Instruct-ERIC in EOSC-Life

Author list: Natalie Haley,[1] Pauline Audergon,[1] Laura Del Cano,[4] Jose Maria Carazo,[4] Edward Daniel,[6] Frank Von Delft,[2] John Dolan,[1] Maarten L. Hekkelman,[5] Robbie P. Joosten,[5] Lari Lehtiö,[6] José Márquez,[7] Anastassis Perrakis,[5] Marcus Povey,[1] Antonio Rosato,[8] Irene Sanchez lopez,[4] Rachael Skyner,[3] Carlos Oscar Sorzano Sanchez,[4] Joel L. Sussman,[9] Hans Wienk,[5] Claudia Alen Amaro,[1]

[1]Instruct-ERIC, Oxford House, Parkway Court, John Smith Drive, Oxford OX4 2JY, United Kingdom
[2]Diamond Light Source Ltd., Harwell Science and Innovation Campus, Didcot OX11 0DE, UK.
[3]OMass Therapeutics, Building 4000, Chancellor Court, John Smith Drive, ARC Oxford, OX4 2GX, UK
[4]Biocomputing Unit, National Centre for Biotechnology (CNB CSIC), Campus Universidad Autónoma de Madrid, Darwin 3, Cantoblanco, 28049 Madrid, Spain
[5] Oncode Institute and Division of Biochemistry, The Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, the Netherlands
[6]Faculty of Biochemistry and Molecular Medicine and Biocenter Oulu, University of Oulu, Oulu, Finland
[7]European Molecular Biology Laboratory (EMBL) Grenoble, Grenoble, France
[8]Consorzio Interuniversitario Risonanze Magnetiche di Metallo Proteine–CIRMMP, Via Luigi Sacconi 6, Sesto Fiorentino 50019, Italy
[9]Department of Chemical and Structural Biology, Weizmann Institute of Science, Rehovot, Israel

1. Introduction

EOSC-Life, an EC-funded H2020 project that lasted from March 2019 until August 2023 (GA 824087), brought together 13 European Life Science Research Infrastructures (RIs) to create an open, digital and collaborative space for data resources and data processing workflows in the European Open Science Cloud (EOSC).

2. Support of Instruct-ERIC pilot projects

As one of the RIs involved in EOSC-Life, Instruct-ERIC was actively engaged in the integration of pilot structural biology projects in EOSC-Life that would help create an open digital space for the life sciences. Through that scheme, seven research pilots led by Instruct partners aimed to develop solutions for the cloudification and FAIR handling of data resources and benefited from EOSC-life financial and technical support. These projects and their main achievements are described below.

2.1 Development of Fragalysis: a web-based platform for fragment-based drug discovery

Diamond Light Source; Instruct-UK
https://www.eosc-life.eu/d1/

The development of Fragalysis was part of a broad integrated project **"Demonstrator 1: European Open Science Cloud resources for Chemical Biology and Structure-Based Drug Discovery workflows**", which was done in collaboration between the RIs Instruct-ERIC, EU-OPENSCREEN, EATRIS, ELIXIR, and Euro-BioImaging ERIC.

Briefly, EOSC-Life supported the development of the Fragalysis platform by Diamond Light Source (DLS, Instruct-UK) to enable rapid access to data from X-ray crystallography-based fragment screening campaigns in a collaborative environment. Fragalysis is a ligand-centric platform that provides information regarding protein targets and their binding ligands. The development allows new screening data to be conveniently shared on the platform, whereas the release of structures from new fragment screens to the worldwide community can be done almost in real time, as could be demonstrated during the Covid19 pandemic. The sharing of data is supported by sophisticated data visualisation tools and Fragalysis now also facilitates the design of molecules. Finally, virtual screening workflows for fragment-based drug discovery were developed using Fragalysis and are available through Galaxy. This turned out very convenient to rapidly screen follow-up compounds that interfere with M^{pro}, the main SARS-CoV-2 protease with a pivotal role in viral replication and transcription.

2.2 Increasing the FAIRness of data and image processing workflows in Cryo Electron Microscopy

This project was part of the "Demonstrator 2 **Increasing the FAIRness of data and image processing workflows in Cryo Electron Microscopy"** and the EOSC-Life WP1 data deployment call.

