D5.6: Evaluation report of service pilots

Author(s)	Giuseppe La Rocca (EGI Foundation)
Status	Final
Version	V1.1
Date	31/03/2019

Dissemination Level

X	PU: Public
	PP: Restricted to other programme participants (including the Commission)
	RE: Restricted to a group specified by the consortium (including the Commission)
	CO: Confidential, only for members of the consortium (including the Commission)

Abstract:

The overall objectives of this report are to:

a. Summarise the activities undertaken, and the available solutions and services that have been used by the Science Demonstrators to implement the pilot services.
b. Evaluate the cross-infrastructure usage and user experience through the executed service pilots.
c. Highlight best practices and lessons learned.
d. Drive and prioritise the integration of solutions and/or services to meet the functional and non-functional requirements requested by the Science Demonstrators.

The European Open Science Cloud for Research pilot project (EOSCpilot) is funded by the European Commission, DG Research & Innovation under contract no. 739563

Document identifier: EOSCpilot –WP5-D5.6	
Deliverable lead	**EGI Foundation**
Related work package	**WP5**
Author(s)	Giuseppe La Rocca (EGI Foundation)
Contributor(s)	
Due date	**31/03/2019**
Actual submission date	**31/03/2019**
Reviewed by	**Tiziana Ferrari (EGI Foundation),** **Joerg Meyer (KIT),** **Michelle Williams (GEANT)**
Approved by	**Mark Thorley (UKRI)**
Start date of Project	**01/01/2017**
Duration	**28 months**

Versioning and contribution history

Version	Date	Authors	Notes
0.1	18/09/2018	Giuseppe La Rocca (EGI Foundation)	Initial ToC
0.2	11/03/2019	Giuseppe La Rocca (EGI Foundation)	First draft
0.3	21/03/2019	Giuseppe La Rocca (EGI Foundation)	Comments from reviewers incorporated
1.0	28/03/2019	Giuseppe La Rocca (EGI Foundation)	Final version
1.1	31/03/2019	Mark Thorley (UKRI)	Final typographic proof-read and edit.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

EXECUTIVE SUMMARY

Fifteen science demonstrators, across different scientific domains, were selected with the purpose of providing insight on technical and policy needs, and cross-infrastructure integration requirements, and to get indications on how the EOSC Service portfolio should be structured. During the project lifetime WP5.4 translated these requirements, collected by WP4, into a set of service capabilities and workflows, underpinning technical solutions and gaps, and service providers. T5.4 contributed to help link research communities to new external providers and to the setup and management of technical infrastructures for co-design and piloting.

The overall objectives of this report are to:

- Summarise the activities undertaken, and the available solutions and services that have been used by the selected Science Demonstrators to implement the pilot services.
- Evaluate the cross-infrastructure usage and user-experience through the executed service pilots.
- Highlight best practices and lessons learned.
- Drive and prioritise the integration of solutions and/or services to meet the functional and non-functional requirements requested by the Science Demonstrators.

During the execution of the WP5.4 activities, some technical requirements and lessons learned were collected during the implementation of the pilot services. For each requirement, WP5.4 produced a recommendation that the developing EOSC should take into account. The full list of recommendations identified by WP5.4 is summarised below:

- Implement a distributed data and compute infrastructure providing high performance access to distributed big data infrastructures and mechanisms for data mirroring and caching, supported by high-speed network connectivity.
- Procure and offer EOSC as a federated infrastructure that integrates existing community resources (data, applications, software, storage and computing) and provides additional adequate capacity to scale up existing in-house IT infrastructures. Procure EOSC as a high capacity system that meets the demands of data intensive science.
- Provide easy to use environments like scientific gateways, VREs and data exploitation platforms as managed services that provide the required integration as a turn-key solution; make service descriptions semantically rich and discoverable; offer ready to use integrated bundles of services with low-barrier procurement processes.
- Provide a managed federated AAI solution to allow users to access services and resources from different infrastructure providers.
- Provide and sustain human networks through competency centres of experts working with scientific application developers in close cooperation.
- Provide support for running standard-based workflows.
- Promote a tightened integration between service and Infrastructure providers for helping research communities to plug into their working environments, the available EOSC services and resources to support their day-by-day work.
- Extend the FAIR concepts currently applied to data also to IT services. Propose a set of recommendations for making services FAIR, or to further enable services to make data FAIR.
- Include analysis of the underlying network requirements, specifically when designing the interoperability of services across sites and organisations.

1. INTRODUCTION

The EOSCpilot Science Demonstrators, selected following the criteria described in the reports D4.2 – Science Demonstrator Selection Process[1] and D4.3 – Consolidated Science Demonstrator progress report[2], represent the early adopters of EOSC and were selected to integrate diverse services and infrastructure in different scientific domains. These Science Demonstrators were selected taking into consideration the following criteria:

- Increase findability, accessibility (through interoperability) and availability of data and services.
- Reduce fragmentation between data providers, compute and storage providers, thematic service providers and network providers.
- Increase trust, usability and effectiveness of services in a hybrid environment.
- Provide the necessary services to conduct excellent science to resource-bound researchers, for higher education and professional training, and for the support of the research data value chain.
- Support of Open Science with new services for scholarly communications and repeatability of Science.

During the course of the project, all these Science Demonstrators, with the exception of DPHEP, received a central funding of 12 PMs for 1 year from WP4.2 to engage with the project, to adopt the solutions provided by WP5 and WP6 to meet their technical requirements and provide feedback on their use. The main objective of WP5.4 was to translate the Science Demonstrators requirements, collected by WP4, into a set of service capabilities and workflows, underpinning technical solutions and gaps, and service providers.

To achieve this goal WP5.4:

- Engaged with the shepherds from each Science Demonstrator via WP5 contacts nominated by the T5.4 Task Leader.
- Analysed the technical requirements from the Science Demonstrators.
- Identified the proper technical solutions and service providers to support the Science Demonstrator.
- Monitored the progress reports of the developing activities arranging, in collaboration with T6.3 task leader, monthly T5.4/T6.3 meetings. During these meetings all Science Demonstrators, especially the WP5 contacts, were invited to report any 'show-stoppers' during the implementation of the agreed work-plans. When necessary, to address the identified gaps, WP5 invited service/technology providers to introduce additional solutions/technologies that may not have been initially considered but that could be beneficial for them. A list of technical talks organized by T5.4, in collaboration with T6.3, to address the identified functional gaps are listed in Annex 1.

To facilitate the requirements gathering and promote the interactions with other work packages and Service/Resource providers, WP5.4 contributed to the organization of different meetings and workshops, including the EOSC Service Provider workshop[3], the EOSCpilot All Hands Meetings[4] and the Topical progress workshop on architecture and federated service management in EOSC[5]. All these meetings offered a good opportunity for helping the selected Science Demonstrators to explain technical aspects and issues they were facing, as well as share best practises that can be reused in the developing of other service pilots.

Towards the end of the project some of the Science Demonstrators were invited to further develop their work-plans for three additional months (January – March 2019) in order to move their pilot services into a

[1] https://www.eoscpilot.eu/content/d42-science-demonstrator-selection-process

[2] https://www.eoscpilot.eu/content/d43-consolidated-science-demonstrator-progress-report

[3] https://eoscpilot.eu/events/eosc-service-provider-workshop-13-15-sept-pisa

[4] https://eoscpilot.eu/events/eoscpilot-all-hands-meeting

[5] https://eoscpilot.eu/events/topical-progress-workshop-architecture-and-federated-service-management-eosc

pre-production phase. This deliverable reports on the selected Science Demonstrators and includes feedback and user experiences from the execution of the service pilots, lessons learnt and recommendations for shaping the EOSC architecture and drive the technological evolution of the EOSC services that meet the functional and non-functional needs of researchers.

This document is organized as follows:

- Section 2 outlines the overall evaluation of the selected Science Demonstrators.
- Section 3 presents a list of EOSC services used by each Science Demonstrator to implement a domain specific pilot service. For each Science Demonstrator a list of technical requirements raised during the running of the pilot services is presented, as well as the recommendations to mitigate/resolve these issues.
- Section 4 draws some conclusions.

2. EVALUATION OF THE SELECTED SCIENCE DEMONSTRATORS

During the course of the project, fifteen Science Demonstrators, across different scientific domains, were selected over three rounds to implement specific use cases. The selected Science Demonstrators were grouped into three sets as follows.

2.1. First set of Science Demonstrators

During the first year of the project, EOSCpilot has been working with the first round of scientific demonstrators. The first set of scientific demonstrators gave a good idea of the broad range of communities and technologies that the project is working with:

- **Environmental & Earth Sciences** – The ENVRI Radiative Forcing Integration science demonstrator focuses on the integration of data and services between the edges of environmental research infrastructures.
- **High Energy Physics** – The Data Preservation for High Energy Physics (DPHEP) aims to bring the Open Data Portal, developed by CERN to provide data preservation (including software and documentation) for the High Energy Physics community, to the wider scientific community. The solution will be deployed on generic cloud services and proved to work with sample preserved High Energy Physics data before being prepared for non-HEP data.
- **Social Sciences** – TextCrowd, a text mining solution developed for the humanities and social sciences communities that enables a semantic enrichment of text sources and make it available on the EOSC.
- **Life Sciences** – Pan-Cancer has developed the Butler genome analysis framework on the EMBL-EBI Embassy Cloud and its EOSCpilot demonstrator is working to deploy it on generic cloud services. This will enable the analysis framework to be used by more communities.
- **Physics** – The Photon & Neutron science community is interested to improve the community's computing facilities, using institutional and public cloud services, by creating a virtual platform where data and analysis tools can be made available to scientists all over the world.

The final evaluation status of the first set of Science Demonstrators is described in the following sections.

2.1.1. ENVRI Radiative Forcing Integration

Objective of the pilot

The core objective of the ERFI Radiative Forcing Integration (ERFI) Science Demonstrator, supported by ICOS ERIC and the IS-ENES Research Infrastructures, was **to build a prototype data and metadata integration framework which can be used to uniformly address and access the Observational and Climate Modelling Environment research Infrastructures**. This prototype was supposed to demonstrate the capability to interconnect heterogeneous datasets from the ICOS and IS-ENES RIs. The scientific focus was on dynamics of greenhouse gases, aerosols and clouds and their role in radiative forcing.

Technical capabilities

To support the creation of the **data integration framework**, and make the initial parts of multi-petabyte climate model data archives hosted at DKRZ and IPSL accessible to the ICOS community, the following technical capabilities were identified:

- Data Management - Access to Onedata Data platform provided by EOSC to offer transparent data access to the IS-ENES datasets.
- Compute - Access to the EGI cloud compute infrastructure to allow members of the ICOS RI to run climate models on the datasets hosted at DKRZ.

