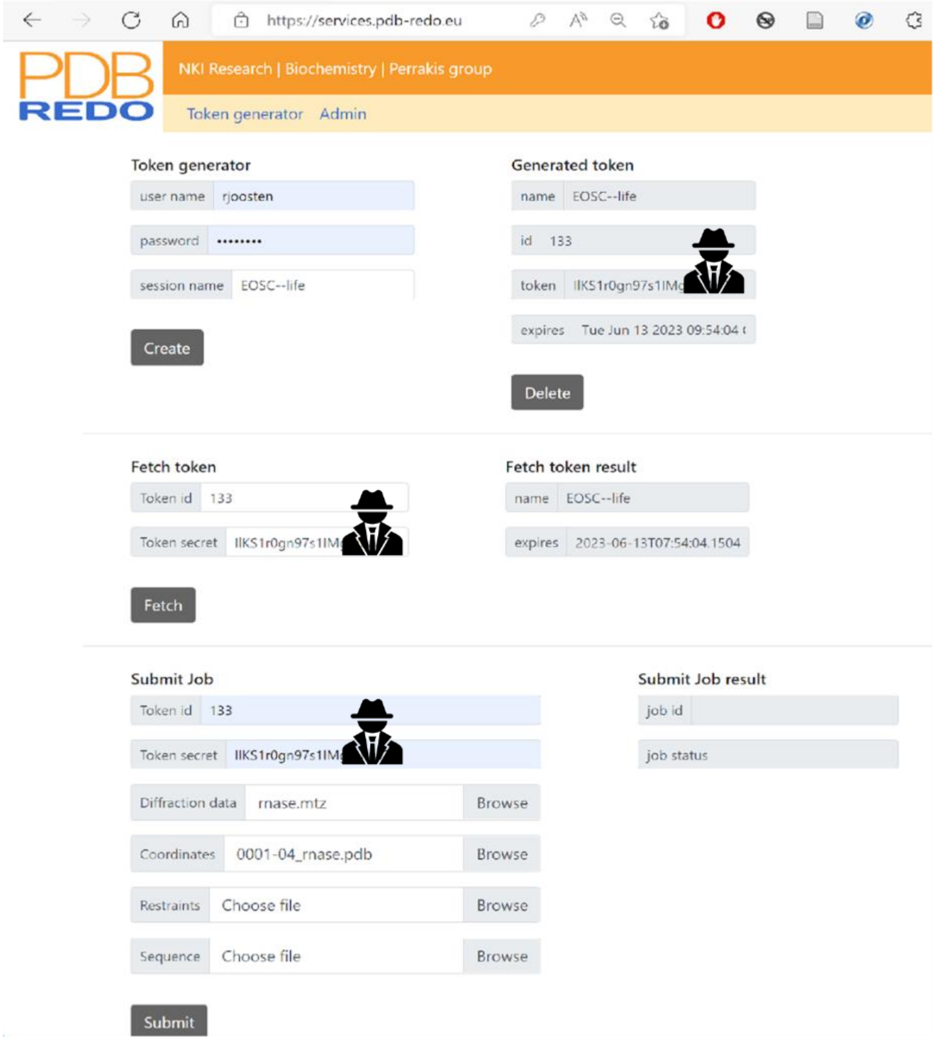
Many elements of the PDB-REDO pipeline were updated to improve deployability in the PDB-REDO-cloud backend. This also entailed the release of our software under a BSD-2-clause licence which permits integration of the code in other projects. The code is available at on GitHub[76] and includes other PDB-REDO related software such as AlphaFill and DSSP.

Because of its long development history, PDB-REDO can be described as consisting of four parts: the core pipeline, the modern libcifpp-based tools, the libwARP-based rebuilding tools, and the libhomol-based tools. The first two have already been released in the new scheme, the last two are in the process of being fully rewritten with new scientific developments and will be released under the same scheme in due course.

The main PDB-REDO script was rewritten to take commands as a single JSON file. This step was important in setting up a PDB-REDO-cloud API in which users have access to all PDB-REDO settings that they would otherwise only have on a local installation. This includes experimental features such as automated fragment fitting. The PDB-REDO output data.json file was updated to explicitly document the use of non-default settings to ensure that users can repeat calculations with the same settings even if they lost their JSON input command files.

---

[76] https://github.com/PDB-REDO



Token generator

user name: rjoosten
password: ••••••••
session name: EOSC--life

Generated token

name: EOSC--life
id: 133 [redacted]
token: lIKS1r0gn97s1IM[redacted]
expires: Tue Jun 13 2023 09:54:04 (

Fetch token

Token id: 133 [redacted]
Token secret: lIKS1r0gn97s1IM[redacted]

Fetch token result

name: EOSC--life
expires: 2023-06-13T07:54:04.1504

Submit Job

Token id: 133 [redacted]
Token secret: lIKS1r0gn97s1IM[redacted]
Diffraction data: rnase.mtz
Coordinates: 0001-04_rnase.pdb
Restraints: Choose file
Sequence: Choose file

Submit Job result

job id
job status

PID 14308 - Figure 1: Testing interface for PDB-REDO session manager

The documentation of the PDB-REDO output was improved to increase the FAIR-ness of the output data. Where needed the data itself was also updated. This was in line with our concurrent WP1 project that focussed on the PDB-REDO data bank. The new documentation, including JSON schemas is available from: https://pdb-redo.eu/download.

An initial session manager for the PDB-REDO-cloud API was created. This session manager allowed users to create and manipulate session tokens and setup calculations. These tokens are used to identify users and sign/encrypt API commands. A visual interface (*Figure* 1) was provided, mostly for testing.

After consulting with testing users and 3rd parties interested in integrating PDB-REDO-cloud in their workflows, the session manager and API were updated to achieve their final form. In addition, the session manager was extended to integrate the PDB-REDO website that hosts the webserver and the PDB-REDO data bank in order to create a uniform look and feel. The PDB-REDO archive, the main result of our WP1 project, is also integrated. The current code of the session-manager is available from GitHub[77], and the PDB-REDO website that hosts PDB-REDO-cloud, the webserver, the data bank and the archive[78]. The API documentation for PDB-REDO-cloud is available online as well[79].

With PDB-REDO-cloud in place the pipeline could be integrated in other workflows. The integration in crystallographic fragment screening workflows in the context of the collaborative Horizon-2020 projects iNEXT-Discovery[80] and FragmentScreen[81] is ongoing. PDB-REDO-cloud was successfully integrated in two workflow managers of the CCP4[82] project, CCP4i2 and CCP4-cloud, both of which are available in the current release of the CCP4 software suite.

A refresh of the PDB-REDO data bank was started focussing on the structure models containing nucleic acids to accommodate the research that is done in the context of the Nucleic Acid Valence Geometry Working Group[83].

SCIENTIFIC IMPACT

The current PDB-REDO-cloud is already making an impact in the structural biology community as is demonstrated by some 1500 calculations performed in 2022, nearly 20% of the total calculations. This is expected to rise significantly with the CCP4 integration released at the start of 2023 and the future integration in iNEXT-Discovery and FragmentScreen workflows. Most calculations performed with PDB-REDO-cloud were part of 3rd party research projects on molecular dynamics for model optimisation and a new method to model alternative side chain conformations. The latter is currently being written up for publication.

- The new PDB-REDO-cloud integration in CCP4 was presented and tested at the CCP4/APS School in Macromolecular Crystallography[84]
- The integration in CCP4 workflow managers is discussed in Agirre et al.[85] which is published in Acta Crystallographica D.

ROLE AND IMPACT OF EOSC-Life

The working experience with EOSC-life was very positive according to the project team. The feedback during the application process was much appreciated as the team was new to the way of working in EOSC. Also, the EOSC-life team was very forthcoming with support where this was needed. The team also benefited from its participation in the EOSC-Life FAIR Hackathon[86].

---

[77] https://github.com/PDB-REDO/pdb-redo-session-manager
[78] https://pdb-redo.eu/
[79] https://pdb-redo.eu/api-doc
[80] https://inext-discovery.eu/
[81] https://fragmentscreen.org
[82] https://www.ccp4.ac.uk/
[83] https://www.iucr.org/news/newsletter/etc/articles?issue=150473&result_138339_result_page=16
[84] https://www.ccp4.ac.uk/schools/APS-2023/
[85] https://doi.org/10.1107/S2059798323003595
[86] https://www.eosc-life.eu/services/training/

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

4.4 Digital Life Sciences Open Call project - PID 14157

Expression Atlas' RNA-Seq and Microarray analysis pipelines migration to workflow environments for cloud deployment and reproducibility

Project team: Pedro Madrigal, Pablo Moreno, Andrey Solovyev, Jonathan Manning, Irene Papatheodorou

PROJECT DESCRIPTION

The EMBL-EBI Expression Atlas[87] serves aggregated, standard re-analysed and curated data for thousands of RNA-Sequencing, microarray and proteomics datasets to ~600 different users daily. Developed during the last 13 years or more, passing through different technical teams, accumulating technical debt due to the funding model. Thus, our pipelines followed old paradigms: scalability issues; lack of portability; did not rely on a workflow environment system, presented proprietary reference layer management and no modern dependency management, etc. As a result, our pipelines were not easily reproducible nor maintainable. This also made the long-term sustainability of the resource vulnerable in technical terms to the departure of key personnel, and to become obsolete.

The main aims of the project were to have for both RNA-Seq and Microarray analysis pipelines one modern workflow environment system that could be portable and easily cloud deployed, with proper dependency resolution. This initiative has allowed Expression Atlas pipelines to shift from being strongly dependent on the EBI cluster and shared file system to a modern, community maintained, explicit workflow environment that can run outside of the EBI infrastructure. More importantly, the migration of these analysis pipelines facilitates the re-use and re-analysis of RNA sequences and microarray data by third parties.

PROJECT RESULTS AND ACHIEVEMENTS

One main achievement through this project has been to refactor RNA-Seq, microarrays and proteomics Expression Atlas analyses into the Snakemake workflow manager to be able to process, reprocess and perform Ensembl updates of existing studies:

- Snakemake workflow for Microarrays
- GTEX workflow for analysis of controlled access data
- Snakemake workflow for Proteomics
- Workflow features to enable deconvolution of RNA-seq samples in EA
- https://github.com/ebi-gene-expression-group/bulk-recalculations

The project team also worked on the Evaluation of existing reference genome managers, and on the setup and evaluation of RefGenie[88] as a reference layer. This workflow is designed to copy all the base genome resources used by the gene expression group to new Refgenie assets.

Important work was carried out on RNA-Seq quantification & QC:

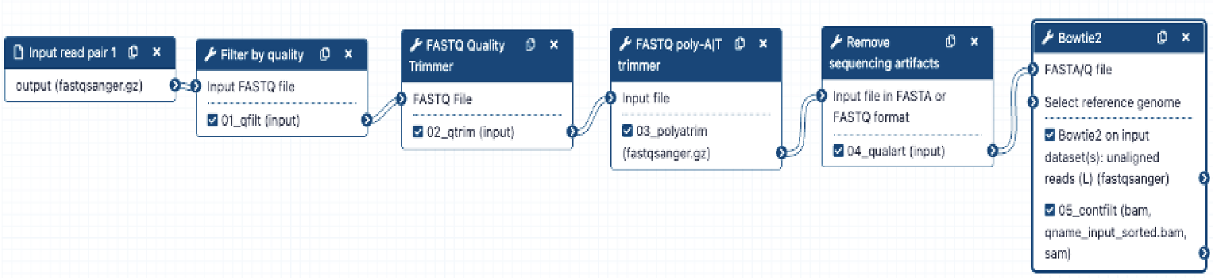
- Initial gathering of dependency requirements from legacy pipelines for RNA-Sequencing
- Software used on legacy systems partly wrapped for Galaxy and Bioconda
- Initial Proof of concept (PoC) pipeline without connection yet to a reference layer (see *Figure 1*)

---

[87] https://www.ebi.ac.uk/gxa/home

[88] http://refgenie.databio.org/en/latest/

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



PID 14157 - Figure 1: Proof of concept pipeline for RNA sequencing

The team also explored some existing Galaxy solutions for Microarrays, but they might require some adaptation for metadata. The team has contacted authors but with no answer. They also gathered dependencies in terms of Conda packages for microarrays and for expression matrix downstream.

