

Towards FAIR CryoEM workflows in EOSC

Monday, 2 November 2020 13:45 (20 minutes)

Scipion is an application framework developed by the Instruct Image Processing Center (I2PC) in Madrid to help the Structural Biology community to process CryoEM data. Scipion is a plugin-based workflow management system that integrates most relevant solutions in the community, allowing scientists to use them without bothering about formats and conversions.

CryoEM processing starts in the microscopes facilities where specific preprocessing workflows are run on Scipion to obtain the first quality data out of the microscope raw images as they are produced. When the acquisition finishes biologists leave the facility with both raw data and Scipion project and ideally continue processing using Scipion on their home labs or computing centers.

In the case of Instruct funded projects, scientists benefit from grants to obtain data in Instruct facilities but they also have to comply with Instruct Data Policy Management Plan, which states that all data obtained has to be made public after a certain embargo period, including both results and raw data.

Due to the recent evolution in CryoEM techniques and algorithms, processing demands more and more CPU and more importantly, GPU power, which implies that research institutions have to invest in expensive hardware and qualify staff to administer it. Cloud computing appears as a natural solution to overcome this problem, providing not only the best hardware but also packed images containing scientific software that allow to deploy a complete processing environment in a very short time.

With the latest in mind and in the context of former European projects, such as WestLife, MoBrain and other Instruct funded projects, I2PC developed ScipionCloud images and made them available in EGI AppDB and AWS. However, users still need to have knowledge on how to deploy and manage instances in the cloud and how to optimize the use of cloud resources.

This has been the main motivation to create a ScipionCloud service within EOSC that will use standard EOSC services to facilitate cloud deployment and user access as well as an optimized usage of cloud resources. This service is currently being implemented as a thematic service in the EOSCSynergy project and in the first phase will only be available to Instruct users, with the potential to be open to other users by means of access control. In parallel and in the context of the EOSCLife project a different aspect is targeted, FAIR data and workflows compliance while addressing the Instruct DPMP to guarantee data publication after the required embargo.

This complex scenario involves moving to containers technology (docker) and using existing EOSC core services such as EGI Checkin for access control and the Infrastructure Manager and EC3 for elastic cluster deployments. Data produced at the facility will be sent to a cloud storage where the service could access it and moreover, to ensure data FAIRness, an existing Scipion plugin used to deposit workflow and data in the EBI EMPIAR database will be enhanced to produce a CWL file with the workflow description and submit it packed into an RO-Crate to the EOSCLife WorkflowHub repository.

Primary authors: DEL CANO, Laura (CSIC); Prof. SORZANO, Carlos Oscar; Mr CONESA, Pablo

Presenter: DEL CANO, Laura (CSIC)

Session Classification: Workflow Management solutions