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

D1.1 – EOSC Cloud feasibility assessment for Demonstrators

WP1 – Publishing FAIR RI data resources in EOSC
Lead Beneficiary: Fraunhofer, EMBL-EBI
WP leaders: Philip Gribbon and Helen Parkinson
Contributing partner(s): EMBL, BBMRI, EATRIS, ECRIN, EMBRC, FZJ, ERINHA, INFRAFRONTIER, UNIMIB, INRA, UNIVDUN, HMGU, CERBM, FLEMING, UOULU, CIRMMP, CSIC, VU, KNAW, UVEG, USMI, IMG, CNR, UNIMAN, UNITO

Authors of this deliverable: **Philip Gribbon, Helen Parkinson, Gesa Witt**

Contractual delivery date: **28th February 2020**
Actual delivery date: **31th Aug 2020**
H2020-INFRAEOSC-2018-2

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

Table of Contents

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

Executive Summary

EOSC-Life involves multiple complex cloud deployments of life sciences data, analyses and services for biomedical research users and infrastructures. This deliverable describes a cloud feasibility assessment performed on the first tranche of eight EOSC-LIFE Demonstrators by WP1, in cooperation with teams from WP's 2, 3 and 7. The feasibility assessment was followed by assignments of individual "Data Experts" to each Demonstrator according to the needs identified. A cloud deployment training event (with WP7) and a hackathon event to bring together the WP1 Data experts and Demonstrator teams was then organised to raise the overall technical capability and capacities of the WP1 supported projects.

Project Objectives

With this deliverable, the project has reached/this deliverable has contributed to the following WP1-specific objectives:

a. Development of cloud compatible FAIR-compliant data resources
b. Assessment of cloud feasibility of data/data resources and publication of these repositories in EOSC for data reuse

Detailed Report on the Deliverable

1. Cloud feasibility assessment

The original tranche of 8 RI-origin Demonstrator projects were identified during the preparation of the EOSC-LIFE proposal. The selected projects were highly diverse in terms of cloud maturity ranging from data resources which were already cloud deployed to those which were only accessible on a local institutional basis. A process was developed to establish the "Cloud feasibility" of the 8 RI- originated Demonstrator projects. This involved a review of the technical and implementation status of the planned cloud deployments (in cooperation with WP2, 3 and WP7).

1.1. Analysis of Demonstrator projects

A series of telephone conferences with demonstrator participants (Chaired by WP3 and coorganised by WP1, 2 and 3) was used to assess the cloud feasibility of the planned deployments. The analysis performed covered the demonstrators' aims, cloudification status and needs, and maturity / cloud feasibility. A summary of the analysis and discussions with the Demonstrator teams is shown in Table 1, each project was ranked on a low-medium-high scale,

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

this classification was designed to prioritise support and interactions with the demonstrator by WP1:

Demonstrator	Aims	Cloudification status and needs	Maturity / Cloud Feasibility
D1- Chemical biology and structure based drug discovery[1]	Dissemination (publically available data + publically available curation, evaluation and discrimination tools/workflows) of 3D structural data from small molecule or fragment-based screening in the context of early-stage drug discovery, including integration of corresponding bioactivity data from RI-linked data sources.	Essential elements of Demonstrators already deployed in the cloud. Fragment-based workflows, need to be for generic beyond Diamond. Will require close links with WP1, especially in terms of data and its management. There is a lack of Documentation both on Instruct and EU-OpenScreen tasks	**Medium-High,** but not using CWL, and dependent upon commercial packages
D2 - Increasing the FAIRness of data and image processing workflows in Cryo Electron Microscopy[2]	Share: i) raw data acquired from Cryo- electron microscope studies and; ii) results of image processing workflows in order to ensure reproducibility and tracing. Requirements are for improved protocols for data transfers and cloud storage	The existing AWC cloud deployed repository of related workflows (http://workflows.scipion.i2pc.es/) can be downloaded and uploaded. However, no formal identification of these workflows exists. The use of parallel FTP protocols and working with WP7 on cloud storage and data transfers is possible. Data sets are large so data compression may be needed. The results will be uploaded and made public via public repositories (EMPIAR), but an intermediate storage may be needed. Team has some experience in cloud services (AWS), but WP7 cloud resources specific to EOSC-Life are relevant.	**Low- Medium**
D3 - Large scale Metagenome	Facilitate metagenomics assembly, binning and	A CWL-based metagenomics pipeline has been implemented. The current	**High**