Link: https://www.eosc-life.eu/d2/
Partners: Instruct-ES: CNB CSIC and Instruct-CZ: CEITEC
Plugin developed: https://github.com/scipion-em/scipion-em-empiar
Example EMPIAR viewer: EMPIAR-10891

Cryo-electron microscopy (Cryo-EM) has evolved extremely rapidly in recent years, with great advances in the quality of the instrumentation and the methods that are used for data analysis, to become a major technique to determine structures of macromolecules and their complexes. This demonstrator project aimed to increase the application of the well-known FAIR principles in Cryo-EM from the moment of acquisition, to make data Findable, Accessible, Interoperable and Re-usable. For doing so, this demonstrator developed a new method to handle Cryo-EM data with the Scipion image processing framework. This includes the deposition of data and workflows from an EM-facility to a public repository for raw data (such as EMPIAR), with the possibility of updating data and workflows throughout data analysis, before submitting the final high-resolution cryo-EM structures to the EMDB database.
The delivery of data and associated workflow to EMPIAR is a streaming transfer. Together with moving the resources from the facility to the public repository, there is a substitution of ownership from the facility to the user that is handled by the workflow engine. Practically, to achieve this ownership transfer, the facility requests the ownership change at EMPIAR, EMPIAR sends an email to the user with the request, and once the user confirms their ownership, the Cryo-EM facility loses access and the user can make updates (see Figure 2.2.1). As part of this demonstrator project, a viewer was integrated in EMPIAR that allows users to view the Scipion workflow used for the analysis of the data. A data viewer was added to allow users to look at the data at different steps of the analysis and to evaluate data quality. This demonstrator was granted subsequent funding from another EOSC-Life WP1 call, allowing the partners to continue to increase the FAIRness and the functionality of the deposition system. This follow-up project delivered a Common Workflow Language (CWL) output, now describing Cryo-EM data and workflow through an ontology, includes capability to deposit the workflows in WorkflowHub, as well as allowing automated submission.

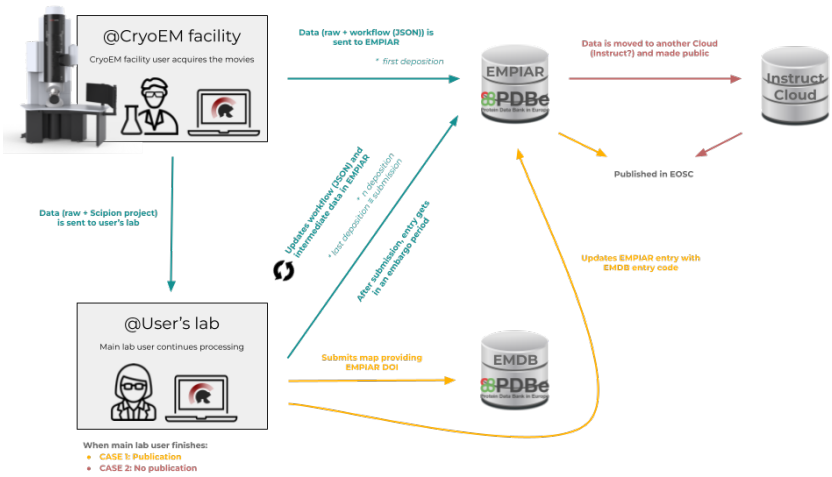

Figure 2.2.1: FAIR data and image processing workflows in Cryo-EM

2.3 Development of ICEBEAR: Macromolecular crystallography data management in the cloud

This project was part of the EOSC-Life WP1 data deployment call.

Partners: Instruct-FI: OULU, Instruct-IL: Weizmann Institute and Instruct-UK: DLS
Website: http://www.icebear.fi/

Publication: https://doi.org/10.1107/s2059798320015223

The software package IceBear (Integrated Crystal-data-tracking Enhancing Biochemistry Education And Research) provides better tracking of the samples, sample preparation and hands-off sample shipment to other research infrastructures for measurements. IceBear records individual crystallization drop experiments at home laboratories through a series of time-dependent images generated by various drop imaging systems and it captures all the relevant crystallization metadata. Modules have been added to exchange the metadata upon sample shipment with ISPyB, a Laboratory Information Management System commonly used for sample tracking and experiment reporting at synchrotron beamlines. Furthermore, IceBear allows inclusion of links to relevant ISPyB pages, as well as links to published raw diffraction data, PDB, Proteopedia and publications (https://doi.org/10.1107/s2059798320015223).

IceBear currently covers:
1) crystallization trials,
2) crystal fishing and shipping,
3) linking crystals, datasets, and PDB depositions for full chain of custody.

The objective of this EOSC-Life project was to make the IceBear software available on a virtual server. Such cloudification would make IceBear accessible for a wide user community, allows administration of IceBear instances by research sites without a need to setup and maintain computational resources, and enables sustainability through centrally managed updates and maintenance. It will also allow addition of new features that will help researchers with evaluating their results from crystallization and data collection.

Indeed, the developed cloud version makes IceBear readily available to a vast user community avoiding the need for in-house servers and resources and providing straightforward management of databases and centralized updating possibilities (https://doi.org/10.1107/s2059798320015223). To enable implementation, communication between the home lab instrument and the cloud instance was developed through uploader scripts running on a crystal imager. Cloud versions are now being incrementally set up for facilities already running IceBear. New instances can be established for interested crystallization facilities, while a standalone version will remain available through Instruct-ERIC for those preferring to maintain their own servers.