Finally, the project team worked on Galaxy workflow[89] Proof of Concept for RNA-Seq IRAP.

SCIENTIFIC IMPACT

Because of the workflow modernisation achieved, the transition of the original pipelines from the LSF cluster to others (e.g., SLURM) is straightforward, not only facilitating deployment in HPC and cloud environments, but also smooth release cycles and the continuation of the service in the future.

Migration to a more modern, community maintained, explicit workflow environment enables a faster turn-over in terms of variations to the analysis workflows, de-couple it from the main EBI HPC, execution on multiple environments, granular tool updates and in general make the workflow more maintainable. An externally maintained workflow environment has a community, code maintenance, training and documentation behind it, facilitating the transfer of expertise should key personnel change. A move to such a workflow environment also facilitates scrutiny and re-utilisation by the external community. The joint maintenance of tool wrappers and dependencies by both Expression Atlas and the community would benefit both with improved handling of gene expression tools.

The project did not include the generation of data as part of the project. However, the migration of the Expression Atlas RNA-Seq, Microarray and Proteomics analysis pipelines to the POCs achieved here will generate data that will be FAIR, as all data currently available on Expression Atlas. The POCs workflows and tools generated during this project have been deposited in public repositories (Galaxy Toolshed, WorkflowHub.eu, Github repositories), which should guarantee the maximum possible level of FAIRness for workflows themselves. Moving the workflows to the new setup, every effort will be made to make sure that intermediate and final outputs are standard interoperable formats that can be easily reused by the community, and avoiding as much as possible any proprietary file layout that would make this difficult.

Most of the dissemination has been done through in-person EBI training at the Genome Campus, with multiple members of our group being involved as instructors.

---

[89] https://training.galaxyproject.org/training-material/topics/galaxy-interface/tutorials/workflow-editor/tutorial.html

- Introduction to RNA-seq and functional interpretation course[90].
- Single-cell RNA-seq analysis using Galaxy course[91].

In addition, an online tutorial on Workflows for Combining tools for data analysis was developed as part of the EOSC-Life project at EMBL-EBI. This tutorial provides a short introduction to workflows for the analysis of life sciences data, providing examples of platforms for creating workflows and the considerations required for developing them.

Team members have joined activities organised by the Workflows Community at the Wellcome Genome Campus in Hinxton, UK, that were mostly oriented to Nextflow and Snakemake use cases and training.

ROLE AND IMPACT OF EOSC-Life

For the past decade, our resource, EMBL-EBI Expression Atlas (EA), has been at the forefront of gene expression data re-analysis, providing tens of millions of users a year the ability to access re-analysed gene expression data for thousands of selected datasets.

The funding, training and support network allowed the team to keep up with this demand for the next 5 to 10 years, as our setup needs now embrace modern ideas in bioinformatics pipeline development such as the use of explicit workflow environments, automatic dependency resolution, containerisation, cloud computing and federated workflow deployments, among others. It is crucial that analysis pipelines used in EA can easily be disseminated with the research community to support reproducible analyses more widely, in different set ups, species and experimental conditions.

4.5 Digital Life Sciences Open Call project - PID 13557

Increasing the FAIRness of Phytolith Data

Project team: Emma Karoune, Celine Kerfant, Carla Lancelotti, Juan José García-Granero, Marco Madella, Javier Ruiz-Pérez

PROJECT DESCRIPTION

The project aimed to increase the knowledge of and the use of the FAIR data principles in phytolith research to improve communication of methods, data sharing and archiving practices within the discipline. Phytoliths are silica bodies that are deposited in or between plant cells during the life-cycle of the plant. They are used in different scientific fields such as archaeology, palaeoecology and plant sciences to address questions of past plant exploitation and long-term environmental and biodiversity changes.

In this project, the project team conducted a community survey to find out about current data sharing and opinions on the use of open research practices. Furthermore, the project looked at published phytolith research and assessed the data and metadata within them in terms of the FAIR data principles.

Using these two new datasets, the Phytolith team has drawn together FAIR recommendations for the phytolith community that will then be reviewed and adapted by the community itself, to

---

[90] https://www.ebi.ac.uk/training/materials/introduction-to-rna-seq-materials/

[91] https://www.ebi.ac.uk/training/materials/single-cell-rna-seq-analysis-using-galaxy-materials/

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

produce FAIR phytolith guidelines. They are currently running a series of training workshops to upskill their community in open research skills including standardised vocabularies and FAIR data.

Future plans include the creation of a phytolith ontology to aid interoperability of phytolith data and an online open repository for phytolith data. More information can be found at the FAIR - phytolith Github page[92].

PROJECT RESULTS AND ACHIEVEMENTS

The project team has conducted a community survey to collect information about current data sharing and opinions on open research practices. This survey had 81 responses and the results from this survey have been written as an article and submitted to Open Quaternary (currently under review).

The team conducted a FAIR assessment of existing phytolith studies by looking at 100 published articles. They collected data on the use of open access publishing, communication of data collection methods, use of standard nomenclatures, communication of identifications through photos, and how data was published and archived. A data paper was written for this FAIR assessment dataset and the article has been published in *Nature Scientific Data* (*Kerfant et al.*[93]).

The project team has also combined the findings of their survey and FAIR assessment in a third article that presents the results of the FAIR assessment and offers recommendations for producing FAIR phytolith data. These recommendations will then be discussed with the community to establish FAIR guidelines for phytolith data.

The Phytolith community building work started by establishing the International Committee on Open Phytolith Science (ICOPS) within the International Phytolith Society. This has brought together a wider global group of 12 researchers to work on different initiatives in the movement towards open research in our discipline. So far, they have worked together to produce the first version of Open Phytolith Publishing guidelines[94].

The FAIR Phytoliths team has given 13 talks about this project at international conferences, meetings and at organisation events including the International Meeting of Phytolith Research, 2nd Virtual conference for Women Archaeologists and Palaeontologists, Computational Applications in Archaeology International Conference, European Association of Archaeology Annual Meeting and the EOSC-Life Annual General Meeting.

The project team also focused on training the Phytolith community in open research skills. In 2021 and 2022 they ran training workshops for the Association of Environmental Archaeology on Open Science skills (Department of Archaeology, Durham University) on reproducibility, and three training sessions on GitHub through ICOPS. All training materials are accessible and archived on Zenodo.

At the end of 2022, ICOPS was awarded an EOSC-Life open training grant. This has allowed the project team to run a series of training workshops on open research skills for our community but also the wider research community, as the workshops are open for all to attend. There are six workshops on: open access, using repositories, GitHub basics, GitHub collaboration, standard

---

[92] https://open-phytoliths.github.io/FAIR-phytoliths/

[93] https://www.nature.com/articles/s41597-023-02296-8

[94] https://zenodo.org/record/6803489

vocabularies and ontologies and FAIR data, running from January to June 2023. All training materials will be archived and made into online self-study courses.

SCIENTIFIC IMPACT

The creation of a committee on open science within our international society is a big step forward in terms of the acceptance of this approach in our discipline. The formation of the committee itself meant that the project team went from a group of six researchers in the FAIR Phytoliths project to a larger group of 12 researchers in the committee. This has expanded the reach of their work in several ways: an increase in the number of institutions involved, an increase in the number of countries and world regions involved (1 to 5 regions including Europe, North America, South America, Asia, Africa) and an increase in the different sub-disciplines represented in our group. This allows the team to progress their work in a number of different ways including taking into account local differences in the ability of researchers to implement open science practices and the ability to provide the project outputs in multiple languages to make them more accessible to the Phytoliths community. The project team has started this dissemination on their Open-Phytoliths Community website[95], which is available in English, French, Spanish and Mandarin.

As a committee the project team worked together to produce a guide to open publishing for phytoliths to be used with a special issue in a journal. Although it was not strictly enforced by the editors, the team did see an increase in the openness of the articles in this special issue and some were published in a very high-quality manner in terms of openness and transparency.

The project team has been receiving positive feedback after talks they have given and also when providing training courses. This is producing more awareness of their work generally and attendance at their training workshops in 2023[96] (70 attendees in March 2023) has been high compared to the initial training sessions. The team has also been providing Spanish and Chinese translation at their workshops and this has widened the impact of this training series to researchers who would not have normally attended due to the inaccessibility of training just in English.

As a result of this greater awareness of the need for FAIR data, the project team has been asked to advise on several projects in related disciplines such as the Rewilding later prehistory project[97] and the Archaeobotany Work Group[98]. In addition, the team has been asked to provide a training session during the next International Meeting of Phytolith Research (IMPR)[99] that will take place in Israel in September 2023.

The project team established a wider committee (ICOPS) in the International Phytolith Society and the work of this project is being taken on by this committee. They meet bi-monthly to discuss different aspects and work on different initiatives such as the phytolith ontology and training workshops.

In terms of scientific career paths, a postdoc employed in FAIR project, has obtained a two-year Juan de la Cierva grant (Spanish Ministry of Science and Innovation) to continue aspects of the FAIR project work along with her own phytolith research.

---

[95] https://open-phytoliths.netlify.app/

[96] https://open-phytoliths.netlify.app/course/open-res-workshop/

[97] https://rewilding.oxfordarchaeology.com/

[98] https://historicengland.org.uk/research/current/heritage-science/archaeobotanical-work-group/

[99] https://conference.adssc.org/

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

ROLE AND IMPACT OF EOSC-Life

The EOSC-Life funding has paid for staff time to work on the project meaning that this work has been driven forward at a quicker pace than would have initially happened. According to the project team, the network created by EOSC-Life has enabled faster progress for the project than expected, and they have benefitted from having experts to advise them on different aspects of the project and the project progressed quickly because of this.

Within the fruitful framework, the project team started to work on a phytolith ontology, which is the first for this discipline. Initially the team was not considering starting this as part of their EOSC-Life project but during the EOSC-Life FAIR hackathon[100] in 2021/2022 they learned more about FAIR resources, tools, ontologies and got connected with an ontology expert from EOSC-Life WP1 (Henriette Harmse). The EOSC-Life expert has been training the team in the ontology software and working with the team to develop the ontology for Phytoliths and using existing Ontology tools developed by EMBL-EBI[101]. This collaboration has moved the project team much further along in terms of addressing interoperability issues by the end of the project.

The team has been collaborating with EOSC-Life WP1 experts (Henriette and Jean-Marie Burel) to provide training around ontologies and FAIR data for our community. There have been two sessions in May and June 2023 with talks and demonstrations from these collaborators. The project team has been also using the FAIR Cookbook[102] to develop data dictionaries for their new datasets and they will be adding a recipe into the FAIR Cookbook from their training in using repositories.