Implementation

To achieve this goal, the following solution was proposed:

- Accessed to the reliable and scalable EGI FedCloud infrastructure.
    - On-demand allocation of VM based on common Linux distribution with at least 2 vCPU cores, 8GB of RAM and 500GB of block storage available in the /mnt partition.

    - Access to the EGI FedCloud Infrastructure was enabled by WP5.4.
- Configured the EGI FedCloud resources to use Onedata[6], the software stack developed in the context of the INDIGO-DataCloud project[7], to **provide a transparent access to ICOS and IS-ENES datasets via POSIX interface**.
    - Cyfronet, one provider of the EGI Federation, set up 2TB of storage on a separate pool on Ceph and configured an instance of oneprovider service for this use case. Restrictive policies at DKRZ prevented to expose the data repository as oneprovider.
    - The provision of the Onedata storage space was enabled by WP5.4.
- Transferred the climate data hosted at DKRZ and IPSL in the EGI FedCloud Infrastructure using synda[8], a command line tool to search and download files from the IS-ENES archives. The tool, running in one of the EGI VM, was used to populate the 2TB of volume space allocated in Onedata using standard protocols such as: http:// and gridftp://. The datasets managed during the pilot were not sensitive and could be freely accessed.

The **metadata integration framework** was proposed to harvest the metadata from the IS-ENES data repository using the EUDAT B2FIND service. Unfortunately, there were no endpoints where climate datasets were accessible via B2STAGE and setting up of a dedicated B2STAGE instance for IS-ENES datasets was considered out of scope for this pilot since this would involve iRODS integration into the existing infrastructure. For this reason, the following approaches was proposed:

- Write a translator component querying the IS-ENES metadata search API and exporting the metadata for the (initial) agreed data subset to Onedata.
- Export data based on IS-ENES metadata search API.

During the implementation of the **data integration framework**, the solution for the data integration framework proposed by WP5.4 had proven to have showstoppers for the ICOS RI (e.g., the service crushed upon trying to visualize and download a few hundred files after having uploaded them as a single tar file). ICOS case provided the means and requirements to increase the scalability of the solution, which resulted in a series of releases. The release rc.13 was used to configure the 2TB of volume space for the ERFI Science Demonstrator and populate the space with the IS-ENES datasets coming from the DKRZ data repositories. The volume space, with the IS-ENES datasets, was mounted in one VM running in the EGI Federated Cloud infrastructure and made accessible to the ICOS community to simulate fluxes of carbon and water using the LPJ-GUESS model. LPJ-GUESS simulated water and carbon fluxes were compared to in-situ measurements from ecosystem flux towers (Fluxnet 2015 dataset) using monthly time resolution.

During the climate model simulations the connection with the Onedata volume space broke down several times making the creation of monthly averages from NetCFD files time consuming. From a technical standpoint, the connection to the IS-ENES datasets stored in the Onedata provider was limited by the network speed available between the VM, where the block space was mounted and the simulations were performed, and the oneprovider where datasets were physically stored. To mitigate this issue, WP5.4 proposed to move the simulation of the LPJ-GUESS model closest to where datasets were available. Moving the simulation in a cloud computing provider close to the oneprovider significantly improved the data access performance. The high-level architecture of the data integration framework is shown in Figure 1:

[6] https://onedata.org/
[7] https://www.indigo-datacloud.eu/
[8] https://github.com/Prodiguer/synda

Figure 1 - The data integration framework for the ERFI service pilot.

Main Achievements

- The pilot tested the data interoperability between institution hosting datasets (IS-ENES) and institution responsible for simulations (ICOS).
- IS-ENES climate models data archived on the ENES Data Infrastructure (DKRZ node of Earth System Grid Federation) were transferred to a OneData EGI node.
- From the EGI FedCloud Infrastructure datasets were fetched from the original data repository into the Onedata volume space.
- ICOS members processed the IS-ENES datasets and extracted data from historical and rcp45 experiments to build input files for a land surface model (LPJ-GUESS).
- LPJ-GUESS simulated water and carbon fluxes were compared to in-situ measurements from ecosystem flux towers.

The lessons learned from the pilot project are:

- **Allocate the needed resources in advance** in order to start the implementation of the planned activities according to the work-plan.
- **Need for increased scalability in distributed access to data.**
- **Network** performance were not well defined in the beginning of the pilot project.
- Need for a **policy** and **processes framework** that facilitates the access to e-Infrastructure by new user communities.

2.1.2. Data Preservation and Re-Use through Open Data Portal (DPHEP/WLCG)

Objective of the pilot

Research Infrastructures such as the ones on the ESFRI roadmap and others are characterized by the very significant data volumes they generate and handle. These data are of interest of thousands of researchers across scientific disciplines and to other users via Open Access policies. Effective data preservation and open access for immediate and future sharing and re-use is a fundamental component of today's research infrastructure and Horizon 2020 actions. In this context, European research stakeholders make increasing use of cloud services to effectively handle such data.

The Data Preservation and Re-Use through Open Data Portal (DPHEP) Science Demonstrator, supported by CERN, aimed to demonstrate "best practices" regarding the long-term preservation and re-use of HEP data,

including documentation and associated software. **The scale of the data volume is one of the key challenges for this Science Demonstrator.** Even if only a small subset of LHC data is released as Open Data, the total volumes (hundreds of PB to tens of EB) mean that several to many PB will be released in the coming years. Other key issues the pilot aimed to address, are: the **"Findability" of data** by providing descriptions that are meaningful to non-experts, and the **"Re-usability" and "Reproducibility" of results in subsequent years.**

Overall, the DPHEP Science Demonstrator tried to deploy services to address the following key points:

- Reference digital objectives in trustworthy (aka certified) repositories via Persistent Identifier (PID).
- Promote scalable "digital library" services where documentation is referenced by Digital Object Identifier (DOI).
- Promote the adoption of a versioning file system to capture and preserve the associated software and needed environment.
- Support the use of a virtual environment.

Technical capabilities
For this pilot, the following technical capabilities were identified:

- Sharing & Discovery:
  - B2SAFE - Distribute and store large volume of data based on data policies.
  - B2SHARE - Store and publish research data.
- Compute:
  - CernVM - Virtual Software Appliance for the participants of CERN LHC experiments. The Appliance represents an extensible, portable and easy to configure user environment for developing and running LHC physics software both locally, on Grids and Clouds, independently of Operating System software and hardware platforms.
  - CMFS - The CernVM File System provides a scalable, reliable and low-maintenance software distribution service.

Implementation
A B2SAFE instance hosted at CINES was configured as Trustworthy Digital Repository (TDR)[9] for the HEP datasets. During the course of the project, the data ingestion/replication of small datasets were demonstrated. Data retrieved from CERN was ingested into CINES repository and replicated to CINECA for open access. The connection between CERN and CINES was fully operation with an average speed of 800Mbits/sec. Data was referenced by a Persistent Identifier (PID). Access to the B2SAFE instance was enabled by WP5.4.

Software was preserved using CernVM and an instance of CVMFS provided by STFC. Access to the CVMFS instance at RAL was enabled by WP5.4. The CernVM was used to allow users to run analysis in Grid and HTC environments.

A B2SHARE instance was used to store the documentation. Access to an instance of B2SAFE at CINECA was enabled by WP5.4.

Main Achievements

- An initial pilot service capable of storing and preserving Open Data was implemented during the project.

The high-level architecture adopted to demonstrate the software preservation is shown in Figure 2, and the Data Archive solution implemented by the pilot is reported in Figure 3. Unfortunately, by the time of writing of this report, the services interoperability is still an open issue. There is no way for the end-users to use the same token to access resources/services from different service providers.

[9] https://www.coretrustseal.org. Note that the FAIRSFAIR project can be relevant for EOSC with respect to improve the trustworthiness of data repositories.

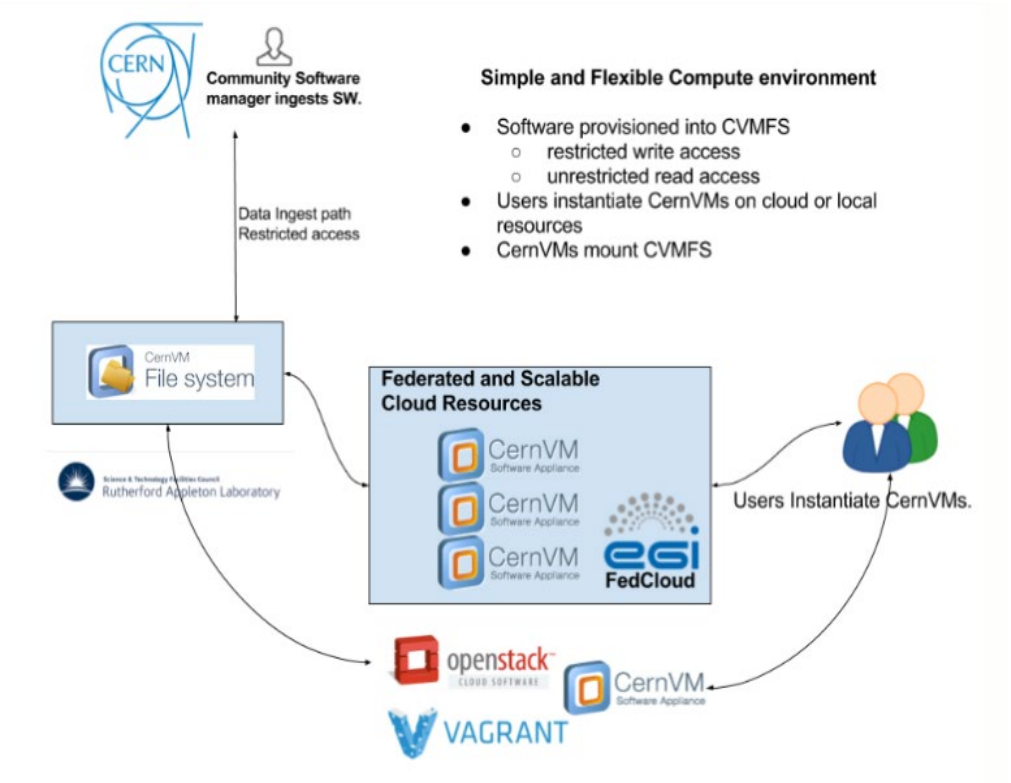

Figure 2 - The software and environment preservation solution in the DPHEP pilot.