[1] https://drive.google.com/drive/folders/1CfXHHxKBGTwAziKkB2U9Sgfe2bzgBCT7

[2] https://drive.google.com/drive/folders/1Z8YCK9fXXNcnh33Ley-qYPYmfAHJoYN6

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

analyses in the cloud[3]	analysis using reproducible workflows described in CWL. Use the Demonstrator to realise deployment on different cloud providers, or using hybrid clouds. Further optimize the distribution and parallelization of jobs	Cloud deployment is via BiBiGrid and Ansible and is successfully deployed on two de.NBI Cloud sites (Bielefeld and Gießen). A WP1 Use case would be to place PFAM data on different clouds for access by different users. There is an opportunity for some license constraints to be removed if needed. Input from WP1 on the area of - orchestration of public DBs used for functional and taxonomic annotation (such as NCBI NR, PFAM, etc) and WP7 for infrastructure	
D4: Marine eukaryote genome portal - access to tools and data-flows for marine genome annotation[4]	Transfer genome annotations between closely related marine organisms using synteny. Provide a focal-point for information on tools, work-flows, digital resources, and services for marine genomic data. Promote Open Science and FAIR data practices, and the use of ontologies and meta-data standards (MIGS). Establish a new tool to update genome annotations for closely related taxa	Team is at a relatively early stage with limited examples of cloudified data sources. Currently, a pilot on the ELIXIR Belgium cloud exists and needs be to adapt this tool into a Galaxy workflow (also documented in CWL) and as a standalone tool (containerized) which can integrate annotations into the Orcae annotation platform.	**Low-Medium**
D5- Development of a novel configurable workflow for processing preclinical images and extracting meaningful information[5]	Establish a database of preclinical biomedical images covering both controlled and free access to preclinical images with searchable/findable tools for automated image processing and storage of quantitative image-derived data. Implement connections with Mammalian Phenotype	Several complex developments necessary: a raw data to DICOM format converter (previously run on MR images from Bruker); an XNAT uploader of multiple imaging DICOM datasets (diverse instrumentations and modalities); Image processing pipelines running in XNAT. Existing processes run on a PC, with substantial amount of human intervention (eg., ROI definition) licensing (eg., MATLAB)	**Medium** Infrastructure robust and OS. Clear plans for cloud portability, not yet implemented

[3] https://drive.google.com/open?id=1E2ZgV7PF1xS9ff2ULJNUVxwG5VtfLT4R
[4] https://drive.google.com/drive/folders/1Ph1Qm27Y6gAo84Z9Kpf3QbhRebBOXSrE
[5] https://drive.google.com/open?id=1b4pTFUxSZzYhwpUqLolVc3gBrVutUpLA

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

	Ontology and with Disease ontology databases	dependencies. Schemas and ontologies for workflow definition and persistent identifiers not yet established.	
D6 - Re-using published microscopy images to study nucleolus biology[6]	Aims for re-analysis of publicly-available images to gain insights into the nucleolus biology. Existing curation is insufficient to allow re-use by external users, due to a lack of metadata standardization and consistency. Data sources are IDR and Human Protein Atlas. Analysis with Cell-Profiler on Embassy Cloud	In this Demonstrator, the IDR is acting as a data repository, and is accessed via the cloud. The novel part is building pipelines and workflows. The initial workflow involves personal computers/ Jupyter notebooks, with a plan to build locally first and then port to the cloud.	**Medium** Data is identified along with a clear analysis pipeline. Currently not cloud based, or using CWL
D7 - An integrative analysis pipeline of genomic and transcriptomic human data for disentangling the genetic origin of a rare-disease in the context of the European Open Science Cloud[7]	Aim is to run sequence analysis workflows on controlled-access data from the European genome-phenome archive. Address research questions such as - *Which genomic variants could explain the observed phenotypic differences?* A cloud environment would avoid users to download data to local systems with benefits in terms of access control.	A related pilot was initially protoyped at BSC. Raw data (including all necessary meta-data) have been deposited in EGA. All phenotypic information (using PhenoTips) is available at RD-Connect and initial standard analysis based on WGS has been performed in RD-Connect. For the cloud solution, a manual access granting step from EGA will be needed. Appropriate data access control mechanisms, (existing process is manual) are necessary. Demonstrator will use CWL and workflows managers such as Galaxy and/or NextFlow. The services have partly been executed using ELIXIR Hybrid Cloud – mainly EBI and CSC, major implementation issues are not forseen.	**Medium-High**
D8: Taking Plant Omics Data through Annotation,	Establish an enhanced resource for the plant community to access molecular and phenotypic	Starting point is that current work involves lots of manual data curation, going back to data producers. Plant domain comprising molecular data	**Medium-High**