4.6 Digital Life Sciences Open Call project - PID 13840

Reference Data Resource

Project team: Ignacio Eguinoa, Frederik Coppens, Björn Grüning

PROJECT DESCRIPTION

This project aimed at deploying a cloud-based service containing reference genome resources using the Refgenie manager[103].

Refgenie is a reference genome resource manager that we have been testing for a while in different contexts and has already been adopted in WfMS such as Galaxy and Nextflow. It's highly flexible both in terms of the methods it provides for accessing the contained data and the possibilities of extending it to accommodate diverse genome derived resources. This makes it a perfect fit to deploy a shared resource that can accommodate the variety of reference data that the community uses.

This deployment of a Refgenie service in a cloud-based platform involved work related with curating data from new independent assembly and annotation projects, including prebuilt indexes and other reformatted assets, and documenting the processes to make it a community-built resource. But the core idea was taking a big step in the (ongoing) FAIRification process:

---

[100] https://www.eosc-life.eu/services/training/

[101] https://www.ebi.ac.uk/ols4/index

[102] https://faircookbook.elixir-europe.org/content/home.html

[103] http://refgenie.databio.org/en/latest/

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

- all assets associated with a genome build should be linked together under a namespace identifying the genome build;
- each reference dataset (fasta files, annotations, indexes, etc) should contain associated metadata with provenance information of how these were obtained as well as file digests. The method used to generate each asset is also detailed which would enable reproducing the resource construction.
- This kind of resource also allows for improved accessibility as it provides several ways of fetching the datasets: command line, web interface and shared (read-only).

PROJECT RESULTS AND ACHIEVEMENTS

The first achievement was the expansion of the starting Refgenie instance, containing a set of previously curated plant reference genomes, also available in the PLAZA[104] data warehouse.

The project team had already created a pilot containing several plant reference genomes files with the help of Refgenie developers which provided us with cloud resources.

This worked as a well evolved proof of concept to provide a service containing a web-accessible entry point[105]. All deployment details are publicly available since the beginning of this pilot deployment[106]. This includes both technical details as well as documentation about how to contribute with reference data to an already running instance.

SCIENTIFIC IMPACT

The main impact of this project is in the FAIRification of reference data in the context of workflow management systems, and platforms like Galaxy[107] aimed at simplifying the data analysis process but which require post processing and sharing of results outside of it, including references to genomic and other derived reference data. The deployment of these resources became considerably easier with the usage of a network of refgenie instances containing community wide used genomic data.

The project results were presented in the last Galaxy Community Conference and were added to a publication in relation with Galaxy management of reference data (VijayKrishna et al.[108])

ROLE AND IMPACT OF EOSC-Life

According to the project team, the best benefit working with EOSC-Life has been to have a constant follow-up for advances throughout the project and receiving help from the multiple EOSC-Life technical work packages to solve different issues arising in the project.

Being part of EOSC-Life has also been important in terms of networking. Indeed, EOSC-Life encouraged the link between projects teams that could benefit from each other for example by allowing the team to attend each other's progress reports, etc.

---

[104] https://bioinformatics.psb.ugent.be/plaza/
[105] http://plantref.databio.org/
[106] https://github.com/refgenie/plantref
[107] https://galaxyproject.org/
[108] https://doi.org/10.1093/bioadv/vbac030

4.7 Digital Life Sciences Open Call project - PID 13888

Open Source Secure Data Infrastructure and Processes for Life Sciences (OSSDIP4LIFE)

Project team: *Martin Weise, Tobias Grantner, Geoffrey Karnbach, Martin Krajiczek, Dietmar Winkler, Tomasz Miksa, Niki Popper, Andreas Rauber*

PROJECT DESCRIPTION

Secure sharing of sensitive data from the data owner to research experts is extremely challenging due to privacy reasons, but also due to the massive risk involved in sharing commercially sensitive data. To address this issue, TU Wien developed a fully open-source based high-security data visiting platform called OSSDIP and is with this EOSC-Life-funded project working on adapting it to the needs of selected data owners in the life sciences, extending its functionality to support more flexible data analytics. This reference implementation allows data owners to provide: 1) highly selective access via data visiting, 2) to specific (fine-granular or aggregated) subsets of data; 3) for identified individuals 4) for limited periods of time; 5) to answer precisely defined questions accepted by the data owner.

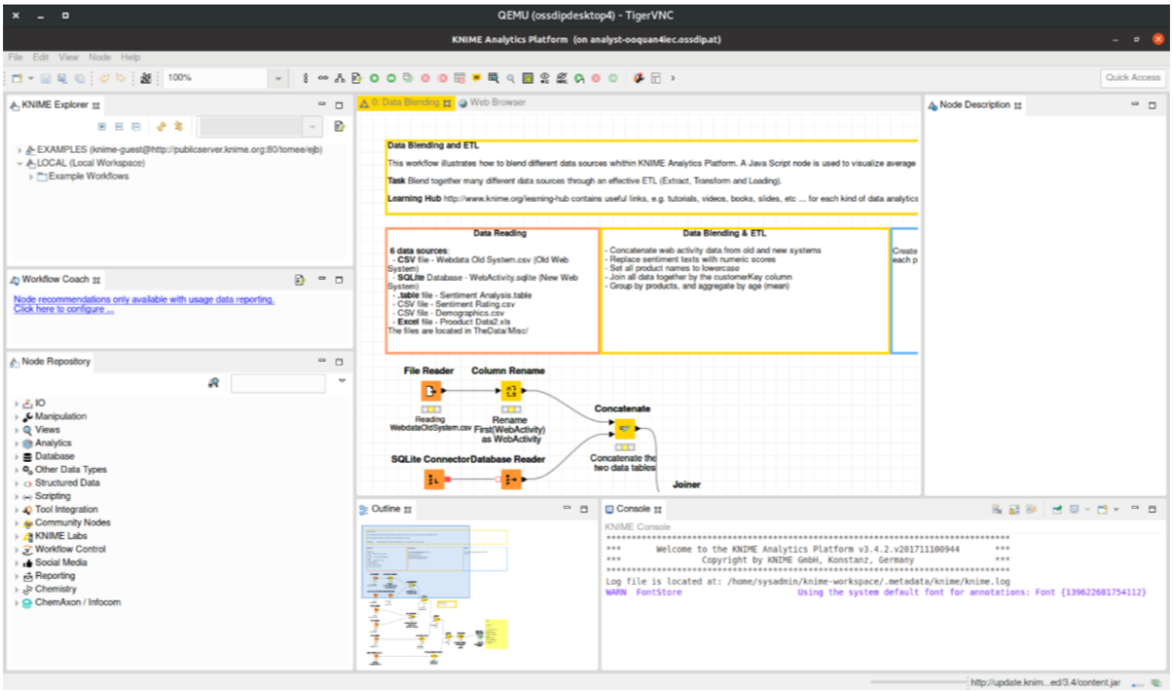
The first prototype of OSSDIP was assembled at the start of the COVID-19 pandemic, where researchers could no longer visit sensitive data physically due to a strict national lockdown without initial foreseen lift of restrictions. In EOSC-Life, the goal was to further extend the quickly put-together implementation, refactor parts where necessary and pave the way for adopters that want to set-up such an infrastructure through extensive documentation while having a demonstration instance.

PROJECT RESULTS AND ACHIEVEMENTS

During the EOSC-Life project, the project team has achieved the following objectives:

1. Automating processes for provisioning of the infrastructure and data ingress and -access processes. The overall system was described in a journal article and disseminated early 2022.
2. Definition and proof-of-concept implementation of a secure software review process to inspect package contents from the software vendor for vulnerabilities and backdoor (calling home) channels.
3. The monitoring was improved, the logging service can be configured to e.g., reduce the log volume or history. The video stream from the Desktop-VM (where the researcher is provided with a full desktop) is now recorded and can be inspected by the data owner at any time.
4. The support for analytics and workflow management tools have been greatly extended, now covering RStudio, KNIME, Jupyter and Galaxy.
5. Documentation of the system components in our Gitlab code repository that is open for collaboration and makes the reference implementation available for reuse to anyone via an Apache 2 open-source license.

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



Data Blending and ETL

This workflow illustrates how to blend different data sources whithin KNIME Analytics Platform. A Java Script node is used to visualize average

Task Blend together many different data sources through an effective ETL (Extract, Transform and Loading).

Learning Hub http://www.knime.org/learning-hub contains useful links, e.g. tutorials, videos, books, slides, etc ... for each kind of data analytics

Data Reading

6 data sources:
- **CSV** file - Webdata Old System.csv (Old Web System)
- **SQLite** Database - WebActivity.sqlite (New Web System)
- **.table** file - Sentiment Analysis.table
- CSV file - Sentiment Rating.csv
- CSV file - Demographics.csv
- **Excel** file - Prooduct Data2.xls

The files are located in TheData/Misc/

Data Blending & ETL

- Concatenate web activity data from old and new systems
- Replace sentiment texts with numeric scores
- Set all product names to lowercase
- Join all data together by the customerKey column
- Group by products, and aggregate by age (mean)

KNIME Console

```
***************************************************************************
***       Welcome to the KNIME Analytics Platform v3.4.2.v201711100944      ***
***                Copyright by KNIME GmbH, Konstanz, Germany               ***
***************************************************************************
Log file is located at: /home/sysadmin/knime-workspace/.metadata/knime/knime.log
WARN  FontStore                          Using the system default font for annotations: Font {139622681754112}
```

http://update.knim...ed/3.4/content.jar

PID 13888 - Figure 1: Analysis tool KNIME running in a secure Analyst-VM that is connected (visually only) to the desktop environment running at the Remote Desktop-VM.

SCIENTIFIC IMPACT

The project has impacted the area by providing a rapid test-deployment for organizations that want to adopt the methodology of data visiting.

The developed code is hosted in the OSSDIP Gitlab[109] code repository along with code documentation. The OSSDIP project website[110] contains information on how-to set up, configure and use OSSDIP. Several articles have been written by the project team (e.g. Weise M. et al.[111]), and the project outcomes disseminated at workshops and conferences (e.g. the EGI Conference[112], the BioImaging and the European Open Science Cloud - Workshop[113], the 20th RDA Plenary[114]). Future project plans aim at hardware proof-of-concept, penetration tests, etc.

ROLE AND IMPACT OF EOSC-Life

According to the project team, the main benefits of joining the EOSC-Life project were the plethora of activities and opportunities to interact and exchange with other project partners and experts to the network. Interesting cross-disciplinary discussions and exchange evolved during the course of the project on topics such as: a) the management of sensitive data and collaborations on sensitive data with researchers, b) what tools and approaches they use, so that the project team can make them available in the reference implementation of OSSDIP. The team presented

---

[109] https://gitlab.tuwien.ac.at/martin.weise/ossdip
[110] https://www.ifs.tuwien.ac.at/infrastructures/ossdip/
[111] http://doi.org/10.5334/dsj-2022-004
[112] https://zenodo.org/record/7265064
[113] https://repositum.tuwien.ac.at/handle/20.500.12708/176772
[114] https://www.rd-alliance.org/trusted-research-environments-sensitive-data-fairness-closed-data-and-processes

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

these implementations to project partners and received positive impressions and feedback, although a production-grade implementation will likely require more effort and adaptation.