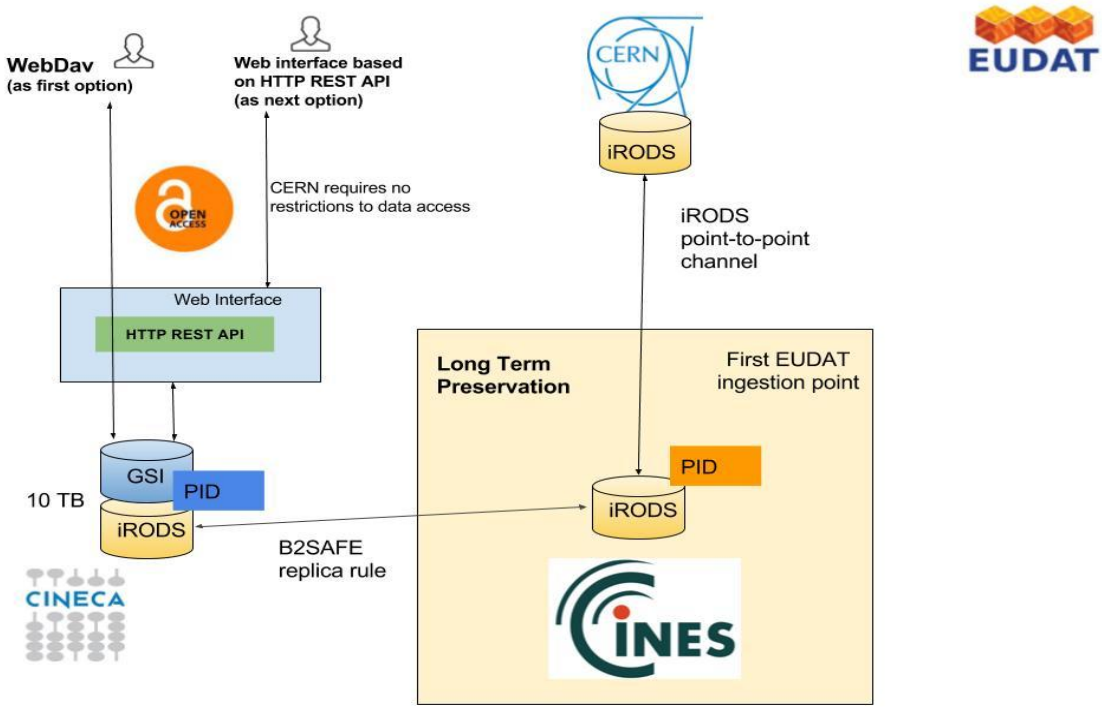

Figure 3 - The implementation of the data archive solution in the DPHEP pilot

2.1.3. Collaborative semantic enrichment of text - based datasets (TextCrowd)

Objective of the pilot

Archaeological documentation are largely based on texts (e.g. excavation diaries, reports, surveys and grey literature). Unfortunately, only the general scope of texts are described in the repositories. Indexing and metadata enrichment of text can be done manually but this a very time-consuming task and it could be unreliable. EOSC can contribute to bridge this gap by offering a cloud-based data infrastructure which can be used by researchers to enable the semantic enrichment of text sources, extract structured semantic information and create standard metadata for archaeological reports.

The main objective of the Science Demonstrator was to **implement a pilot service that archaeologists can use to process text reports and produce, through natural language processing, semantic metadata**. The resulting metadata can be easily integrated with information coming from other domains, establishing a machine-learning scenario. Better indexing means also more researchers can benefit from the data.

Technical capabilities

For this pilot, the following technical capabilities were identified:

- Compute - Access to the EGI cloud compute infrastructure to host the web services needed by the pilot to perform the semantic enrichment of text. These include, machine learning models for Natural Language Processing[10] (OpenNLP), Named Entities Recognition[11] (OpenNER) and the GATE[12] open-source toolkit.
- Processing & Analysis - D4Science[13], a platform for data validation, data enrichment and efficient data analysis.
- Sharing & Discovery - Access to secure and trusted data exchange service to store datasets.

Implementation

To support the semantic enrichment of texts GATE, an open-source toolkit developed by the University of Sheffield, was used. During the course of the project this software toolkit has been further customized in order to run as a cloud service and extract semantic information from Italian archaeological texts. To facilitate the interaction of archaeologists with this software toolkit, since no transparent and easy-to-use GUIs were available, WP5.4 liaised with the CNR provider which is the main responsible of D4Science. D4Science offers a number of services, and Virtual Research Environments (VREs), for users willing to acquire concrete understanding of the major features and capabilities. In this specific context, this interoperable framework was used to set-up a custom VRE including all the services needed by the digital heritage community, including the GATE toolkit, the OpenNLP and OpenNER web services.

A view of the VRE set-up for the archaeology community is shown in Figure 4:

[10] https://opennlp.apache.org/
[11] https://www.opener-project.eu/
[12] https://gate.ac.uk
[13] https://www.d4science.org/

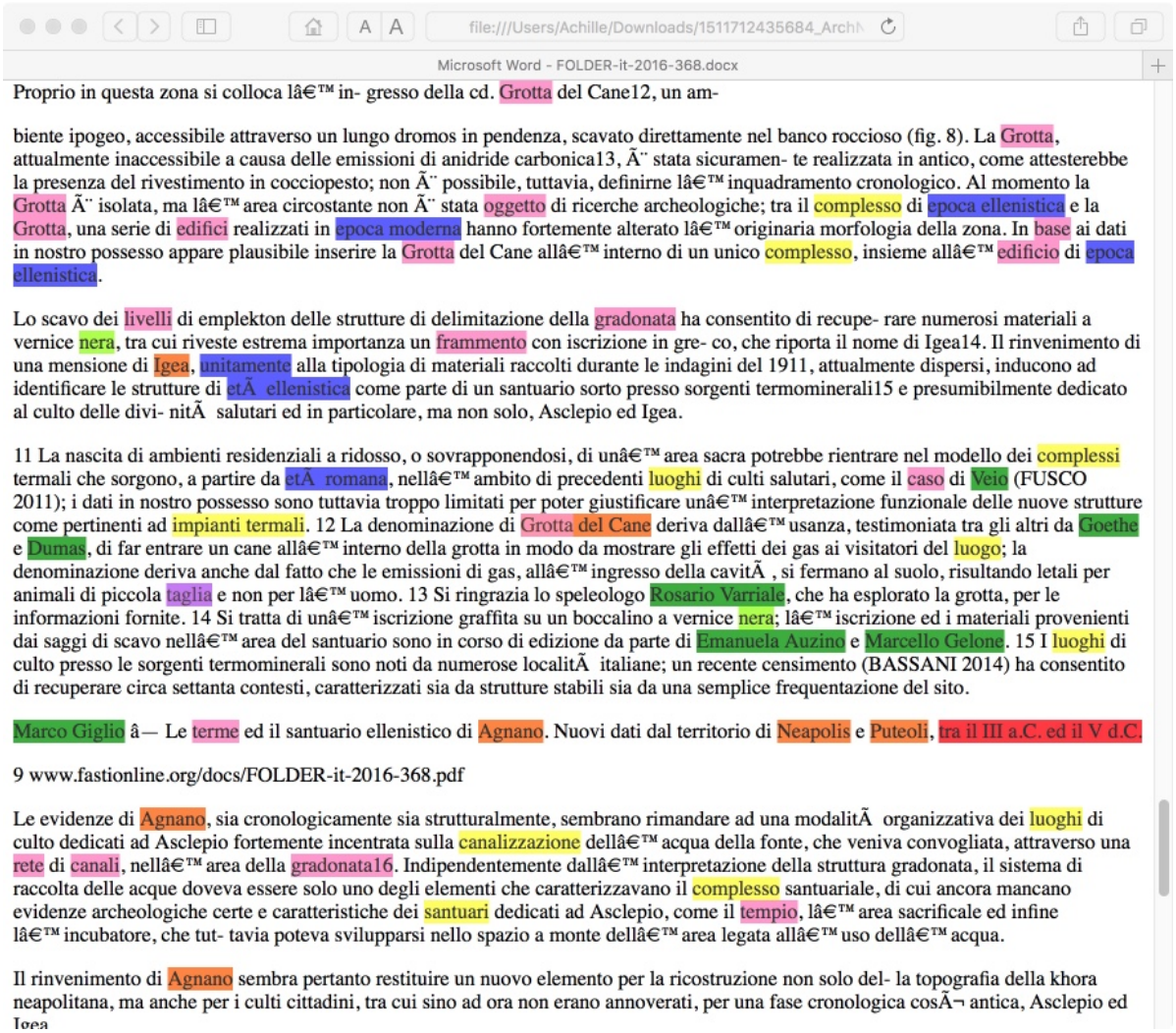

Figure 4 - Text analysis through natural language processing.

Achievements

- Semantic enrichment of text documents (e.g. archaeological excavation).
- Semantic information produced by the pilot are in CIDOC CRM format to be interoperable with other Cultural Heritage semantic data.
- Enrichment of the metadata schema to improve indexing, discoverability, accessibility and reusability.
- Metadata stored in various registries for easy findability and accessibility (e.g.: ARIADNE Registry, PARTHENOS Registry).
- Set-up of a Virtual Research Environment for the member of the community based on the D4Science framework.
- Extended support for Machine learning capabilities to integrate information coming from other scientific domains.

Further development during the project extension

During the funding period the Science Demonstrator reached an excellent level of development, attracting interest for potential applicability in different ways and fields. For this reason, during the extension of the WP5.4 activities, TextCrowd was selected to further improve the interoperability with other Science Demonstrators (e.g.: VisualMedia) and register the resulted solution in the EOSC Marketplace and EOSC portal. One of the key-points of the new submitted work-plan was to use the EUDAT B2DROP service to store the text reports produced during the natural language processing and scale-up the text processing with

additional cloud resources. The technical details to extend the TextCrowd VRE, which is based on D4Science framework, to deposit text reports into the EUDAT B2DROP instance was addressed in a dedicated meeting. The meeting between the scientific community and the B2DROP experts was enabled by WP5.4.

During the last three months, the development of TextCrowd has moved in different directions to face technical and integration issues, and to improve the quality of its framework. Integration with other EOSCpilot Science Demonstrators was evaluated. In particular, interoperability was established with the VisualMedia pilot, developed by ISTI-CNR for automatic metadata extraction from controlled lists or textual documents for 2D and 3D models, and with the Frictionless Data Exchange, intended for synchronization of annotated documents repositories across the Web. A possible interoperability with external services, like B2DROP, B2SHARE and OAI-PMH compatible repositories, has also been envisaged but not yet implemented.

