Showcase Integration Story:
PanOSC Dynamics of Biological Processes, Serial Crystallography (Photon) (8)

Demo for M24 review (13/14 June)

Science project/cluster	Dynamics of Biological Processes, Serial Crystallography/PaNOSC (8)
Contact(s)	Jordi BODERA SEMPERE, jordi.bodera@esrf.fr ILL?
Shepherd(s) or e-Infra contacts	From EGI: Giuseppe La Rocca, giuseppe.larocca@egi.eu
Brief description of story "Recipe"	Say what the scientist wants to do - the problem to be solved and how it can be solved with EOSC, ending up with the output and what the scientist can do with it Establish and develop a data processing platform for Serial Crystallography analysis on EOSC. Serial crystallography requires processing a large amount of data to solve the structure of a protein. Due to availability of beamtime, not all the data are collected at the same experimental facility. Users have to learn a lot of different transfer protocols as well as processing methods. Using our platform, scientists could login with their federated AAI to transfer data from public repositories (Zenodo, ESRF data portal just to give 2 examples). Data are then processed using the Jupyter interface to crystfel on EGI resources. The processing on the platform will be composed of: 1. Transfer data 2. Find peaks and evaluate noise (fundamental for reliable structure solution) 3. Indexing, using crystfel software 4. Integration, merging and export. MErging step could eventually involve multiple datasets from different sources and thus be the high point of the centralized platform. Visualisation of data statistics will be also provided in the notebook, while the final solution of the protein structure will be performed externally by the users, as this requires less resources and uses more established software. Analysis of crystallography data using software/workflows which are accessible to people outside ESRF. Demo can be done using data on data.esrf.fr and the processing can be done on the VISA tool (VM with python+cvmfs with data repos) VISA access is done via institutional AAI/Umbrella which is part of EOSC AAI. (The user if s/he doesn't have an

	identity can create one OR can access open data anonymously)
Input data and location	Data portals / public data ● ESRF: https://data.esrf.fr/ (already onboarded, but the URL has changed)
Core Resources needed if known (e.g.: EOSC Portal, EOSC AAI, Helpdesk)	EOSC AAI, EOSC Portal
Horizontal Resources needed if known (refer to SIS Athens presentation by Matt)	● ~~EGI Cloud Compute~~ ● ~~EGI Online Storage~~ ● Zenodo for output data/visualisation data
Output data and expected hosting location	● Output data will be a map file, containing the electron density map for the subsequent modelling. (on Zenodo) ● The map is then uploaded by the users on the protein databank database, once the modelling and refinement are done. (on Zenodo)
Notes and timeline	● UmbrellaID is already integrated in the EOSC Federation AAI and being tested. Needs to be onboarded: VISA tool

Status updates

please aim to provide an update at least every month

Date	Who	Notes on status/next steps/problems
21/10/2022	Gianluca Santoni, Jordi Bodera Sempere	Updated SIS
3/4/23	Matt, Jordi, Jean-Francois , Giuseppe	Discussed usecase and agreed on next steps: - onboard VISA (Jean-Francois) - check whether there is any interesting data that could be used for the demo is already open access, that doesn't require authentication (Jordi will check with Daniele) - Discussed about potential use of EGI

		Fedcloud/storage to address scalability but the problem is the lack of a workable data transfer service that works today.