4.8 Digital Life Sciences Open Call project - PID 14366

Towards FAIR data for X ray-based structure-guided drug design

Project team: Jose A. Marquez, Sameer Velankar

PROJECT DESCRIPTION

Macromolecular crystallography (MX) is commonly used for the structural analysis of biological molecules, with direct impact for scientific communities on all domains of life sciences, chemistry and chemical biology as well as in drug design and biotechnology applications. Crystallographic models are deposited in the open access Protein Data Bank which is managed by the wwPDB consortium. The deposition includes structure models along with a limited set of metadata, which is manually collected often months to years after the initial experiments took place. Recent technological developments at Instruct-ERIC[115] facilities for X Ray-based ligand and fragment screening have enabled experimental screening of very large chemical libraries generating large amounts of data. This data is of key relevance for drug design. However, manual data deposition in this context is very challenging and the current level of FAIRification in the field is limited. This represents a barrier for data reuse and particularly for the development of new approaches for drug design based on AI. The goal of this action is to develop tools to contribute to improve the level of FAIRness in MX.

PROJECT RESULTS AND ACHIEVEMENTS

The project team has developed new, standardised metadata models in order to enrich PDB depositions with new classes of data and metadata to ensure higher compliance with FAIR principles in the field. This was achieved by extending the PDBe[116] PDBx/Macro Molecular Coordinate Information File (PDBx/mmCIF) with new data and metadata categories from early steps of the MX workflow. This work was carried out in consultation with key stakeholders within the Instruc-Eric and Inext Discovery networks, including members of other MX and NMR facilities in Europe. The new data and metadata classes are now part of the standard PDB mmCIF dictionary[117] used by the wwPDB.

The team has developed a new pilot metadata model capable of providing a full description of a prototypical fragment screening experiment. This model is based on the new wwPDB Investigation file format, and contrary to the current PDB standard, which only supports individual entries, makes it possible to associate multiple PDB coordinate files along with their corresponding data and metadata into a single experiment. This new investigation file format provides appropriate links between the fragment libraries, crystals, soaking experiments, raw crystallisation and X-ray diffraction data, and the resulting structural models in the form of PDB coordinate files for a complete fragment screening campaign, potentially including hundreds to thousands of individual experiments - something that was not possible with the current wwPDB model (see *Figure 1*).

---

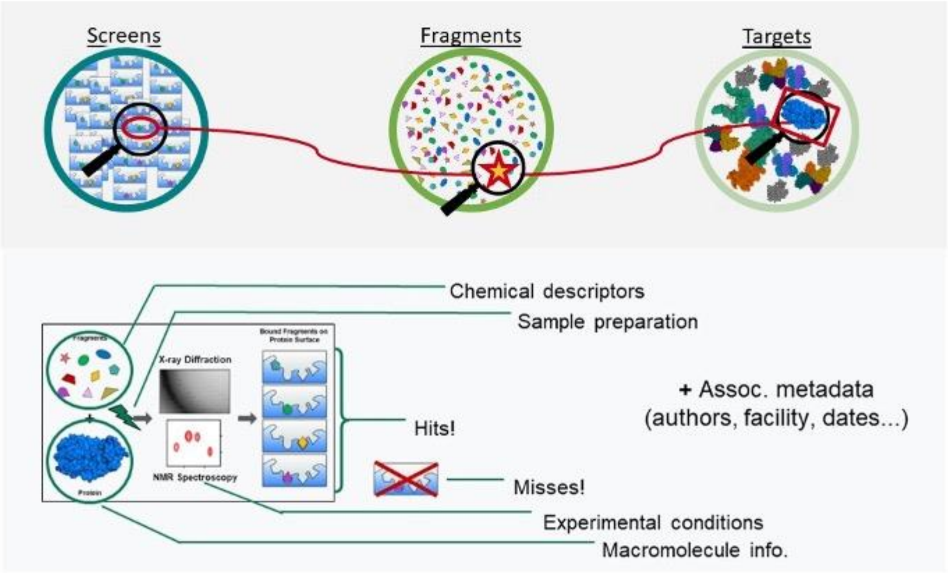
[115] https://instruct-eric.org/

[116] https://www.ebi.ac.uk/pdbe/

[117] https://mmcif.wwpdb.org/

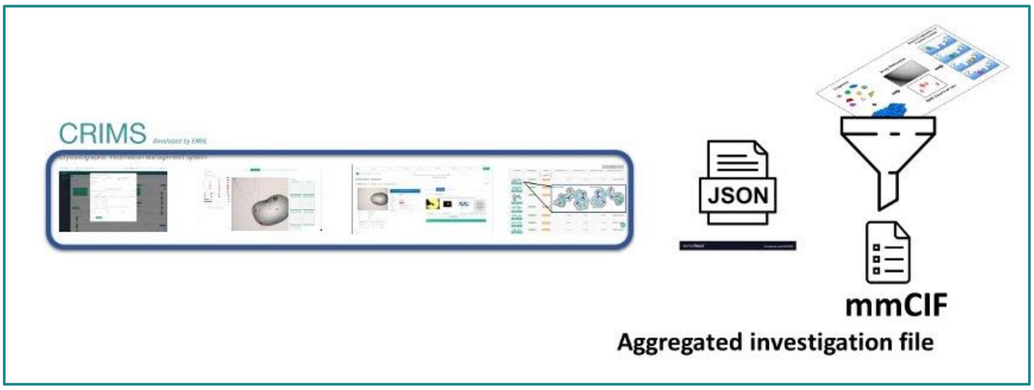
This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

The new investigation file format makes it possible to deposit all the results from a typical fragment screening campaign, including both positive and negative results instead of only the positive hits. It was developed in consultation with 5 different European facilities offering fragment screening services as with the wwPDB. This new format is not exclusive of X-ray based fragment screening applications but can also be used for any type of small molecule screening project and represents a major step towards the availability and reusability of structure-based drug design data. The pilot investigation file is currently being tested with prototypical fragment screening datasets generated at different European facilities.



PID 14366 - Figure 1: The new Investigation file format developed through this action makes it possible to represent a full fragment screening campaign. It groups multiple entries under a single experiment, with appropriate links between small-molecules, crystals, datasets and the resulting coordinate files. It enables the deposition of complete experiments rather than positive results and represents a major step towards the FAIRification of SBDD data.

The project team has initiated the development of software tools for automated data and metadata harvesting from the most commonly used crystallographic data management software in order to support automated deposition of SBDD campaigns under the Investigation file format described above. These tools collect data from the CRIMS and ISPyB software, used in many crystallisation and synchrotron facilities in Europe and will support translation into the mmCIF format supported by the new Investigation file format (*Figure* 2). This activity has involved the organisation of 4 different hackathons with the participation of multiple European facilities offering fragment screening services. Several representative SBDD datasets produced at the different facilities have been identified and will be used to test and validate both the new ned Investigation file format and the automated data deposition tools in the near future.



PID 14366 - Figure 2: Routines for automated data and metadata harvesting for automated deposition of SBDD experiments into the new mmCIF aggregated investigation file.

Through the Instruct-ERIC Network the team has been able to extend the research team and mobilise additional European facilities towards the FAIRification of SBDD. Support from the E.C. INFRATECH project Fragment Screen, coordinated by Instruct-ERIC in the form of additional funding has been obtained to continue these activities towards the implementation of an open pilot repository for SBDD. These activities are also being extended to other relevant experimental disciplines like Cryo-EM and NMR with the participation of expert groups in these areas.

SCIENTIFIC IMPACT

MX generates large amounts of data through complex experimental workflows at distant facilities. Moreover, new applications like X-ray-based fragment screening have led to a considerable increase in the volume of data produced and a bigger dependency from users on advanced facilities. While facilities are in the process of developing repositories for raw diffraction data through a number of initiatives, no tools are available for automated and comprehensive MX data harvesting, validation and deposition. As a consequence, the responsibility for data deposition and publication lays solely on the individual facility users. This represents a major obstacle both to populate EOSC with FAIR compliant data and for the consistent implementation of Open Science policies. The developments carried out here lay the foundation to improve FAIR compliance in structural biology and in particular structure-based drug design. In the long term this will:

- Facilitate implementation of FAIR data policies at MX facilities within Instruct-ERIC and elsewhere.
- Equip structural biology facilities with tools to support hundreds of users with FAIR compliant data.
- Enable the implementation of Open Science policies at Instruct-ERIC facilities, by enabling automated data and metadata deposition under embargo with automated public release according to generally accepted policies.
- Contribute to populate the EOSC with large amounts of highly interconnected and interoperable structural biology data
- Enable the future development of large-scale Open Data repositories for SBDD. Until recently, SBDD projects were mostly carried out in industry and results not publicly shared. Recent breakthroughs at multiple Instruct-ERIC facilities have made it possible to support a large number of academic, translational research-oriented projects with advanced compound and fragment screening technologies producing a wealth of data. FAIR-compliant, Open Science repositories for this type of data will enable the development of innovative

approaches for drug design based on advanced computing and machine learning with impact in drug design.

The dissemination of the team activities was facilitated with contributions from the Instruct-ERIC, iNEXT Discovery and EOSC Life networks. 4 meetings with representatives from European facilities providing SBDD services were held during the duration of the project. The project was also presented at several international conferences, including among others the iNEXT Discovery webinar series, the Inext Discovery workshop on fragment-based drug design (Berlin, 2022) as well as the EOSC Life AGM.

ROLE AND IMPACT OF EOSC-Life

According to the project team, the EOSC Life project has provided a great opportunity to initiate and structure their activities. The whole EOSC Life Team has been very proactive and supportive both at the technical and scientific levels. While the amount of funding was limited, it has been instrumental in spearheading this initiative and in enabling the nucleation of a larger research team. In the future the team recommends the continuation of this type of programs but with an increased level of funding for these types of activities. This will enable a transition from exploratory to more ambitious proposals supported by larger, mature research teams with a durable impact in EOSC.

4.9 Digital Life Sciences Internal Call: Sensitive Data project - PID 17410

A cloud-ready data management and processing platform for sensitive cancer-genomics data

Project team: Sven Twardziok, Valentin Schneider-Lunitz, Philipp Strubel, Philip R. Kensche, Ivo Buchhalter

PROJECT DESCRIPTION

With the ongoing fast technical advancements in the life sciences and healthcare sector, the amount of data is growing exponentially. This led to a significant challenge in sharing datasets, particularly whole genome sequencing data generated from cancer tissue. The data is extremely large and complex, but its content is very helpful in improving diagnostics and patient treatment. The sequencing information aids in the identification of patient-specific mutations as well as the classification of different types of cancer. However, the results produced from these data contain highly sensitive and personal information, which is protected by law. This makes data sharing challenging but also impractical. Despite the challenges, sharing data and knowledge remains a critical tool in basic cancer research. The combination of large genetic and clinical information from many patients increases the power to identify rare conditions and improve our understanding of the disease's nature. To address these requirements, we proposed the development of a cloud-ready management and processing platform for sensitive genomics data.

Rather than sending data between collaborating institutes, the aim is to utilise a management platform to send workflows to the data. This makes data sharing superfluous, saves time and enhances data security by keeping the data in one place. To ensure standardised processing of the data the Global Alliance for Genomics and Health[118] (GA4GH) develops standards for collecting, storing, analysing, and sharing genomics data. One example is the GA4GH Workflow

---

[118] https://www.ga4gh.org/

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

Execution Service (WES) that describes a programmatic way to run and manage workflow executions.

The project team has been using the “One Touch Pipeline” (OTP), a platform originally developed by the German Cancer Research Center (Deutsches Krebsforschungszentrum, DKFZ)[119]. OTP was used as a part of the International Cancer Genome Consortium (ICGC)120 to manage and process sensitive cancer genomics data.

Currently, OTP is used at the DKFZ and Berlin Institute of Health (BIH) for various projects including next-generation sequencing (NGS) data but also management of single-cell genomics data from COVID-19 patients[121].