Tests for its applicability outside the EOSC framework was also carried out: the tool was deployed as part of the NLPhub[14] framework, a Virtual Research Environment developed under the PARTHENOS[15] initiative. Overall, the initiative aims to combine the potential and the specific features of different NLP tools in order to create advanced software pipelines for processing streams of documents of various language, topic, provenance, in a recursive way.

From a technical standpoint, a new way to improve the TextCrowd algorithms, and make them more efficient, was analysed. To empower the tool, adapt it to other scientific domains and extend its capabilities to other languages, the new prototype will be equipped with innovative Supervised Machine Learning algorithms (Conditional Random Fields and Neural Networks) to learn an entity recognition model from examples of documents annotated by experts. The entity recognition model can efficiently process streams with large amounts of document, populating a structured knowledge base. At any moment the learned entity recognition model can be used to drive the selection of the documents to be annotated by experts, so as to better exploit the human annotation effort (Active Learning).

The new prototype is still not ready to be deployed in a cloud infrastructure, nevertheless, the developments of TextCrowd will also continue beyond the specific scope of the EOSC initiatives. Within the ARIADNEplus[16] framework, for instance, it will play a fundamental role for indexing excavation reports and enriching the ARIADNEplus Catalogue. Planned improvements will concern user interfaces and the implementation of user profile management facilities, extension of the interoperability with external services (e.g.: B2DROP, B2SHARE and OAI-PMH compatible repositories), the extension of its functionalities to other domains to analyse existing corpora.

The lessons learned from the project are:

- The D4Science VRE framework is a promising model to interoperate with different services and present them in transparent way to the users.
- **Create an efficient interface between the science community and the e-infrastructure is a necessity.** It is also necessary to preserve the valuable assets developed by research communities in the process of porting them in the EOSC framework without upsetting their functionality. This may require more flexibility than the one envisaged so far, and it is the EOSC that must adapt to the research needs, not vice versa.

[14] http://nlphub.cilenis.com/

[15] http://www.parthenos-project.eu

[16] https://ariadne-infrastructure.eu/

2.1.4. Pan-Cancer Analysis in EOSC

Objective of the pilot

The Pan-Cancer Analysis of the Whole Genomes (PCAWG)[17], supported by EMBL-EBI, aimed to **set-up a cloud-based workflow system**, using public and private clouds. This workflow system can be used **to execute cancer genomics analysis that, ultimately, can contribute to improve patient health care and cancer outcomes.** To deliver standardized pipelines for processing patient genomes via Docker containers and Common Workflow Language (CWL) and overcome the challenges of orchestrating analyses of thousands of human genomes (~10,000) on public and private clouds it was used Butler[18], the computational framework developed in the context of the PCAWG project.

Technical capabilities

For this pilot, the following technical capabilities were identified:

- Compute - Access public and private cloud infrastructures to perform cancer genomics analysis on whole genome DNA sequencing data.

Implementation

During the first months of work, WP5.4 approached several potential cloud service providers, both partners within the project and organizations external to the project, to elicit interest in contributing resources to Pan-Cancer. From the service providers approached, twelve were excluded as they were either unable to provide the required amount of resources, did not run the required software, or declined for other reasons. Four service providers were able and willing to provide resources - Amazon Web Services (AWS) and Compute Canada[19] who were interested in providing resources for free, and two EGI members (CSIC and Cyfronet) who were interested in providing resources on a pay-per-use basis.

At last, WP5.4 managed to identify two additional cloud infrastructures interested to support the Pan-Cancer Science Demonstrator. The three cloud infrastructures (EMBL-EBI Cloud Embassy[20], Compute Canada, and Cyfronet) provided, to this demonstrator, **a total of 2500 compute cores, 11 TB of RAM, and 1.5 PB of disk storage**. In more detail, the total resources provisioned are reported in Table 1:

Provider	Cloud Middleware	vCPU cores	RAM	Storage space	Others
CYFRONET	OneNebula (access via OCCI standard)	700	2.6 TB	4.9 TB	32 floating IPs
ComputeCanada	OpenStack	1000	3.9 TB	62 TB of volume + 200TB of NFS storage	
EMBL-EBI	OpenStack	800	4.5 TB	1.4 PB	

Table 1 - The computational infrastructure allocated for the PanCancer Science Demonstrators.

Subsets of the data from the ICGC Pan Cancer Analysis of the Whole Genomes (PCAWG) project was staged at these clouds and shared over NFS. The solution used to implement data sharing of the genomes datasets

[17] https://www.sevenbridges.com/case-studies/pcawg/
[18] https://github.com/llevar/butler
[19] https://www.computecanada.ca/
[20] https://www.embassycloud.org/

across different cloud infrastructures was the Onedata[21] software stack developed by Cyfronet during the INDIGO-DataCloud[22] project and further improved, in terms of stability and performance, in the context of the HNSciCloud[23] project. Oneprovider instance was configured at EMBL-EBI to pre-cache ~30TB of datasets. 100TB of data-sets were also tested in ComputeCanada cloud infrastructure.

Achievements

The Butler framework for scientific analysis on the cloud was set up and configured to manage the full analysis lifecycle, from creating virtual infrastructure, to configuring networking and security, to scheduling and executing workflows, to operational management during the execution phase. **Over 800 high-coverage whole genome samples (~200 TB of data) were analysed (genome alignment and variant calling) via Butler on the three cloud infrastructures over the course of the project.** All of the resultant configurations and analysis settings were made available via GitHub for use by the scientific community.

The high-level overview of the multi-cloud architecture used to run Butler is shown in Figure 5:

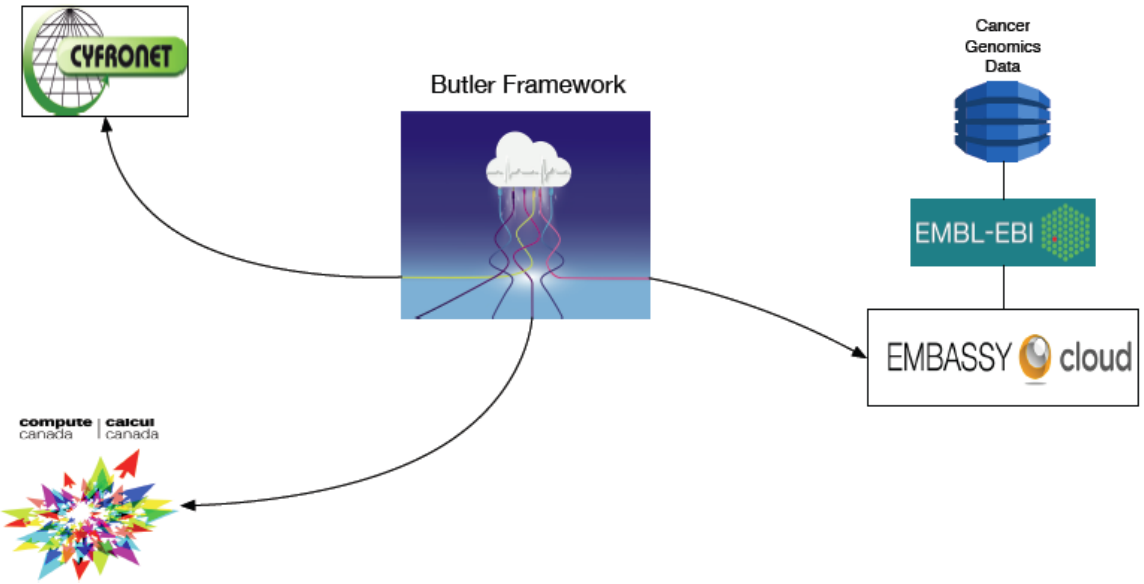

Figure 5 - Overview of the multi-cloud infrastructures used by the PanCancer Science Demonstrator.

As further recommendations provided by the Science Demonstrator, it would be beneficial if EOSC can:

- **Provide a procuring framework to** facilitate **the allocation of computational resources**. PCAWG, a study with <3000 cancer patients, has generated ~1PB of data and utilized ~16000 compute cores for processing. Modern genomics studies aim to process on the order of 100,000-1,000,000 genomes, thus requiring 2-3 orders of magnitude more computational resources to accomplish.
- Expose a **centralized identity and access management solution is desirable** to establish a consistent level of security necessary for handling of private genomic research data within the context of EOSC, and to simplify information systems that are distributed over multiple computing environments.
- Promote the adoption of **high**-performance **shared network storage and object-storage solutions to alleviate storage performance concerns when it comes to processing ever-larger genomics data sets.**
- Provide a **catalogue of datasets would be beneficial** as the inability to move cloud-scale datasets will largely dictate environment choice by scientific groups in the future.
- Offering other Platform as a Service type services, such as databases and queues, would be beneficial.

[21] https://onedata.org/

[22] https://www.indigo-datacloud.eu/

[23] https://www.hnscicloud.eu/

2.1.5. Research with Photons & Neutrons

Objective of the pilot

The Photon Neutron Data Science Demonstrator aimed to demonstrate the use of EOSC resources to support the technical requirements of a specific demanding application. CrystFEL[24] is a software suite created to address the processing needs of serial femtosecond crystallography (SFX). The raw data involved during the analysis ranges between 1-100TB resulting very difficult to be moved around. Efficient analysis of large datasets requires bringing together the data with workflows and the need of computing resources in an integrated platform.

Technical capabilities

For this pilot, the following technical capabilities were identified:

- Compute - Access cloud compute infrastructures to deploy and test software used by a large community in structural biology at Free Electron Lasers and Synchrotrons.
- Storage - Integration of middleware solutions for mass storage systems like dCache.
- Processing & Analysis - Adopted Jupyter Hub technology to wrap the scientific applications into Jupyter kernels.
- Networking - Local VPN was used to provide access to VMs and container in multi-cloud environment. Possible interest to use the GÉANT Multi-Domain Virtual Private Network (MD-VPN).

Implementation

During the course of the project a significant amount of time was spent to containerize the application, using Docker container technology. **The resulting software suite is now more cloud-compliant** thanks to testing that took place on the EGI Federated Cloud infrastructure under the technical support of WP5.4. A first PoC of the pilot service, which integrates datasets with relevant metadata, workflows and functions, was implemented. This PoC used the local cloud resources provided by DESY, examples of datasets provided by XFEL and ESS and some examples of functions and workflows to analyse and visualize data. With the local cloud infrastructure going to production, in the context of the project, WP5.4 considered the opportunity to further support the implementation of the service pilot during the 3-months extension of the WP5 activities in 2019 to move the pilot into a fully operation and deployable service.