[6] https://drive.google.com/open?id=1MD38cS8m76Q3mSE4A2WIP1bxNPCQh3Sx

[7] https://drive.google.com/drive/folders/1UcCgsTNVPTDZkZBmQ08teP1meDt5UBdu

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

Acquisition, and Analysis to Application[8]	data sets, via interactive and API-driven search/retrieval. Use the MIAPPE (Minimal Information About a Plant Phenotyping Experiment) recommendations built on ISA-TAB as a container and Breeding API as a web service. Implement for solanaceae and maize (+1) data sets annotated to MIAPPE standards. Test, refine and extend of the MIAPPE standard to accommodate new data types as required. Development of tools for the analysis of data in the context of ontologies.	(gene expression, metabolomics) and phenotypic data sets exists across multiple different experiments and labs. Genome data will be annotated using the GBOL ontology. The API is based on BrAPI, (conform to MIAPPE standards). Tools used include: Mercator/MapMan, BrAPI REST API, ISA tools and R. Tool registries Bioconda, BioContainers, Galaxy tool shed. Following initial processing, data should fit into, automatic downstream processing via machine learning. At outset, tools and hosting environment for tool not in place, matrices exist as data files, metadata as spreadsheets. Ontologies are at different levels for the 3 plants.	

Table 1: Cloud feasibility assessment of the EOSC-LIFE Demonstrators

A survey tool was also used to collect a wide range of information from the demonstrator teams on data types, sensitivity of data, formats and cloud deployment status of the data resources planned to undergo cloud deployment. The extended survey results can be in project drive. The summarised version of the survey is shown in table 2.

Demonstrator Number	D1	D2	D3	D4	D5	D6	D7	D8
Short Name	Chemical biology	CyroEM Workflows	Meta-genomics	Marine Eukaryote	Preclinical image workflows	Image repository and mining	Rare Disease	Plant Omics
Public data	Y	Y	Y	Y	Y	Y	Y	Y
Private data	Y	N	Y	Y	y	N	Y	Y
Raw data	Y	Y	Y	Y	Y	Y	Y	Y
Sensitive data e.g. human	N	N	N	N	N	N	Y	N

[8] https://drive.google.com/open?id=1oMKzDuEVKIFjVzOSMo2BwnXfmO_hSyVd

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

Added value data	Y	Y	N	Y	Y	N	Y EGA	
Dynamic data access	Y	Y	N	N	N	N	N	N
Data type(s)	Structural Data	Stuctural Data- EM	Genomes for assembly	Genomes from marine organisms	Multiple imaging modalities X-ray, MRI and others	Images	Omics data for rare disease	Plant structured data
Data format	pdb, sdf, mol2, smiles, etc.	Image data	Diverse	Fasta, gff	Image data types, uses XNAT As backend	tiff	fastQ	ISA-Tab
Containerised data	Y	N	Y	N	N	N	Y	N
Containerization of tools/workflows (WP2)	Y	ND	Y	Y	ND	Y	Y	N
Data security	Y	Y	N	N	Y	N	Y	N
Data registry	Y	N	Y	Y	N	N	N	N
Established workflow(s)	Y	Y	Y	N	N	N	Y	N
Data accessed by	XCDB local repository defined by django models, accessed by Luigi parameter classes	Users, analysis pipelines, Scipion SW, EMPIAR or instruct database	CWL pipelines fully containerised	Galaxy workflows wished	python notebooks	python notebooks, custom analysis pipelines	NextFlow, CWL, Galaxy	Galaxy
Data resources accessed	WebPortals	direct from the microscope /EM facility	PFAM and others		WebPortal	IDR	NextFlow?	

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

Data resource cloud accessible	Y	Y	Y	N	N	Y	Y	N
Is workflow cloud-ready (WP2)	N	Y	Y	Y	N	N	Y	N
Curation	Y	Y	N	Y	Y	Y	Y	Y
Cloud instance	Y	Y	Y	N	N	Y	Y	N
Maturity / Cloud feasibility	Medium-High,	Low-Medium	High	Low-Medium	Medium	Medium	Medium-High	Medium-High

Table 2: Summary of Demonstrator technical status survey

Overall, there was a range of maturity status for various demonstrators which reflects the relative positions of the data resources within the Life science RI-infrastructure community. To address the needs of the Demonstrators and further facilitate their deployment plans, two events were planned and executed by WP1. Firstly, in collaboration with WP7, a training course in cloud deployment was delivered, bringing together the WP1 data experts and the demonstrators and secondly a “Hackathon” for Demonstrator teams and Data experts from the WP1 members. The training event will be reported by WP7 (EOSC-LS Cloud Observatory Report D 7.1).

1.2 Hackathon and training events to support Demonstrator cloud deployments and increasing overall maturity and feasibility

The event was held between 13th and 15th November 2019 at the Fraunhofer Forum in Berlin (meeting report is available on the project drive). Summary is given here

Attendance

- A total of 20 attendees received training in cloud deployment from WP7
- A total of 42 attendees took part in the hackathon, representing each Demonstrator

Hackathon Aims

Equip attendees to support their infrastructures/demonstrators with necessary technical skills to implement their proposals. This involved knowledge exchange, for example with image based methods share expertise on a thematic basis - with data experts from in-vitro, in-vivo and clinical imaging domains. Teams were encouraged to get involved in networking - build a team to help deliver Demonstrator and Internal projects based upon best practice solutions.