OTP contains workflows designed to analyse whole genome sequencing (WGS), exome sequencing, RNA-seq, whole genome bisulfite sequencing (WGBS), and single-cell RNA sequencing (10x_scRNA-seq) data. The platform allows orchestration of individual users based on specific projects, automatic execution of selected workflows, and inspection of the results.

The current development of OTP aims to integrate an independent implementation of the workflow execution service called “WESkit” which is based on the GA4GH WES standard. WESkit supports the execution of Snakemake and Nextflow workflow engines. Additionally, it allows scalability to process up to hundreds of jobs in parallel.

The goal of the project is to deploy both OTP and WESkit on the cloud to process genomic data. Finally, the aim was for the software stack to be integrated with corresponding EOSC-Life services as well as made available as an EOSC service for other researchers. This will enable the processing of sensitive cancer data within their home institutes.

PROJECT RESULTS AND ACHIEVEMENTS

During the initial phase of the project, both OTP and WESkit were prepared for deployment in the cloud. Independence from the underlying infrastructure was ensured by creating Docker images for both systems and deployment in a Docker swarm along with their dependencies. To demonstrate the deployment in a cloud system, a SLURM cluster was established on the Berlin node of the de.NBI cloud and individual VMs were generated for OTP and WESkit.

Registered users are required to authenticate with OAuth2.0 within OTP, to maintain data security. A separate login to WESkit is not necessary, as it functions as a backend in OTP. The stability of the system and reproducibility of the executions is ensured through excessive automatic testing in the continuous integration (CI) in GitLab for each branch and merge-request. For the identification of common security issues ESLint (OTP) and Bandit (WESkit) were used. Data security and protection of sensitive genomic data are ensured through state-of-the-art automated hardening specifications. Currently, WESkit is configured to be ready to serve as a backend in OTP without loosening previous security settings in data storage and processing.

Initially, OTP utilised Roddy as a backend for developing and managing workflows. As a result, the current workflows are either Roddy-based or standalone, such as CellRanger. However, to reduce dependency on Roddy and improve compatibility with WESKit and cloud environments, the DKFZ is in the process of removing this dependency and rewriting the workflows using established

---

[119] https://doi.org/10.1016/j.jbiotec.2017.08.006
[120] https://doi.org/10.1038/s41586-020-1969-6
[121] https://doi.org/10.1038/s41587-020-00796-1

workflow engines such as Snakemake or Nextflow. Additionally, Conda environments and Docker will be used to execute individual workflows, improve reusability and ensure high reproducibility.

To prevent any unexpected behaviour of the workflows during processing, automatic workflow testing would be beneficial. It would allow quick identification of potential bugs and follow-up debugging and ensure high reproducibility. This procedure was suggested by the nf-core team for all Nextflow workflows.

Approved workflows are stored on GitLab in the OTP repository (https://gitlab.com/one-touch-pipeline/workflows) and are available to the community. Storage in GitHub and workflow registration at WorkflowHub[122] is considered to achieve high findability.

OTP was expanded with the ability to support OAuth2.0 authentication to the software stack, enabling end-users a single-sign-on authentication. The integration of OAuth2.0 ensures compatibility with Life Science Login (LS-AAI)[123], an authentication service from EOSC-Life which combines ELIXIR AAI and EOSC-Life Login. For the extension of cloud compatibility, the WESkit team implemented an initial, experimental GA4GH TRS (Tool Registry Service) support during the Biohackathon 2021. TRS requires the installation of the workflow into a shared directory. However, TRS support is not production ready. During the ELIXIR Biohackathon 2021, we further improved the cloud deployment of WESkit by extending the system with a prototypical S3 data export feature based on MinIO. This feature allows storage and access to individual results using the OAuth2.0 system and S3-URIs that are listed in the WES REST responses. MinIO can be co-deployed with the WESkit stack.

For demonstration purposes of cloud usage, the de.NBI cloud was used. Here the software stack for both the OTP and WESkit were deployed on separate virtual machines. A small Slurm cluster is used to execute workflows. A public genomics cancer data set from cell culture sequencing will be used as input. Interested users are able to login to the demo site at https://otp-demo.bihealth.org/. All steps are documented and will be made publicly available in a newly generated comprehensive Gitlab OTP Documentation[124]. This documentation also includes an operating manual and a data security checklist and background information that are useful for communication with local data security offices.

The OTP demonstrator platform is currently (25.07.2023) in the registration process and awaits approval by the legal department of the home institution. Despite previous time constraints and project status, significant progress has been made towards achieving this goal.

SCIENTIFIC IMPACT

Most projects tend to keep their sensitive patient sequencing data at their respective institutes, with data sharing only among cooperating partners within the limits of national and legal boundaries. However, this project offers new possibilities in terms of global data access. Deployment of OTP together with WESkit allows registered users to submit validated workflows on sensitive genomic data and obtain summarised results without accessing the actual data.

As a result, drawing conclusions on individual patients based on e.g., private mutations become impossible and data sharing becomes obsolete since it remains at its home institution. This concept enables groups worldwide to use condensed information from validated workflows to

---

[122] https://workflowhub.eu/
[123] https://lifescience-ri.eu/ls-login.html
[124] https://gitlab.com/users/sign_in

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

answer specific questions such as the effectiveness of a particular therapy on a specific cancer type or the occurrence of a specific mutation in a cohort.

The team participated in the ELIXIR Biohackathon 2021 to implement a prototypical GA4GH TRS support to automatically retrieve workflows, as well as an improved cloud deployment and a prototypical S3 data export feature based on MinIO. Already during this Biohackathon, the team intensified the communication with the GA4GH project[125] and presented WESkit at the GA4GH 10th Plenary Meeting in 2022 and also presented a user demonstration at the GA4GH connect meeting in April 2023. In the ELIXIR Biohackathon 2022[126], they worked on the interaction of the Snakemake workflow engine with the TES API using TESK (ELIXIR Cloud & AAI, 2023) to run workflow jobs on Kubernetes and on implementing the required changes in WESkit, e.g., to manage authentication tokens. The results from this project were presented in July 2023 as a contribution to the ECCB/ISMB 2023 conference with the title: "*OTP@EOSC: A cloud-ready data management and processing platform for sensitive cancer-genomics data*".

ROLE AND IMPACT OF EOSC-Life

Overall, the experience of the team with EOSC-Life was very good. They mentioned that they received a lot of important and productive feedback and suggestions during the project. The meetings in WP3 and with the other projects were always professionally organised. In addition, the people involved were always available for questions.

4.10 Digital Life Sciences Internal Call: Academia - Industry Collaboration - PID 17989

QSM4SENIOR: "Ultra-high Field MRI Quantitative Susceptibility Mapping reconstruction of the healthy Brain-ageing cohort"

Project team: *Miguel Guevara, Alexandre Vignaud, Jean-Francois Mangin, Yann Cointepas, Ludovic de Rochefort, Michael Bottländer*

PROJECT DESCRIPTION

Given the perspective of current demographic trends, neurodegenerative diseases are expected to affect a rising number of people in the future. Detection of early biomarkers predicting brain disease is moving more into focus of current research. SENIOR[127,128] (BAOBAB[129], Neurospin[130], CEA) is a longitudinal observational study including 100 cognitively healthy elderly people between 50 and 70 years old. Subjects are followed annually for 10 years which involves multimodal brain imaging data, as well as several biological and neurological data. BAOBAB is in the process of making this cohort data available in a databank for scientific research, with the ambition to help investigate inter-individual physiological ageing, identifying predictive biomarkers and risk factors of brain diseases.

---

[125] https://doi.org/10.37044/osf.io/2z6nu

[126] Kanitz, A., & Kensche, P. (2022). *ELIXIR Biohackathon 2022 project 30: The ELIXIR:GA4GH Cloud* https://github.com/elixir-europe/biohackathon-projects-2022/tree/main/30

[127] https://alzres.biomedcentral.com/articles/10.1186/s13195-020-00642-1

[128] https://joliot.cea.fr/drf/joliot/Pages/Actualites/Scientifiques/2020/Cohorte-SENIOR-vieillissement-cerebral.aspx

[129] https://joliot.cea.fr/drf/joliot/Pages/Entites_de_recherche/neurospin/baobab.aspx

[130] https://joliot.cea.fr/drf/joliot/recherche/NeuroSpin

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

Iron load represents the most promising predictive biomarker in this context. The accumulation of iron in the brain during ageing was shown to be correlated with the occurrence of neurodegenerative disease. The SENIOR imaging protocol includes iron-sensitive imaging data at ultra-high-field Magnetic Resonance Imaging, offering great spatial resolution as well as improving the sensitivity to some specific contrasts such as iron load. The best way to extract this data is to perform an image reconstruction technique called quantitative susceptibility mapping (QSM).

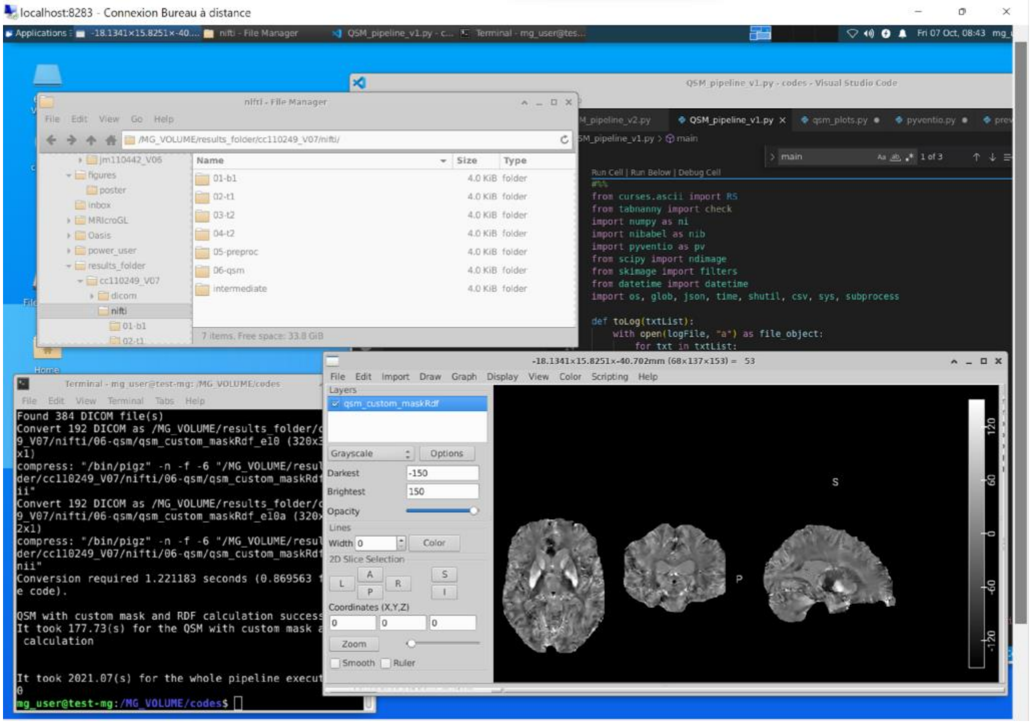
The main objective of the QSM4SENIOR project is to reconstruct QSM maps with the cloud based MEDI software, the most commonly used method in literature. The MEDI licence was provided by industry collaborator Ventio. Another key objective is the FAIRification of the imaging data in order to make them accessible. The project involves cloud processing of sensitive human brain images, therefore ethical and legal aspects and the consequences on data security were considered. In the meaning of GDPR, the BAOBAB will act as a controller, and Ventio as well as the cloud resource providers as processors. The DICOM data, currently stored in CEA premises in a secure network, includes the scans for each volunteer. The data is transferred to a virtual server, set up by Ventio, on openstack-deployed cloud resources (in the de.NBI cloud in this project).