The goal in this extension phase was to consolidate and extend the services of the EOSC PaN platform and to provide more examples and documentation for scientific use-cases for function-as-a-service and event-driven computing. During this phase the two applications: CrystFEL (Photon Science) and Mantid[25] (Neutron Science) were offered as scientific software frameworks for end-users.

The high-level architecture of the Photon & Neutron platform is depicted in Figure 6:

[24] http://www.desy.de/~twhite/crystfel/

[25] http://www.mantidproject.org

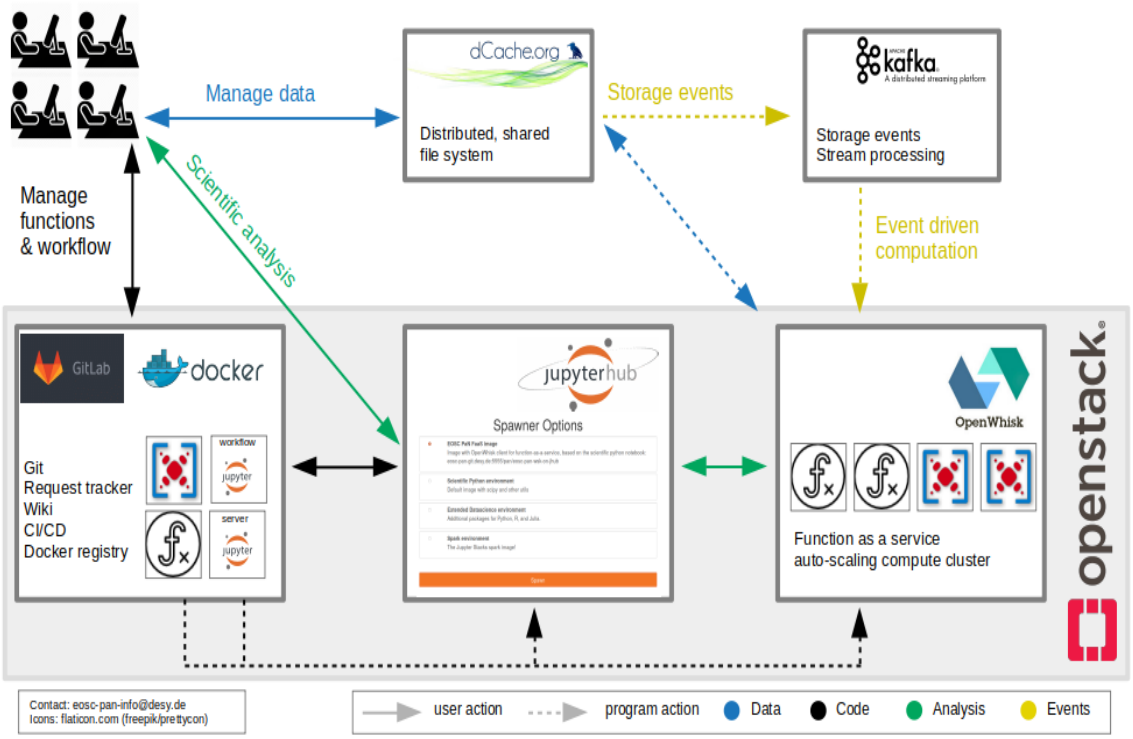

Figure 6 - The overall architecture of the resulted Photon and Neutron platform.

Achievements

During the extension of the WP5.4 activities, the Science Demonstrator was selected to consolidate and extend the services of the EOSC PaN platform and register the resulted solution in the EOSC Marketplace and EOSC. In the next sections are reported what was achieved during the 3-months additional extensions.

2.1.5.1. Source code management, CI/CD, docker registry

On the DESY OpenStack Cloud, access to GitLab[26] with federated AAI via EGI Check-In production instance was granted to end-users. In the extension phase, the auto-scaling Continuous Integration and Continuous Delivery (CI/CD) via group runners was enabled. These runners, available as Heat templates for rapid provisioning on OpenStack, were used to offload container building to cloud VMs, and to test the integrity of codes, especially the two essential building blocks for this project:

- Containers for Jupyter Servers to run Jupyter Notebooks in custom environments;
- Containers for OpenWhisk blackbox actions to run functions as services.

The built in Container registry is now accessible from the internet. Authorized users can upload/edit Docker files, build containers through CI/CD pipelines, deploy them on the platform and also pull them from other sites in the EOSC.

2.1.5.2. Collaborative computational environments, notebooks

The continuous integration of custom images as Jupyter Servers was enabled. In addition, it was created an image with pre-installed client for OpenWhisk Cloud Functions. When the user login, they can choose from a list of available images, or upload their own custom environment (via GitLab CI). The P&N Jupyter Hub[27] instance was registered as Service Provider (SP) in the AAI EGI-Check-in production instance.

[26] eosc-pan-git.desy.de

[27] eosc-pan-jhub.desy.de

2.1.5.3. Function-as-a-service, micro-services

OpenWhisk[28] was configured to use the Container Registry on the GitLab server. GitLab access tokens were used for secure access to the registry.

2.1.5.4. Backend storage and storage events

The Kafka Message Broker and the dynamic stream processing module were pulled into the OpenStack project and, additional examples were provided to run functions in response to incoming data were provided. When data sets are written into the dCache backend storage, configurable messages are sent to dedicated Kafka topics. These messages contain embedded macaroons token to limit the scope of who can access data.

2.1.5.5. CrystFEL

It was analysed how the Photon Science framework, for serial femto-second x-ray crystallography, performs for typical batch processing jobs and data reduction. During this analysis, it was reported that the tight integration of the CrystFEL framework into the GUI framework GTK was not ideal for cloud deployments. A clear separation between compute backend and data visualization is desirable.

2.1.5.6. Mantid

A container[29] version was used to integrate this neutron science framework as a micro service and provide example notebooks how to run it on data sets. In addition, an experimental event-driven workflow, to investigate its potential in automated data processing pipelines, was set-up. As for CrystFel, we gained experience with different job profiles, data volumes and processing times.

The knowledge gained from configuring and running Mantid as a micro-service will also help in adding other similar software suites that could be useful for Photon and Neutron Science community (eg. McStas and Sasview).

Recommendations

Some final recommendations from the project are:

- Since many Photon/Neutron applications require visual inspection of intermediate results produced during the analysis, it would be beneficial, for the end-users, whether **EOSC can provide** a reliable and effective **inspection/surveillance service**, across a global network, **to validate intermediate analysis**.
- Many Photon/Neutron applications are “free to use for academic purposes” but subject to restrictive licensing conditions. This means that to provide access to services based on such applications the user has to agree to licensing terms. It would be more scalable whether EOSC can provide such attributes in a central/federated way.
- To facilitate the reproducibility of data analysis, **scientific workflows are provided as Jupyter kernels** and offered to the P&N community as web services. A better integration of Jupyter notebooks with EOSC services would be beneficial.
- Since Docker container is one of the *de-facto* solutions adopted to “dockerize” applications, it would be beneficial if EOSC could provide a **centralized Docker registry providing trusted solutions**.

[28] eosc-pan-faas.desy.de (access only from eosc-pan-jhub.desy.de)

[29] mantidproject/mantid:latest

2.2. Second Set of Science Demonstrators

The first EOSCpilot Open Call for Science Demonstrators in April 2017 resulted in five new Science Demonstrators with execution from July 2017 to June 2018.

- **Energy Research – PROMINENCE**: HPCaaS for Fusion - Access to HPC class nodes for the Fusion Research community through a cloud interface.
- **Earth Sciences – EPOS/VERCE**: Virtual Earthquake and Computational Earth Science e-science environment in Europe.
- **Life Sciences/Genome Research**: Life Sciences Datasets: Leveraging EOSC to offload updating and standardizing life sciences datasets and to improve studies reproducibility, reusability and interoperability.
- **Life Sciences/Structural Biology**: CryoEM Workflows: Linking distributed data and data analysis resources as workflows in Structural Biology with cryo Electron Microscopy: Interoperability and reuse.
- **Physical Sciences/Astronomy**: LOFAR Data: Easy access to LOFAR data and knowledge extraction through Open Science Cloud.

The final evaluation status of the second set of Science Demonstrators is described in the following sections.

2.2.1. HPCaaS for Fusion (PROMINENCE)

Objective of the pilot

From a scientific perspective, the PROMINENCE Science Demonstrator aimed to **enable users of the fusion community to reproduce Science in a shorter time and in most efficient manner**. Access to HPC facilities are vitally important to the fusion community, not only for plasma modelling but also for advanced engineering and design, materials research, uncertainty quantification and advanced data analytics for engineering operations (e.g. condition monitoring). Unfortunately, the access to HPC facilities have some restrictions preventing the community to conduct scientific research in a reasonable time.

To remove these barriers **the Science Demonstrator developed a platform which provided a simple way for users to submit batch-like jobs, including multi-node MPI jobs, across different cloud resources.** The key point of this new platform was the ability to burst the computation on national research clouds, and then into the EGI Federated Cloud sites, as necessary to meet peak demands when other clouds are busy. Thanks to this platform users in the fusion community gain a transparent access to a wider variety of resources worldwide.

Technical capabilities

For this pilot, the following technical capabilities were identified:

- Compute - Used the EGI cloud compute resources, via opportunistic access, to burst the execution of scientific applications when local resources are busy.
- Processing & Analysis - Used high-level solutions such as: EC3/IM[30] to deploy complex and customized virtual infrastructures on multiple back-ends.
- Security & Operations - Proposed solutions to extend the AAI framework of the fusion platform.

Implementation

To support the implementation of this platform, and the orchestration of different cloud infrastructures, WP5.4 provided access to an instance of the Infrastructure Manager (IM), the solution developed by UPV in the context of the INDIGO-DataCloud[31] project. The IM is a key tool **that eases the access and the usability of IaaS clouds** by automating the selection, deployment, configuration, software installation, monitoring and update of Virtual Appliances[32]. The IM already supports a large variety of APIs for different IaaS making user's

[30] http://www.grycap.upv.es/im/index.php

[31] https://www.indigo-datacloud.eu/

[32] https://en.wikipedia.org/wiki/Virtual_appliance

applications cloud-agnostic. During the project the IM was used for provisioning resources across different cloud infrastructures, including the EGI FedCloud, OpenStack, Google Cloud Platform and Azure. The IM was used to deploy static SLURM clusters with OpenMPI and Singularity installed and to run MPI-based applications on the cloud resources of the EGI Federation. During the development of the pilot, up to **300 vCPU cores of the EGI cloud providers** were used for testing fusion jobs. The access to the EGI cloud providers (e.g.: CESNET-MetaCloud, IN2P3-IRES, RECAS-BARI and CESGA) through the IM was enabled by WP5.4.