2.4 Development and refinement of PDB-REDO Cloud for FAIR experimental macromolecular structure models

Partner: Instruct-NL: NKI
Website: https://pdb-redo.eu
API: https://pdb-redo.eu/api-doc
Publications: https://doi.org/10.1107/S2059798323003595; https://doi.org/10.1016/j.jmb.2022.167599,

This project contained two parts, one from the EOSC-Life WP1 call (**"PDB-REDO Cloud: FAIR protein structures with deep versioning for scientific reproducibility and data provenance tracking"**) and one from the WP3 open call for user projects (**"PDB-REDO-cloud: A flexible and scalable engine for computational structural biology"**)

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 views of protein function rely on atomic detail in three dimensions, provided by experiment-based molecular structure models.

The PDB-REDO software pipeline[1] for crystallographic structure optimisation is 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 and also when many datasets need to be processed. The PDB-REDO databank contains over 170,000 'redone' Protein Data Bank entries on https://pdb-redo.eu.

At the start of the WP3 project, the PDB-REDO software pipeline was available as a webserver on https://pdb-redo.eu and a more flexible version as installable software for advanced users. However, local deployment appeared too challenging to most users, also owing to the fact that its local installation requires continuous maintenance to incorporate ongoing developments. Both issues were addressed by creating the PDB-REDO-cloud API which allows access to all settings of the pipeline, similar to a local installation, without any local deployment or maintenance. The pipeline was upgraded to manage and document user commands in JSON data structures, allowing calculations to be more easily set up and reproduced. The PDB-REDO server software was rewritten to provide output as a self-contained web component for easy integration in 3rd-party workflows. PDB-REDO-cloud was successfully integrated in two workflow managers of the CCP4, one of the major crystallography suites (https://doi.org/10.1107/S2059798323003595). In terms of impact, the number of PDB-REDO

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

calculations has increased from 8,000 in 2022 to over 10,000 in 2023, mostly through use of the new API.

In parallel to the new PDB-REDO pipeline, the PDB-REDO databank was completely overhauled in the context of the WP1 project. All output is now provided in FAIR data formats, notably mmCIF (https://doi.org/10.1016/j.jmb.2022.167599) and JSON. Provenance tracking in PDB-REDO was extended assuring that both the input data from the Protein Data Bank as well as all software used in the process is tracked with full version information to maximise reproducibility. In addition, a new *attic* for PDB-REDO entries was created allowing key (meta)data for previous versions to remain available whenever a PDB-REDO entry is updated. A search interface was added to create datasets with explicit references to versioned PDB-REDO data, which allows long-term recreation of original data sets. Owing to these new developments we experienced an increased use of the PDB-REDO website with an 8% rise of unique visitors from 2022 to 2023 and a 71% increase of website hits.

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

This project was part of the WP3 open call for user projects (**"Towards FAIR data for X ray-based structure-guided drug design"**)

Partners: Instruct-EMBL: EMBL-Grenoble and ELIXIR
Links: https://www.ebi.ac.uk/pdbe/; https://mmcif.wwpdb.org/

Recent technological developments at Instruct-ERIC facilities have enabled experimental X-ray crystallographic screening of very large compound libraries generating large amounts of data of key relevance for drug design. However, the manual deposition of data in public repositories in this context is very challenging and the current level of FAIRification in the field is limited. Through this project, new standardised models were developed to enrich PDB depositions with new classes of data and metadata, to ensure higher compliance with FAIR principles. The PDBe[2] PDBx/Macro Molecular Coordinate Information File (PDBx/mmCIF) was extended with new data and metadata categories from early steps of the screening workflow, in consultation with key stakeholders from Instruct-ERIC and iNEXT-Discovery. The new data and metadata classes are now part of the standard PDB mmCIF[3] used by the wwPDB.
Moreover, a new pilot metadata model was developed, capable of providing a full description of a prototypical fragment screening experiment. This model, based on the new wwPDB Investigation file format, makes it now possible to associate multiple PDB coordinate files, with corresponding data and metadata, into a single experiment. This new format provides links between the used fragment libraries, crystals, soaking experiments, raw crystallisation and X-ray diffraction data, and the resulting structural models represent a complete fragment screening campaign, potentially including hundreds to thousands of individual experiments, something that was not possible with the earlier wwPDB model (*see Figure 2.5.1*).

This new format is not even exclusive of X-ray based fragment screens but can be extended to any type of small molecule screening project, thereby making a major step towards the

[2] https://www.ebi.ac.uk/pdbe/
[3] https://mmcif.wwpdb.org/

availability and reusability of structure-based drug design data. The pilot investigation file from this project is currently being tested with prototypical fragment screening datasets from different European facilities.