Expectations of attendees

Attendees were expected to have an intent to deploy a cloud based solution covering a Demonstrator or an RI internal data resource. Each came to the event with a use case and were requested to have access to data sets and or associated workflows where the projects were sufficiently mature. It was necessary for attendees to have completed the training event prior to the hackathon or to have prior experience that is equivalent. The event was structured in order to encourage participate in groups to enable knowledge exchange, to design their deployment scenarios for each demonstrator, to identify areas for data expert input and to move to the next phase of each demonstrator aligned with the Findable, Accessible, Interoperable and Reusable (FAIR principles). The overall aim was to spend 60% of the time coding.

Hackathon Topics

Hackathon topics were structured along the FAIR principles in order to have establish common feature elements across the various Demonstrator teams

Topic	Content
Findable	● Data Registry ● Identifiers ● Meta Data - format and implementation choices ● Use case specific problem solving
Accessible	● Cloud provision and access
Interoperable	● Metadata interoperability ● Access to ontologies and ontology mapping services. ● Writing data mapping workflows in Python ● Running ontology services in the cloud ● Repository deployment in the cloud - EuBI, link to analysis environments via gitlab analysis pipe; ● Meta data formats: Incl. Imaging - BioFormats and Flexible meta data
Reusable	● CWL workflows, different CWL execution engines and their peculiarities as an example of a difference between specification and real world ● JupyterHub ● Documentation; “bus factor” ● OMERO ● R, Galaxy. Python ● git, versioning and code packaging in general

Table 3 Topics for the WP1 Hackathon event

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

Hackathon Achievements

Each of the demonstrator projects summarised their achievements in the framework of FAIR, i.e. what technical advancements had been made at the event to make their respective data resources more Findable, Accessible etc. The main achievements for each Demonstrator are summarised on the google project drive. Overall significant progress was made in the majority of Demonstrators (D1, D2, D3, D5, D6 and D8). In some cases not all relevant Demonstrator members were not able to attend, so the efforts focussed on a subset of the FAIR headings in these cases.

Hackathon Feedback

A post event survey was carried out in cooperation with WP5. Overall rating of attendees for the hackathon was: Poor (0%); Good (36%), Very good (36%), Excellent (24%). In terms of the hackathon event directly allowing deployment of a cloud based solution, the survey results were: Agreed (20%), Maybe (36%) and No (44%). The latter results demonstrate the need for additional training and networking opportunities for these teams and these information and other feedback will be used to plan future events, in collaboration with WP1, 2 and 7.

2. Next Steps

The next steps for WP1 are to continue to provide ongoing support of Demonstrator projects during their operation through the direct participation of Data Experts and to work towards the next set of deliverables. At the first consortium Annual General Meeting, technical gaps in terms of implementation of cloud services across the EOSC-LIFE project were identified and discussion are ongoing to bridge these gaps through the recruitment of additional technical personnel.

3. Background

N/A

Abbreviations

Abbreviation	Full name
AWS	Amazon Web Services
OMERO	Open Microscopy Environment Remote Object https://www.openmicroscopy.org
CWL	Common Workflow Language https://www.commonwl.org/

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

PFAM	https://pfam.xfam.org/
NCBI	The National Center for Biotechnology Information https://www.ncbi.nlm.nih.gov/
MIGS	Minimal Information about a Genome Sequence
ROI	Region of interest
IDR	Image Data Repository https://idr.openmicroscopy.org/
MIAPPE	Minimal Information About a Plant Phenotyping Experiment https://www.miappe.org/
BrAPI	Breeding API
GBOL	Genome Biology Ontology Language
ISA-TAB	Investigation Study Assay Tabular format https://isa-tools.org/index.html
PDB	Protein Database
SDF	structure-data file
XNAT	eXtensible Neuroimaging Archive Toolkit https://www.xnat.org/
fastQ	http://maq.sourceforge.net/fastq.shtml
EGA	European Genome-phenome Archive https://ega-archive.org/

Delivery and Schedule

The delivery is delayed:

Yes

The delivery was delayed for a period of 7 months due to the COVID-19 situation leading to multiple members of the WP1 and Demonstrator teams being engaged in the preparation of cloudified datasets related to characterisation of the SARS-Cov2 virus and identification and validation of therapeutics and vaccines treatments. The work for the deliverable was largely completed prior to due date, however, the WP leads were assigned to Covid19 projects and were delayed in writing up the deliverable.

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

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.