Encryption at rest and on transit is used, with access control (user-specific password and ssh keys). A script running periodically checks the reception of data and triggers QSM processing on the virtual server. Image processing results can be visualised remotely, and the data is then sent back to CEA premises using secure transfer.

PROJECT RESULTS AND ACHIEVEMENTS

Regarding ethical, technical, organisational, legal aspects for the processing of the sensitive data: most ethical aspects were already covered by CEA in setting-up the SENIOR study. A collaboration contract was set-up between BAOBAB (CNRS and CEA) and Ventio. Access to the SENIOR database was granted.

The infrastructure interoperability was demonstrated thanks to the access to Tübingen University HPC resources (EOSC-Life research infrastructure): de.NBI Cloud. A contract was established between Tübingen University and Ventio. Access to the adapted bioinformatic resources with the ISO27001 certification and the GDPR compliance needed for human sensitive data was obtained. The computing resources correspond to an openstack project space with allocated resources (32 vCPUs, 34 Go RAM, 200 Go volume storage) on which a virtual machine of 14 cores, 32 Go RAM was set-up. A client-side Infrastructure-as-a-code deployment strategy was used to deploy and configure the server with all needed installed software and functionality. An implementation of the security needed to provide transnational resources in Europe and the imaging data encryption on transit and at rest was also included. Encryption at rest was implemented by mounting user-controlled encrypted data containers. The necessary image processing software as well as Python packages for image processing were installed in the virtual environment. A view of the virtual desktop is shown in *Figure 1*.



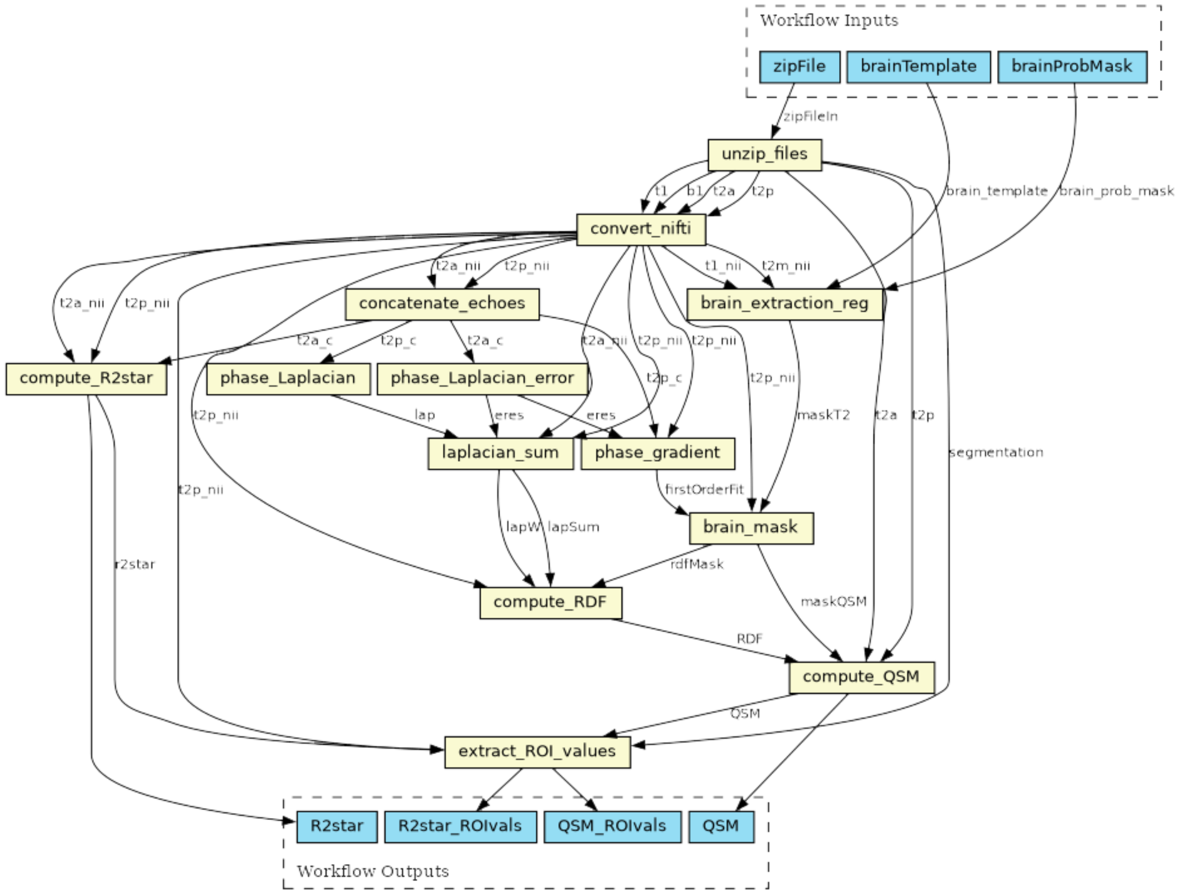
```
Run Cell | Run Below | Debug Cell
#%%
from curses.ascii import RS
from tabnanny import check
import numpy as ni
import nibabel as nib
import pyventio as pv
from scipy import ndimage
from skimage import filters
from datetime import datetime
import os, glob, json, time, shutil, csv, sys, subprocess

def toLog(txtList):
    with open(logFile, "a") as file_object:
        for txt in txtList:
```

```
Found 384 DICOM file(s)
Convert 192 DICOM as /MG_VOLUME/results_folder/c
9_V07/nifti/06-qsm/qsm_custom_maskRdf_e10 (320x3
x1)
compress: "/bin/pigz" -n -f -6 "/MG_VOLUME/resul
der/cc110249_V07/nifti/06-qsm/qsm_custom_maskRdf
ii"
Convert 192 DICOM as /MG_VOLUME/results_folder/c
9_V07/nifti/06-qsm/qsm_custom_maskRdf_e10a (320x
2x1)
compress: "/bin/pigz" -n -f -6 "/MG_VOLUME/resul
der/cc110249_V07/nifti/06-qsm/qsm_custom_maskRdf
nii"
Conversion required 1.221183 seconds (0.869563 f
e code).

QSM with custom mask and RDF calculation success
It took 177.73(s) for the QSM with custom mask a
 calculation


It took 2021.07(s) for the whole pipeline execut
0
mg_user@test-mg:/MG_VOLUME/codes$ 
```

PID 17989 - Figure 1. Virtual desktop deployed in the de.NBI Cloud with an example of installed application to launch specific processing with python / jupyter notebooks and to visualise the QSM 3D results.

Regarding the data preparation for processing on the cloud, the data was recovered from the CEA premises in DICOM format. First, scripts checking raw QSM data quality (acquisition parameters, image quality, absence of motion) are run, which then prepare the dataset for transfer to Ventio's QSM services by organising and compressing all the necessary data. The zipped file is then securely transferred to the Cloud environment over ssh. A pipeline to compute QSM maps and to extract the values for regions of interest (ROIs) was implemented. This pipeline was structured using Common Workflow Language (CWL), which specifies every step of the pipeline and calls Python scripts that perform the required computation step. A schematic of the pipeline is presented in *Figure 2*.

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



PID 17989 - Figure 2. Schematic of the pipeline implemented using Common Workflow Language and Python/bash scripts.

The pipeline also incorporates the computation of R2* maps, by means of Gauss-Newton algorithm, which is another parameter linked to iron load in order to compare to QSM values. For the QSM computation, it incorporates a pre-filtering of the phase data from the MRI multi-gradient echo (MGRE) acquisition, by using the information of the magnitude and phase from the ten echoes. For both R2* and QSM maps, deep grey matter nuclei values were extracted using Regions-of-interest (ROIs) obtained with VolBrain.

Regarding the FAIRification of the imaging data, the imaging results (as DICOM and Nifti files) were prepared for delivery (compression, file signature) and then were sent back to the CEA server hosting SENIOR database, along with the tables containing the values in the different ROIs.

Concerning the workflow and sustainability of the code, Git versioning was used for script and pipelines. The developments made by the partners BAOBAB and Ventio are in process of being made accessible in a Git repository. The CWL workflow management system facilitates the reusability of the code.

SCIENTIFIC IMPACT

The project sets a precedent regarding the possibility of remote computing to process medical sensitive data. Moreover, the computation of QSM as a biomarker of the iron load could be a

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

powerful tool for the early diagnosis of neurodegeneration. Thanks for the results of this work it was observable that QSM seems to be more sensitive than R2* to changes in the iron load over the years. The fact that the QSM computation was applied to a cohort of healthy senior subjects allowed the team to establish a range of values for the "normality", in terms of the quantification of the iron load. This data could be used in the future to help the physicians to determine if a patient follows an abnormal course and is prompt to suffer from a degenerative disease associated with iron load, such as Alzheimer's disease. The automated pipeline can be applied to other databases. In particular, BAOBAB will apply it to SHATAU database, acquired with the same imaging parameters as SENIOR, which contains data of Alzheimer's patients. This will help studying quantitative differences between healthy subjects and patients.

The project outcomes have been disseminated in several important conferences in the research field (QSM and MRI):

- QSM workshop 2022, Lucca, Italy (Guevara et al.[131])
- SFRMBM 2023, Paris, France (Guevera et al.[132])
- ISMRM 2023, Toronto, Canada. (Accepted for publication)

The manuscript *Guevara et al. "Study of the iron load in normal ageing brain by using an automated pipeline for the QSM and R2* computation"* is being finalised (August 2023) and will be soon submitted to *Neurology*.

ROLE AND IMPACT OF EOSC-Life

According to the team, the interaction with EOSC-Life was extremely fruitful. The team appreciated the effort to be initiated in the use of the necessary tools to fulfil the objectives of the project.

The team participated in several Workflow-hub meetings (WP2), which gave us a broad view of the different scopes that this community has to offer. The interaction with de.NBI cloud representatives (WP7) at Tubingen were also extremely positive. The interaction with them was very efficient, allowing the team to take advantage of the infrastructure very quickly. The project follow-ups were very appreciated too. That allowed the team to have useful input from the different WP as the project progressed. The EOSC community facilitated the networking between the project team and the different members at several events.

The team attended the EOSC-Life FAIR hackathon[133], where the different WPs gave a very good view about what should be done and how. Moreover, it also gave a vision on the use-cases and applications to the industry. Meetings such as the BioImaging and the European Open Science Cloud Workshop[134] held in Heidelberg the 19th and 20th April give visibility of the work of the imagining in life science communities. This allows sharing experiences and to see/understand better other use-cases. Therefore, more meetings like this should occur, maybe in online format.

---

[131] https://zenodo.org/record/7303615#.ZEquQXZBzEY

[132] https://zenodo.org/record/7871306#.ZEqDIHZBzEY

[133] https://www.eosc-life.eu/services/training/

[134] https://www.eurobioimaging-industryboard.com/workshop_report_2023/

4.11 Digital Life Sciences Internal Call: Academia - Industry project - PID 17836

Accessible and scalable detection and identification of foodborne pathogens

Project team: Bérénice Batut, Engy Nasr, Ralf Seyfarth, Anna Henger

PROJECT DESCRIPTION

Food contamination with pathogens is a major burden on our society. In the year 2019, foodborne pathogens caused 137 hospitalisations in Germany (BVL 2019)[135]. Globally, they affect an estimated 600 million people a year and impact socioeconomic development at different levels. These outbreaks are mainly due to Salmonella spp. followed by Campylobacter spp. and Noroviruses, as studied by the Food safety - World Health Organization (WHO)[136].