The high-level overview of the platform developed to support the fusion community is shown in Figure 7.

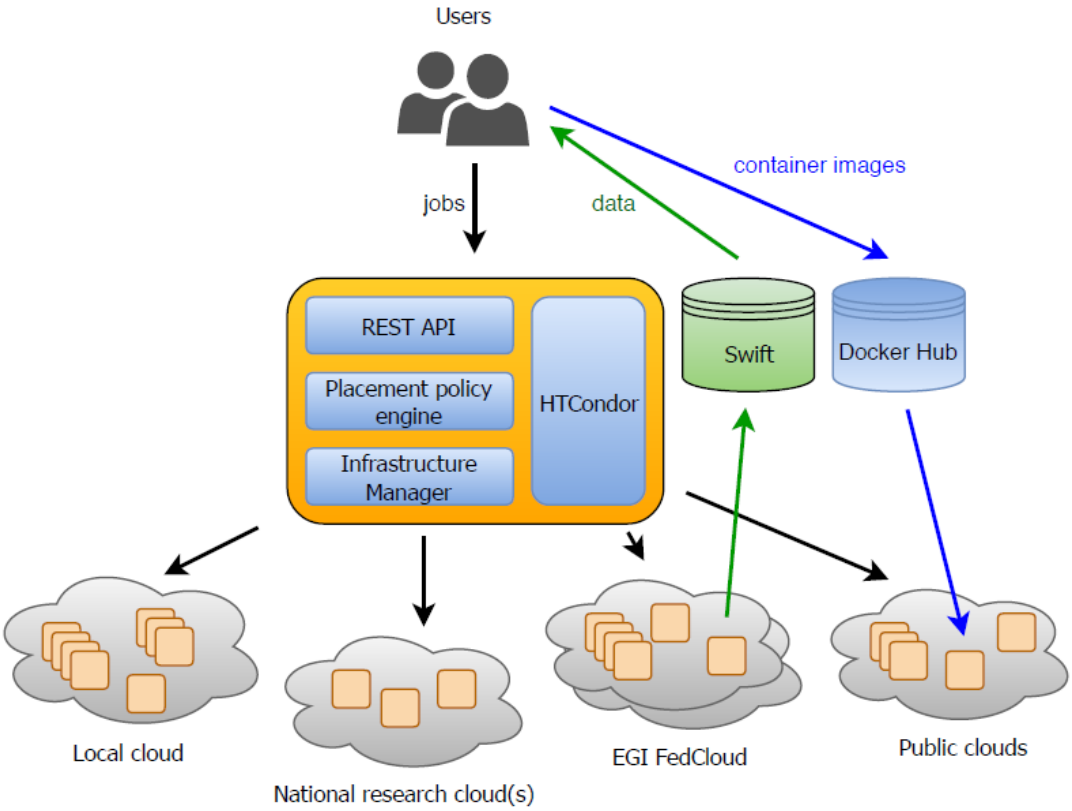

Figure 7 - Overview of the platform set-up for the fusion community.

Achievements

To move the pilot into a fully operational and deployable service, the Science Demonstrator was sponsored by WP5.4 to be further developed during the three month project extension. In the new work-plan, the Science Demonstrator focused on the handling of data, the integration of a transparent AAI, and enhance the scalability on multiple computing infrastructures.

Authentication

The integration with the EGI AAI Check-In was tested as an alternative to IAM[33]. The next step would be to set up a formal collaboration with EGI to make use of the production Check-in service. The workflow is as follows:

- Users use the prominence login command.
- Prominence login requests EGI Check-in for access using device flow.
- Prominence login returns a URL and device code and then waits for the user to authenticate.
- The user uses that link in a browser where they authenticate using their credentials and enter the device code.
- After user has authenticated, the prominence login command queries Check-in and retrieves the user's token, which is saved in the user's home directory.

[33] https://www.indigo-datacloud.eu/identity-and-access-management

- Users can then submit requests to the PROMINENCE service without any further authentication (until the token expires).

From a technical standpoint, there are open issues with the interoperability of the AAI solutions offered by EOSC. These solutions are not talking to each other and this prevents users to easily access services from different infrastructure providers.

Scalability Enhancements

To address the scalability issues the following model has been adopted:

For each applicable cloud provider, a single small VM was created by PROMINENCE and then used by the INDIGO Infrastructure Manager (IM)[34] to contextualize VMs on the same cloud provider. This Ansible VM was persisted and stored within a local database in PROMINENCE for future re-use (its existence is tested if it is to be reused). This VM was used by Ansible to construct the required cluster nodes using the internal private network of the cluster, thus removing the scalability issues associated with using public floating IPs since only one required per cloud instance. This workflow is shown in Figure 8.

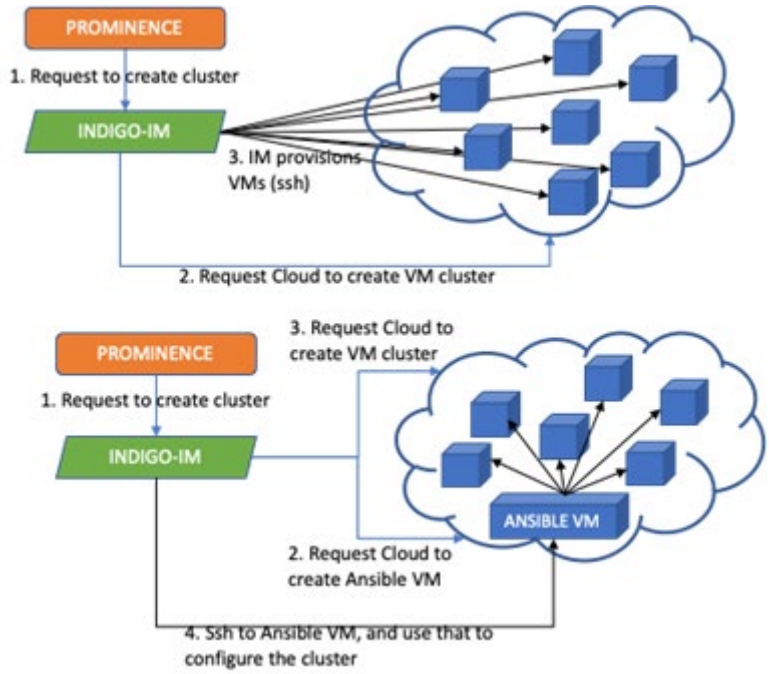

Figure 8 - Original (upper) and new (lower) diagrams showing the reduction in the number of public floating IPs required.

Integrating Data management

In the original science demonstrator, large scale data management was specifically omitted since the original target was using modelling codes which required only temporary storage for the output of these models. In this case, the CEPH storage at STFC was used to store the user's data using the PROMINENCE CLI, and return a temporary Swift URL to the user to access the data.

The PROMINENCE platform was also extended in order to allow different tasks within a workflow to be executed on different clouds (see later). Since generally steps within a workflow communicate via file, the ability to make use of shared storage which is accessible from anywhere is vitally important. The following technologies were tested (or are under testing):

[34] https://www.grycap.upv.es/im/index.php

B2DROP

B2DROP[35], the secured and trusted data exchange service, would appear to be ideal since often the intermediate files for an analysis workflow can be deleted. While APIs are somewhat limited within B2DROP, it is possible to use a browser to obtain an obscured 'app token' consisting of a shared username and password. This token can then be passed to VMs running jobs to allow access to the service. By default the B2DROP itself is mounted as a WebDAV FUSE filesystem by installing the davfs2 package and using the documented method defined in the B2DROP documentation. Within PROMINENCE this is taken care of automatically, and users just have to provide their credentials when submitting the job.

B2SAFE

By the time of writing this report, the evaluation of this solution was still in progress. Several issues experienced during the integration made this solution unattractive for long term integration. These issues are outlined below:

- Each site is slightly different and offers different access mechanisms.
- Access to some sites is only available via plain username/password representing a security risk. Integration with B2ACCESS has yet to be fully rolled out.
- Coordination between sites is a very effortful process; is has taken many months to get access to two sites so far (out of three requested).
- The only common API is the iRODS iCommands which we reject as being a proprietary API which goes against FAIR principles.
- Each replica gets its own PID which means making use of these for publication is fraught with danger.

Performance Testing

Table 2 shows the results of performance testing different storage technologies and the APIs they provide using different cloud providers. Note that the CEPH instance used is hosted at STFC and the general instance of B2DROP is hosted at Jüelich.

VM Site	Storage Technology	Protocol	Network Performance (MB/s)
Google, us-east1-b	CEPH	curl	21 (write) 24 (read)
	CEPH	S3FS	15 (write) 26 (read)
	B2DROP	WebDAV	14 (write) 26 (read)
IN2P3-IRES	CEPH	curl	34 (write) 90 (read)
	CEPH	S3FS	17 (write) 34 (read)
	B2DROP	WebDAV	25 (write) 30 (read)

[35] https://www.eudat.eu/services/b2drop

VM Site	Storage Technology	Protocol	Network Performance (MB/s)
STFC	CEPH	curl	114 (write) 195 (read)
	CEPH	S3FS	38 (write) 137 (read)
	B2DROP	WebDAV	4 (write) 3 (read)

Table 2 - Performance testing in different storage providers.

Whilst in general these results are in line with what is expected, accessing B2DROP from VMs at STFC was significantly slower. Before being used operationally, additional testing of B2DROP would be required to investigate this anomalous behaviour and ensure that they are not repeated at other cloud provider sites.

Additional work carried-out

Token-based access to the EGI FedCloud Infrastructure
As part of the extension, it was tested the access to the IN2P3-IRES cloud provider using the standard OpenStack APIs and standard token-based authentication. The deployment workflow used within PROMINENCE was extended accordingly in order to ensure that a valid access token is available and new tokens are obtained (using a refresh token) as needed.

Several GitHub issues for IM were submitted due to issues found while carrying out this work (#804, #807, #808) and we have continued to receive good support from the IM developer.

MPI improvements
Previously PROMINENCE only supported OpenMPI, and specifically only a limited number of versions of OpenMPI (1.10.7, 2.1.1, 3.0.2 and 3.1.0). This meant that users had to create container images compatible with one of the supported versions of OpenMPI. The reason for this is that we were using the common approach of using the mpirun command on the host to launch udocker or Singularity containers containing the user's MPI application, and this requires the same version of MPI on each host as in the container.