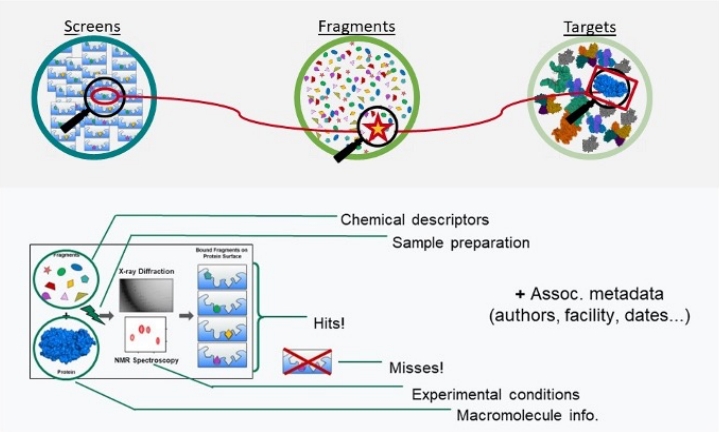

Figure 2.5.1: New Investigation file format to represent a full fragment screening campaign. It groups multiple entries under a single experiment, with appropriate links between small-molecules, crystals, datasets and the resulting coordinate files.

Finally, the project initiated development of software tools for automated (meta)data harvesting from commonly used crystallographic data management software to support automated deposition of structure-based screening data under the Investigation file format described above. These tools collect data from the CRIMS and ISPyB softwares and will support translation into the mmCIF format supported by the new Investigation file format.

3. Other Instruct-ERIC activities with EOSC-Life

3.1 "Exchange of Experience"-workshop on remote access

Supported by several Instruct-ERIC partners, the Instruct-Hub organised with EuroBioImaging-ERIC in the frame of the EOSC-Life WP9 training call the *"Exchange of Experience Workshop to Develop Remote Access, Virtual Teaching and Remote Training in Research Infrastructures".* This one-day virtual workshop brought together 124 participants from different Research Infrastructures from the life sciences and beyond to discuss challenges and exchange experiences regarding the development of remote access procedures and provision and remote user training.

3.2. EOSC-Life support for Cloud and compute resources

Three Instruct partners: Instruct-ES, Instruct IL and Instruct-hub benefited from a grant from the EOSC-Life WP7 call offering support for cloud and compute resources.

3.2.1 Cloud deployment support for the validation of Cryo-EM entries in EMDB (Instruct-ES, CSIC)
Publication: https://doi.org/10.1039/d2fd00059h

The Electron Microscopy Data Bank EMDB is the world's largest public repository of biological macromolecular maps solved by cryo-EM. The number of maps deposited is rapidly growing and their quality is a growing concern. CSIC recently developed a validation server reporting on each EMDB entry using most, if not all, validation methods published by the cryo-EM research community (DOI: 10.1039/d2fd00059h). The server is currently hosted by Instruct-

ES, but clearly lacks computer power to process the many thousands of entries in the database. With an EOSC-Life WP7 resources grant a group of machines in the AWS computing cloud was deployed for processing a large number of EMDB entries. The exposure to such amount of diverse data allowed the team to refine and improve their software for such needs.

3.2.2 Support for the development and the sustainability of the ARIA access managements system developed by the Instruct Hub and Proteopedia, the 3D-encyclopedia of proteins & other biomolecules developed by Instruct-IL.

3.3 Towards improved Tools and Workflow management:

EOSC-Life WP2, co-led by Instruct-IT, worked on identifying interoperable computational tools and workflows for European life scientists to build a roadmap. They also developed the WorkflowHub (https://workflowhub.eu/), a federated workflow registry thus facilitating the discovery and re-use of tools and workflows over a fragmented ecosystem.

3.4 ARIA developments in EOSC-Life

To support cross-RI access management, significant improvements were made to Instruct-ERIC's access administration software ARIA as part of Instruct's involvement with EOSC-Life. To support also the EOSC-Life open calls, ARIA's in-built "call" functionality was extended to align the moderation and review process with the standard e-Access pathway. Significant improvements were made for visit management, the access pathways for internal and external service access were unified into a common interface, and granular machine level workflow support was added. A new login system was deployed supporting various authentication mechanisms, including Life Science Login, allowing for ARIA's growth as a platform.

4. Conclusions

The Instruct-ERIC involvement in EOSC-Life was very productive, not only for networking but more importantly for delivering cloudified tools, workflows, databases and support the on-going transition of the Structural Biology community towards better and FAIRer handling of research outputs. We believe that the achievements that were made possible from this collaboration will be very useful for the structural biology community. In addition, they will strongly help to connect research from different life science communities and will show extremely beneficial for future endeavours with other European programs.