During the investigation of a foodborne outbreak, a microbiological analysis of the potentially responsible food vehicle is performed in order to detect the responsible pathogens and identify the contamination source. By default, the European Regulation (EC)[137] follows ISO standards to detect bacterial pathogens in food: pathogens are detected and identified by stepwise cultures on selective media and/or targeting specific genes with real-time PCRs. The current gold standard techniques have some disadvantages, and single organism targeted Whole Genome Sequencing (WGS) has been proposed as an alternative. With just one sequencing run, one can: (i) detect all genes, (ii) run phylogenetic analysis to link cases, (iii) get information on antimicrobial resistance genes, virulence, and other critical factors. Despite its recommendation by several Public Health organisations. Single-organism WGS is limited by isolation of the targeted pathogen: it is a time-consuming process, the execution is not always straightforward, nor success is guaranteed.

Sequencing methods without prior isolation as metagenomics could solve this issue. They give an overview of the genomic composition in the sample, including the food source itself, the microbial community, and any possible pathogens and their complete genetic information.

It does not need prior DNA isolation as required by single-organism WGS, it is not limited to specific genes as in the real-time PCRs approach, and it is more accurate based on the entity of the genetic information contained in samples. Combined with Oxford Nanopore (ONT) sequencing, it becomes easier, more accessible, and more practical to quickly identify pathogens.

Biolytix, the industry partner of this project, developed a procedure to extract sequence using ONT sequencing, and analyse a food sample for foodborne pathogen detection and contamination source tracking. However, the bioinformatics pipelines are not following the best practices for open data science, are not straightforward to use, and can only be manipulated by the original author, making it hard to scale and reuse.

Within this project, the project team modernised and opened up the current Biolytix' foodborne pathogen detection pipelines using modern paradigms of open data science, ensuring they are accessible, and scalable but also portable, cloud deployable, maintainable, and shareable.

The workflows use state-of-the-art tools to detect and track pathogens from any - not only food - metagenomics Oxford Nanopore sample. The team successfully tested the workflows on 124

---

[135] https://www.bvl.bund.de/SharedDocs/Berichte/10_BELA_lebensmittelbed_Krankheitsausbruechen_Dtl/Jahresbericht2019.pdf?__blob=publicationFile&v=4

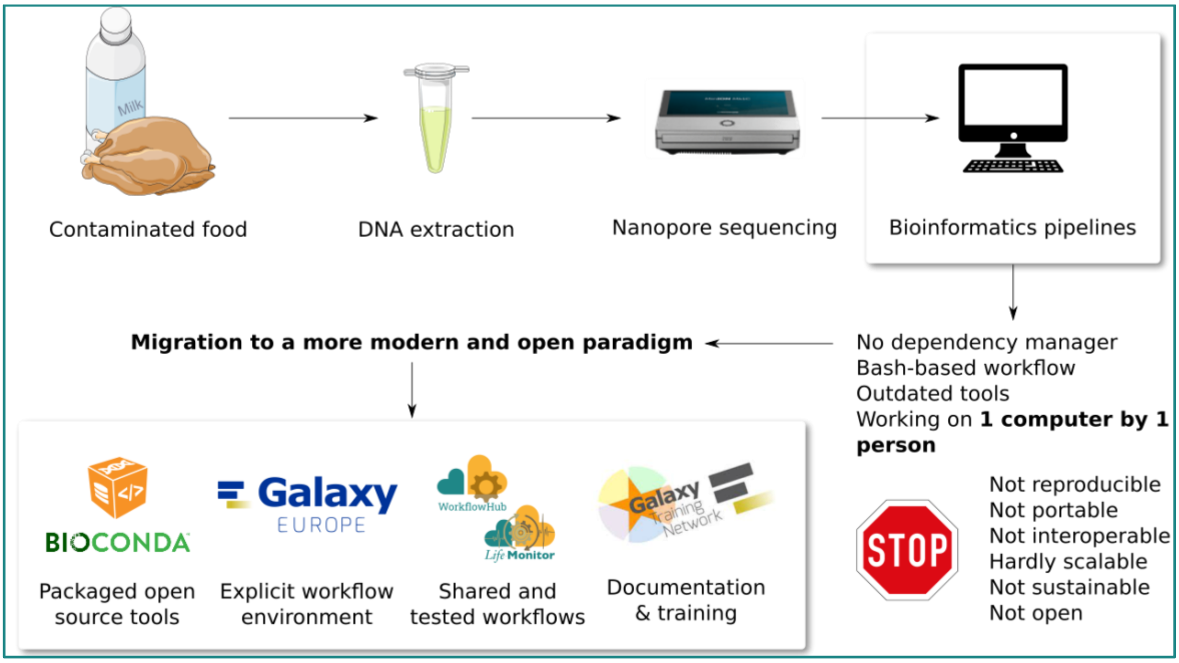
[136] https://www.who.int/publications/i/item/9789241565165

[137] https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2005:338:0001:0026:EN:PDF

samples of different pathogens: (1) spiked food with different *Salmonella enterica* strains enriched using non-targeted enrichment techniques at different concentrations, and (2) samples from humans and chicken stool as well as beef collected in Palestine containing *Campylobacter jejuni*.

PROJECT RESULTS AND ACHIEVEMENTS

Biolytix developed pipelines for foodborne pathogen detection from ONT metagenomics data. This bash workflow neither used a dependency manager nor was it possible to use parallel computing techniques, which made the workflow non-portable, non-interoperable, and difficult to scale. Any changes in underlying versions of tools were difficult and prone to break the setup process. The long-term sustainability was jeopardised by the departure of key personnel, as the pipelines could only be changed by one person, making them quickly obsolete. Workflow testing was not in place either and the documentation could be improved considerably.



PID 17836 Figure 1: Update of the the pipelines for foodborne pathogen detection

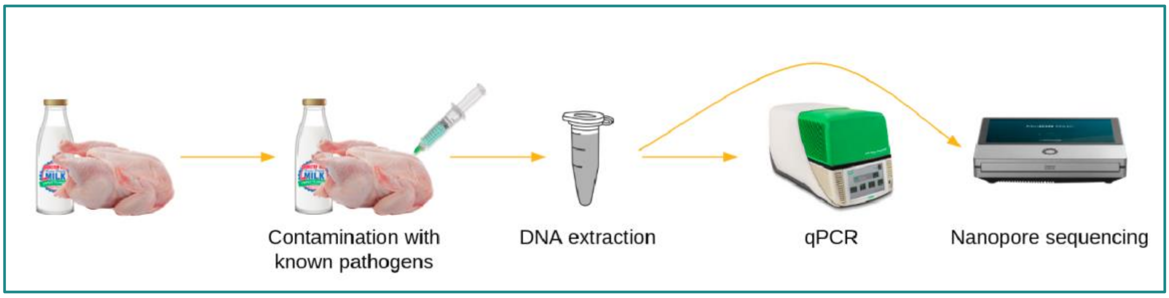
The project team took the pipelines for foodborne pathogen detection from ONT metagenomics data developed by Biolytix. They evaluated them given the used tools and their licences, custom scripts, and outputs to re-engineer them (*Figure 1*) and have:

- split them into 5 workflows: Pre-Processing, Taxonomy Profiling, Gene-based pathogenic identification, SNP-based pathogenic identification, Pathogen Tracking among all samples;
- replaced some tools and custom scripts by state-of-the-art tools;
- packaged all tools as Bioconda/BioContainers packages;
- made sure all tools are available as Galaxy tools by creating and updating Galaxy wrappers
- migrated the workflows as Galaxy workflows
- added steps of report generations

To validate the workflows, Biolytix generated test sequencing data by spiking chicken meat with different Salmonella enterica strains at different concentrations (*Figure* 2). The raw data will be

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

deposited on the European Nucleotide Archive (ENA)[138] and the protocols for spiking food in an open protocol database.



PID 17836 - Figure 2: Workflow validation

Using data, the project team generated a test suite for automatic workflow testing on the European Galaxy server and EOSC-Life LifeMonitor[139].

They also validated the workflows on samples collected from humans, meat, and chicken in Palestine containing *Campylobacter jejuni*. The workflows are available to be directly used on the European Galaxy server and can also be deployed on other Galaxy servers.

To make the workflows FAIR, the team annotated and deposited them in the Intergalactic Workflow Commission (IWC) GitHub repository. Here, they were reviewed by the community and once accepted, pushed to EOSC-Life Workflow Hub.

The project team also created an e-learning tutorial available on Galaxy Training Network to document the workflows and organised a dedicated workshop in which 90 people registered and 35 participated.

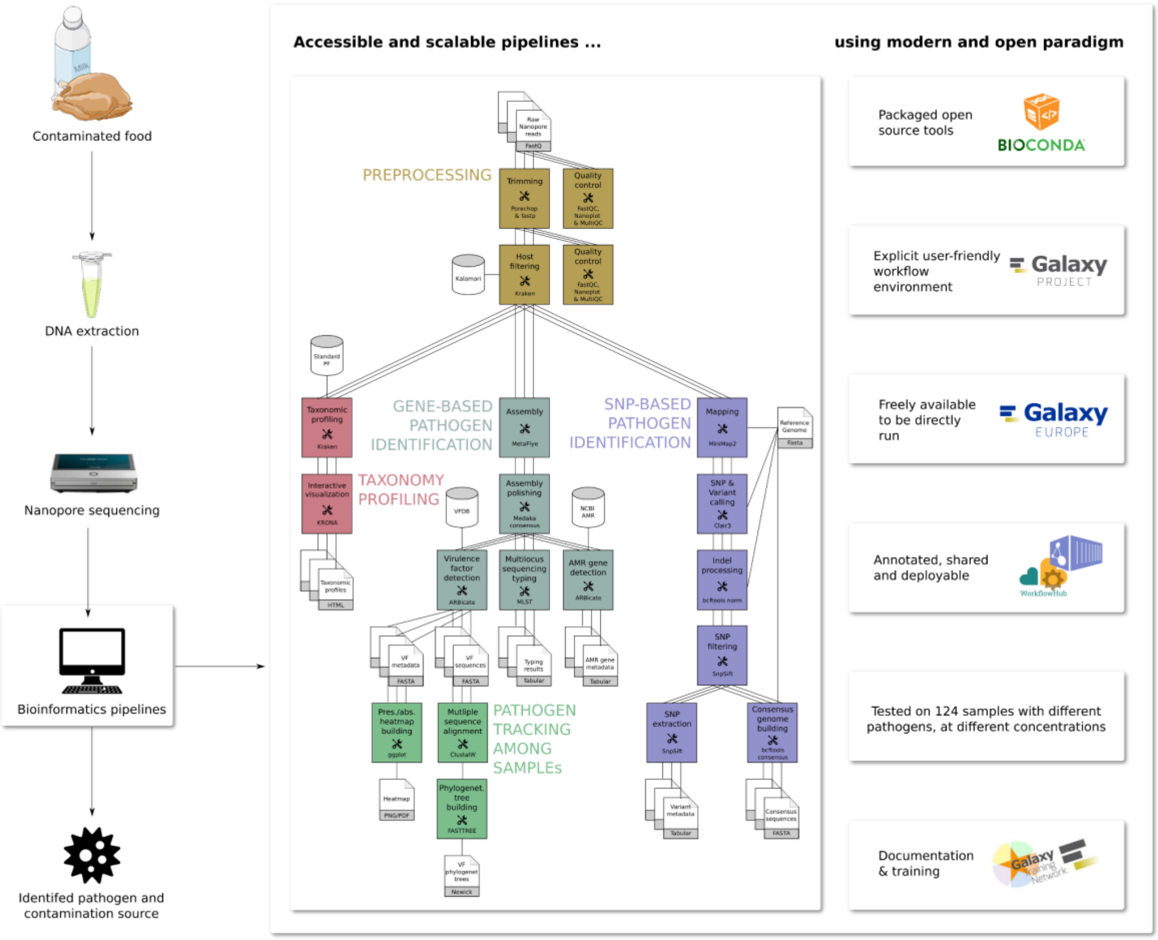
The result of this project is a series of accessible and scalable pipelines for fast and easy (foodborne) pathogens detection and tracking using modern and open paradigms (see *Figure 3*). A manuscript is currently in preparation and will be submitted to BioRxiv after the end of the EOSC-Life project.