The MPICH support, in addition to OpenMPI, was added and there is no longer any restriction to what version of MPI is being used. This was achieved by using special wrapper script which enables to run mpirun from within a container, which is then able to launch containers on the other hosts in order to run the user's application. Commands internal to MPI, such as orted, are also run inside containers. This approach ensures that the versions of MPI used are identical everywhere (as the same container image is used) and ensure MPI jobs are more likely to run reliably.

Command line interface
The PROMINENCE CLI is now available on the Python Package Index, PyPi, enabling users to install it as simply as typing "pip install prominence".

Support for HPC systems
Through the EOSC-hub project it was obtained access to AVITOHOL, a HPC system in Bulgaria. This contributed to demonstrate how PROMINENCE could make use of existing HPC resources in addition to clouds, and enable users to transparently use both HPC and cloud resources.

Because PROMINENCE internally uses HTCondor for managing jobs, existing functionality within HTCondor were used for submitting jobs to external batch systems. A single ssh connection was used to connect to a standard login node to launch a HTCondor process called the BLAHP, which submits jobs to the batch system

and monitor their state. This is shown in Figure 9. It is important to note that only a standard account is required on the HPC cluster and no special modifications are required.

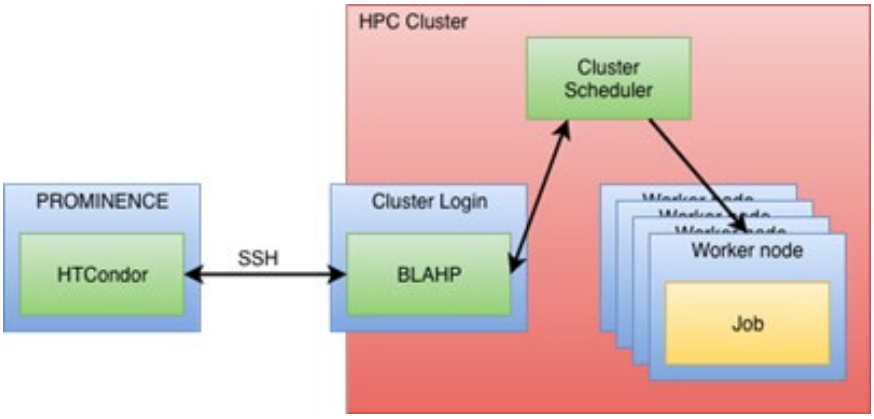

Figure 9 - Architecture diagram showing how PROMINENCE can submit jobs to an external batch system.

A proof-of-concept of PROMINENCE submitting containerized single-node jobs to AVITOHOL was demonstrated.

Workflows

Originally PROMINENCE was designed to handle only individual batch jobs. Leveraging the Directed Acyclic Graph (DAG) Manager in HTCondor, the framework was extended to allow users to submit DAG workflows. This means that users can specify a sequence of jobs, where each step depends on the previous step(s). Users can specify different resource requirements for each job in the workflow, meaning that a single workflow can contain both single-node HTC jobs and multi-node HPC jobs, and each job will only use the required resources. Workflows are not constrained to run in only a single cloud - the jobs comprising a single workflow can be run across multiple clouds.

Combining the workflow functionality with the new integration with B2DROP it was possible to demonstrate a running a workflow which involves rendering a movie. The frames were rendered across both IN2P3-IRES and STFC, and the final movie was generated from the individual frames as a second step in the workflow on IN2P3-IRES.

The lessons learned from the project are:

- The importance of technical support and the availability of experts helping with technical enabling activities is required in complex workflows.
- Even if different levels of integration are available in EOSC services (e.g. it is possible to log in B2DROP instance using an EGI SSO credential), these are still not well integrated with each other. The burden to make these services interoperable is now on the researchers' shoulders. EOSC should take care to provide a more tighter services integration to the end-users.
- The AAI interoperability in EOSC is still in the early stage. This has to be tackled effectively to provide a simplified access to services and resources.

2.2.2. Virtual Earthquake and Computational Earth Science e-science environment in Europe (EPOS/VERCE)

Objective of the pilot

The Science Demonstrator aimed to make the EPOS/VERCE Virtual Research E-Infrastructure (VRE) integrated and interoperable with e-Infrastructures of the EOSC. The final scope of this pilot was to support researchers with the study and the improvement of regional and global Earth models. The demonstrator addressed challenges that emerged from real-life computational scenarios in seismic hazard assessment, where each experiment involves 100s of GB of data and metadata (~20TB storage). The heterogeneous datasets will be generated in HPC and Clouds, acquired from federated archives, and moved across computational and

dedicated storage resources. The application of FAIR principles is guaranteed by the generation of provenance information during the execution of the models.

Technical capabilities
For this pilot, the following technical capabilities were identified:

- Compute - Allocated sufficient cloud capacity in the EGI cloud compute Infrastructure to execute Earth model simulations.
- Processing & Analysis - Used high-level solutions such as: EC3/IM to deploy complex and customized virtual infrastructures on multiple back-ends.
- Data Management - Allocated resources for storing and publishing large volume of datasets.
- Security & Operations - Enhanced the AAI framework of the EPOS/VERCE VRE to support federated authentication mechanism.

Implementation
Overall, during the course of the project, the back-end services of the Misfit earthquake application was enhanced in order to extend the computational capabilities of the VERCE platform[36] with the EOSC cloud resources through a required management middleware. To support this enhancement and migrate part of the scientific workflows running on HPC resources on the cloud resources, a new release of the DCI Bridge service used by the WS-PGRADE portal to be interoperable with the EOSC cloud infrastructure was developed. The latest release of the DCI Bridge service, which is now registered in the EGI Cloud Marketplace (AppDB)[37], was configured to use the latest gUSE release[38], includes additional libraries to support the Misfit Analysis Workflows and enables the contextualization of the VM following the good practices provided by EGI. This work, supported by WP5.4, the WS-PGRADE technical staff and members of the Science Demonstrator, contributed to use the EOSC cloud infrastructure as the main data-intensive computational service provider for Misfit Analysis Workflows.

From a technical standpoint, through the EGI FedCloud Virtual Organisation (verce.eu), the workflows for the evaluation of Earth models, including the processing and the comparison of data resulting from simulations of seismic wave propagation following a real earthquake, and real measurement recorded by seismographs, utilised cloud resources on currently three sites: SCAI, GRNET and IN2P3.

During the runtime of the Science Demonstrator, WP5.4 also contributed providing solutions to extend the AAI framework of the VERCE portal and enable single sign-on access to the EOSC infrastructure. The interoperability with the AAI EOSC infrastructure was achieved registering the portal as service provider of the EGI AAI Check-In service[39] and using the OIDC module[40], developed by the INFN, to allow users of the WS-PGRADE portal, based on Liferay portal framework, to login using social media accounts (e.g.: Facebook, Google, LinkedIn, ORCID), eduGAIN, or home institutional accounts. Last but not least, the VERCE portal has been extended to allow the retrieval of Per-User Sub-Proxy[41] certificates from the eToken proxy certificate service. The pbjectives of this are the execution of scientific workflows and the generation of provenance data.

[36] http://www.verce.eu
[37] https://appdb.egi.eu/
[38] https://sourceforge.net/projects/guse/files/3.7.5
[39] https://wiki.egi.eu/wiki/AAI_guide_for_SPs
[40] https://github.com/csgf/OpenIdConnectLiferay/tree/EGICheckIn
[41] https://wiki.egi.eu/wiki/Usage_of_the_per_user_sub_proxy_in_EGI

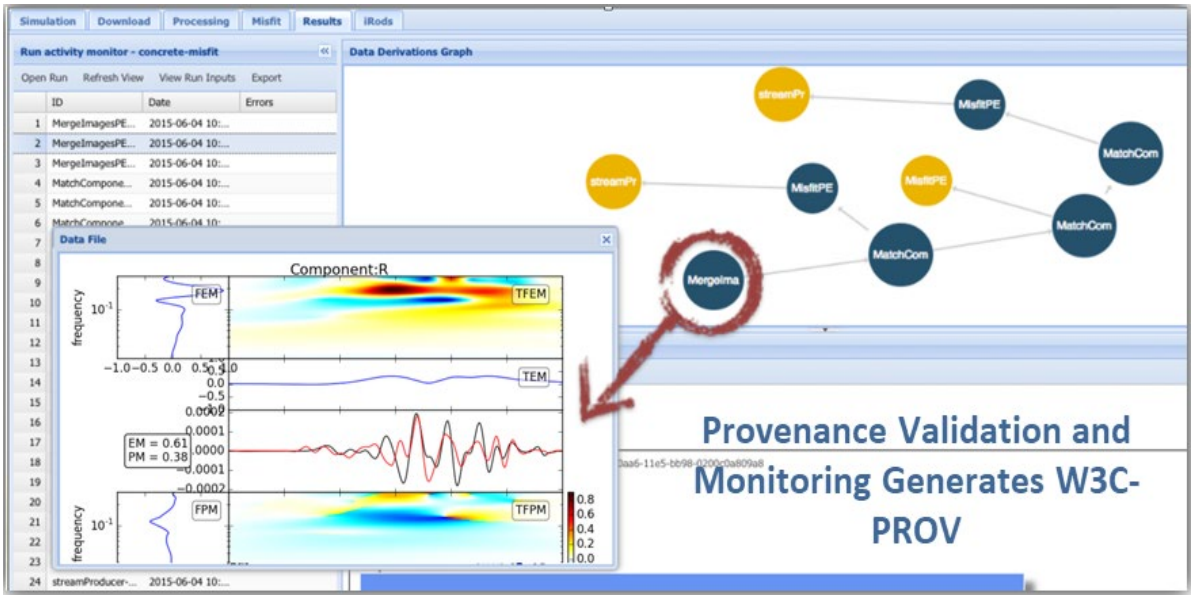

Figure 10 - Execution of a Misfit workflow.

Achievements

- The processing workflow was refactored in order to support the execution of Earth models also in the EGI cloud computing infrastructure. Three cloud providers of the Federation allocated the initial set of resources needed by the pilot.
- A dockerized version of the provenance service was developed. The provenance exploitation tools in the VERCE portal was upgraded.
- The EPOS/VERCE portal AAI framework was enhanced to support federated authentication mechanism and allow users to login with home institutional accounts from accredited IdP registered in eduGAIN, or via social media accounts.
- Support to robot certificate to run Earth model simulations in the computational infrastructure was enabled.
- Simulation results were sent to iRODS. The metadata, of the simulation results, were extracted during processing and stored within the S-ProvFlow system.