---

[138] https://www.ebi.ac.uk/ena/browser/home

[139] https://lifemonitor.eu/

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



PID 17836 - Figure 3: Pipelines for fast and easy (foodborne) pathogens detection and tracking

SCIENTIFIC IMPACT

Foodborne pathogens and their impact on human health is a major problem in the food industry. Advanced detection and tracing beyond classical methods would have a tremendous impact on large and small-scale food production. The identification of the exact pathogenic strains and their genomic information directly in the food without pathogen isolation would lead to a new level of food security. The source of food contaminants could be faster identified and avoid outbreaks of foodborne illnesses. Future outbreaks of foodborne pathogens with all their negative implications like cost and disease could be reduced.

This project gives access to the EOSC and scientific community to resources (FAIR data, tools, workflows, and training) to trace back food pathogens in the food chain while fostering collaboration with other organisations doing pathogen detection analysis.

The workshop organised by the project team reached more than 90 registrations and had really positive feedback, which shows the need from the community for such resources.

Test data and their metadata as well as protocols to generate such data will be deposited in the next month on public databases for others to use and validate their workflows.

The maintenance and development of Bioconda packages and new tools benefit the wider community that currently runs or aims to run pathogen detection analysis on different platforms. Moving workflows to a cloud-deployable workflow environment like Galaxy increases the potential compute capacity where these analyses can run. Moreover, Galaxy makes these resources more accessible: anyone, even without computational knowledge, will be able to run them.

The training community around pathogen detection also benefits from the updated and improved workflows and their accompanying online training material on real-case scenarios.

More researchers and industries will be able to trace back pathogens using metagenomics sequencing of food samples directly on the field and to do their own analyses by re-using and adapting these standardised workflows. These workflows enable faster foodborne pathogen identification and improve scrutiny of how such analyses are run. They are a great benefit for food safety, outbreak detection, and beyond. Other projects (e.g. EuroScienceGateway) and communities (e.g. Biodiversity) will also profit from the work there.

The team started to disseminate the outcomes of this project, mainly by writing a paper that will be submitted soon. In parallel, the team submitted abstracts for talks and posters at several conferences: Applied Bioinformatics & Public Health Microbiology[140] in May (accepted as a demo), ELIXIR All Hands meeting in June 2023 (accepted as a poster), German Bioinformatics Conference in September, and the Bioinformatics Open Source Conference[141] at ISMB/ECCB in July 2023.

The team organised a 1-day practical workshop for (foodborne) pathogen detection from (direct Nanopore) sequencing data using Galaxy[142] at the beginning of February 2023. During this workshop, the tutorial developed for this project was presented.

The tutorial was also part of the High-Throughput Data Analysis workshop with Galaxy (HTS) workshop in Freiburg[143] in February 2023 and is recorded to be freely available online and used as part of the GTN 2023 Smörgåsbord[144], a free, global, online week of training around Galaxy.

The project and the tutorial were also presented during the OSA2micro workshop, in July 2023.

This project was also disseminated on the Galaxy Community Hub145 with several blog posts: one about the workshop and one will be published about the project in general once a preprint of the paper will be available.

Finally, this project was disseminated to the community of Nanopore users in Switzerland, mainly the cantonal laboratories, the Swiss Tropical Institute of Health, and in Germany the Chemical and Veterinary Investigation Offices (*CVUA*).

ROLE AND IMPACT OF EOSC-Life

The main benefit of joining the EOSC-Life project was the connections to LifeMonitor, WorkflowHub[146], and other EOSC-Life services. The team especially appreciated the orientation

---

[140] https://coursesandconferences.wellcomeconnectingscience.org/event/applied-bioinformatics-public-health-microbiology-20230503/

[141] https://www.open-bio.org/events/bosc-2023/

[142] https://galaxyproject.org/events/2023-02-02-foodborne-pathogen-detection-workshop/

[143] https://galaxyproject.org/events/2023-02-20-freiburg-workshop/

[144] https://gallantries.github.io/video-library/events/smorgasbord3/

[145] https://galaxyproject.org/

[146] https://workflowhub.eu/

and interim review meetings, the discussions, and the resources shared during these meetings, in particular around data management plans. Getting the needed seed funding to start working on this project full-time has enabled to strengthen the connections between the Freiburg Galaxy team and Biolytix.

Through its participation in EOSC-life, the team also developed relationships with MIRRI[147] via Paolo Romano. The team is currently evaluating what Galaxy could provide to MIRRI. Bérénice Batut was also invited to talk about Galaxy for microbial data analysis, including this project, at the OSA2micro workshop, organized by Paolo Romano in July in Turin.

This project is a nice example to illustrate how to develop FAIR, open and sustainable infrastructure using Galaxy. It helped reinforce the interactions within ELIXIR, especially with the emerging Microbiome or the Biodiversity communities and we plan to use it to develop interactions with the Food & Nutrition community.

---

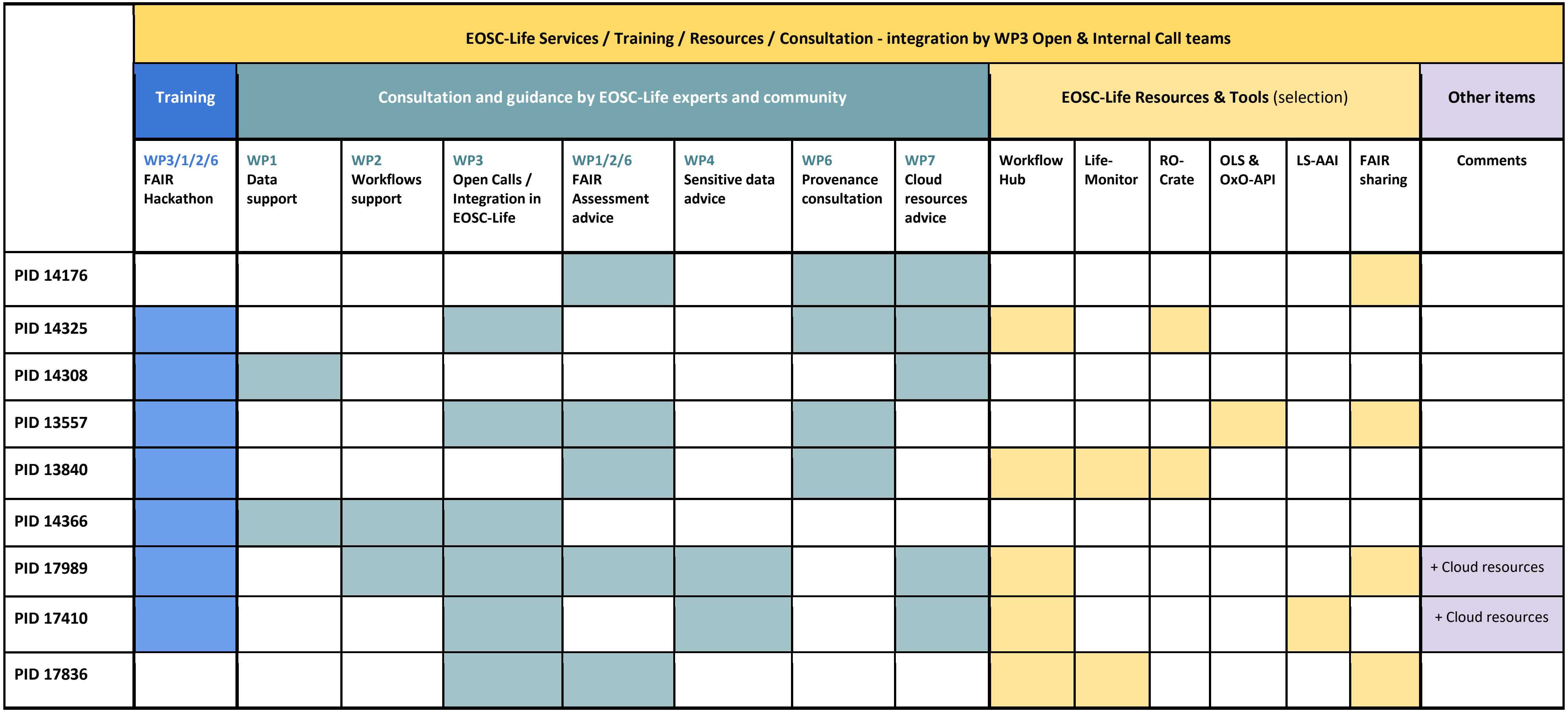
[147] https://www.mirri.org/

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

5. EOSC-Life services integrated by Open & Internal Call teams

EOSC-Life Training Initiatives	EOSC-Life Experts Advice & Consultation	EOSC-Life Resources & Tools
WP3/1/2/6 FAIR Hackathon ● Data life cycle for the distributed provenance chain ● Resources and tools for data FAIRification ● Tools and workflows FAIRification **WP9** Training Open Calls	**WP1** - Data support **WP2** - Workflows support **WP3 -** Calls/Integration in EOSC-Life **WP1/2/6** - FAIR Assessment advice **WP4** - Sensitive data advice **WP6 -** Provenance/data life cycle consultation **WP7** - Cloud resources advice	WorkflowHub FAIRsharing LifeMonitor RO-Crate OLS & OxO-API LS-AAI Cloud Resources

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



	EOSC-Life Services / Training / Resources / Consultation - integration by WP3 Open & Internal Call teams													
	Training	Consultation and guidance by EOSC-Life experts and community							EOSC-Life Resources & Tools (selection)					Other items
	WP3/1/2/6 FAIR Hackathon	WP1 Data support	WP2 Workflows support	WP3 Open Calls / Integration in EOSC-Life	WP1/2/6 FAIR Assessment advice	WP4 Sensitive data advice	WP6 Provenance consultation	WP7 Cloud resources advice	Workflow Hub	Life-Monitor	RO-Crate	OLS & OxO-API	LS-AAI	FAIR sharing	Comments
PID 14176					■		■	■						■	
PID 14325	■			■			■	■	■		■				
PID 14308	■	■						■							
PID 13557	■			■	■		■					■		■	
PID 13840	■				■		■		■	■	■				
PID 14366	■	■	■	■											
PID 17989	■		■	■	■	■		■	■					■	+ Cloud resources
PID 17410	■			■		■		■	■				■		+ Cloud resources
PID 17836				■	■				■	■				■	

Delivery and schedule

The delivery is delayed:

Yes

The deliverable was delayed from February 2023 to August 2023 to take into account all the results from the 2nd and 3rd Open Call projects.

Adjustments

Adjustments made:

None

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