Recommendations:

- EOSC should provide a centralized catalogue of services from the different e-Infrastructure providers. By the time of running of this Science Demonstrator the catalogue was not ready and it made a bit difficult to get a clear picture of the list of services available in EOSC.
- There is no a clear endorsement from EOSC to support high-level services such as: portals and science gateways. In some cases, communities are developing their own open-source solutions.
- EOSC services should be able to manipulate provenance information (e.g. generation, acquisition, exploration and exploitation) to facilitate the FAIR principles.

2.2.3. EGA Life Science Datasets Leveraging EOSC

Objective of the pilot

In Biological Sciences, services like the European Genome-phenome Archive (EGA)[42] offer Life Sciences datasets for re-analysis at the requester computing facilities. Usually, these datasets have been processed with reference genomes and analysis pipelines that were current at the time but become obsolete quite fast. The main objective of this Science Demonstrator was to **test the feasibility of data reproducibility and re-mastering in genomics**.

[42] https://ega-archive.org

Technical capabilities
For this pilot, the following technical capabilities were identified:

- Compute - Allocated sufficient cloud capacity in the EMBL-EBI Embassy Cloud[43] to support the pilot.
- Processing & Analysis - Combined docker and workflow technologies to reproduce scientific analysis.

Implementation
During the project lifetime, the Science Demonstrator managed to reproduce and re-master biological pipelines using Nextflow[44], an emerging language aimed to ease the interpretation of scientific workflows.
To do so, a dedicated technical meeting with the Nextflow experts was arranged to discuss how biological pipelines could be converted with this framework. Results from rebuilding pipelines were compared to original datasets. Early January 2018, the biological pipelines were successfully converted to run through Nextflow, and in March 2018 newly updated pipelines were executed at the Barcelona Supercomputing Centre (BSC)[45]. In April 2018 the Science Demonstrator started to explore how B2FIND, the EUDAT Metadata catalogue, could be used to manage metadata produced by the pilot service. An initial technical meeting with B2FIND experts was organized by WP5.4. During the meeting were discussed the guidelines for helping the Metadata catalogue to harvest data from the data provider. The next step was to agree on the metadata scheme used to implement the harvesting.

Achievements
- The reproducibility of biological pipelines was demonstrated using the genome hg19 and hg38. Differences existing between the data obtained, were analysed both quantitatively and qualitatively.
- Containerized versions of the pipelines were successfully executed in a third-party infrastructure, demonstrating the portability of the pipelines implemented in this Science Demonstrator.
- Pipelines were deposited in Dockstore[46].

2.2.4. CryoEM workflows

Objective of the pilot
CryoEM is a framework to facilitate the understanding of biological functions. Among the Structural Biology techniques, the microscopy under cryogenic conditions ("cryo-EM") is currently the fastest growing area, having been nominated "Method of the Year (2015)" by Nature. The main objective of this Science Demonstrator was to **support the reproducibility of Science by enabling the sharing of detailed information on cryo-electronic microscopy image processing workflows**.

Implementation
Current practices in electron-microscopy is the following:

- User visits the facility and generates data.
- Data is pushed from electron-microscope into a local storage.
    - Approx. 1 TB/day, usually for 2 days/user.
- At large facilities (synchrotrons):
    - The facility stores the raw data for longer term.
        - User performs analysis and generates processed data using Scipion, the application framework deployed at the facility.
    - User takes home the workflows and the processed data (output of workflows).
        - Raw data stays at facility, with an embargo period (years).
        - Open access after embargo.

[43] https://www.embassycloud.org/
[44] https://www.nextflow.io/
[45] https://www.bsc.es/
[46] http://dockstore.org

- At small facilities:
    - User is given with the raw data and processed data on a disk which they bring home.

During the 12 months funding, a significant amount of time was spent to extend the Scipion application framework in order to make the resulting workflows, used to describe the image processing steps, shareable, re-playable and the outputs reproducible. From a technical standpoint, this goal was reached by extending the Scipion framework to generate a workflow file which linked together raw data, metadata, tools and also a list of movies used for the analysis. The workflow file used to describe image processing steps was exported in JSON format and deposited in the CryoEM major databases. Each workflow file contained all the relevant information to describe and enable the reproducibility of image processing steps.

For testing, the Science Demonstrator used example of non-embargoed raw datasets from the Diamond and ESFR databases, and showcased how a user A, who has access to this data, can develop and execute a workflow on system A, and publish the workflow reproducibility file. This resulted file can be picked up by user B who can repeat the same execution and result the same output on a system B. The assumption is that both user A and B has access to the same raw dataset.

To support the creation of reproducible workflows to describe the image processing steps, WP5.4 invited the scientific contacts to explore the use of the Common Workflow Language de-facto standard.

Achievements

- Cryo-electronic microscopy image processing workflows were extended to support FAIR principles. Thanks to this integration scientists can now reproduce and reuse each other's workflows and open data.

As further recommendations provided by the Science Demonstrator, it would be beneficial whether EOSC can:

- Create a public repository of acquisition metadata and image processing workflows for new acquisitions, a temporary repository until the data is finally analysed and deposited in the standard public databases (EMDB and EMPIAR).
- Provide a cross-infrastructure AAI solution to allow biologists coming out from an EM facility to perform image processing in the EOSC computing providers.
- Support Workflow Language systems (e.g. CWL[47] or NextFlow[48]), for helping research communities to better organise and execute analytical and computational pipelines.

2.2.5. Astronomy Open Science Cloud access to LOFAR data

Objective of the pilot

The main objective of this Science Demonstrator was to **allow science community to locate, access, and extract science from the LOFAR archive without being an expert on data retrieval and data analysis tools**. As a result, users will be able to create new scientific results based on archived data products.

Technical capabilities

For this pilot, the following technical capabilities were identified:

- Compute - Allocated HPC resources for scaling the execution of the pipelines.
- Processing & Analysis - Use of Workflow Languages for describing analysis workflows and execute scientific pipelines.
- Security & Operation - Offered solutions to enable federated authentication mechanism in the community.

[47] https://www.commonwl.org/

[48] https://www.nextflow.io/

Implementation

To scale up the execution of LOFAR pipelines, access to the SURFsara HPC cluster was granted. An initial discussion to use CVMFS for software distribution in the Science Demonstrator was initiated by WP5.4. The portable deployment and execution of LOFAR pipelines (e.g.: Prefactor[49], Presto[50] and Spiel[51]) on EOSC infrastructure was achieved through the adoption of container-based solutions (e.g.: Docker[52], Singularity[53] and uDocker[54]), and the utilization of the Common Workflow Language (CWL) [55] standard for defining processing workflows. This combination proved to be extremely flexible and allow running the same workflow on systems ranging from a personal laptop to large scale High Throughput computational clusters. The three LOFAR pipeline were executed in the SURFsara HPC cloud environment and Cartesius[56], the Dutch national supercomputer, also hosted by SURFsara, but WP5.4 already liaised with the PSNC and FZJ Jüelich to scale-up the HPC resources. Metadata from the LOFAR archive database was imported in a Virtuoso RDF triple store.

During the funding period the Science Demonstrator reached an excellent level of development, targeting open and FAIR access to data in the LOFAR long term archive. For this reason, during the extension of the WP5.4 activities, LOFAR was selected to implement a Proof-of-Concept AAI integration, e.g. setting up a LOFAR Collaboration in a COManage instance, and use it to provide access to data in the LOFAR archive as well as to compute resources.

Achievements

To move the pilot into a fully operational and deployable service, the Science Demonstrator was awarded by WP5.4 to enable the further exploitation of, and integration with, EOSC infrastructure and services. The proposed work focused on the two subjects:

Federated access to LOFAR services

To address this objective, a team was formed with representatives from EGI, GRNET, ASTRON, and SURFsara. Through bi-weekly telcos, information was exchanged and a plan devised and implementation coordinated for setting up federated user management and service integration. ASTRON has documented the organization of LOFAR science teams. GRNET has configured a Collaborative Organization (CO) environment in the EGI-Check-in COmanage Registry service (development instance).

The functionality of COmanage was also extended to support the LOFAR collaboration requirements with respect to self-organization of science teams, central organization of LOFAR support, and role/group membership-based entitlements to be used by services for authorization purposes. The functionality added to COmanage supports:

- Using the title of COU membership records as roles from a vocabulary specific to the ASTRON CO.
- Sending email notification to all members of a group/science team upon deletion.
- Searching filtering groups/science teams based on their name, description and status.
- Sending email invitation to users for joining ASTRON CO.

The following services have been successfully integrated with this environment to allow federated user access:

- SVN Code repository with group-based web-access to different coding projects based on SAML.
- JIRA User Helpdesk and ticketing system based on SAML.

[49] https://github.com/lofar-astron/prefactor

[50] https://www.cv.nrao.edu/~sransom/presto/

[51] https://github.com/gijzelaerr/spiel

[52] https://www.docker.com/

[53] https://singularity.lbl.gov/

[54] https://github.com/indigo-dc/udocker

[55] https://www.commonwl.org/

[56] https://userinfo.surfsara.nl/systems/cartesius

- Web framework to be used for implementation of a staging service (see below) based on OpenID Connect.

Other services considered to be integrated are:

- SVN code repository with group-based client access based on public SSH key distribution.
- SPIDER compute cluster based on public SSH key distribution.
- LOFAR web-based archive catalogue (integration mechanism to be investigated).

The objective was to implement a proof-of-principle public SSH key distribution through COmanage, and have an implementation plan for the other services, before the end of the project.

In parallel, SURFsara and ASTRON have started work on an approach to improve data access for end users by replacing existing X509 certificate-based access and application of custom web-based download servers which do not easily scale to address growing data access demands. The new service to be developed will provide access to scalable WEBDAV-based data access, natively provided by the underlying dCache middleware, using 'macaroon' tokens to enforce the applicable authorization policy. In this part of the activity, the following tasks have been completed:

- Document and analyse existing data staging/transport practices, identifying bottlenecks and options for improvement.
- Create an implementation plan for the dCache Macaroon setup and utilization of dCache WEBDAV for data transport.
- Configure a LOFAR VOMS user role to be used for the generation of Macaroon authorization tokens.
- Perform groundwork (select applicable libraries, setting up the federated web framework) for implementation of the staging service.

The objective was to be able to demonstrate a functional federated staging service before the end of the project.

Demonstration of portable LOFAR processing pipelines at large scale

With respect to activities towards demonstration of portable LOFAR processing pipelines at large scale, preparatory work on the SURFsara SPIDER cluster for support of LOFAR processing pipelines is being undertaken and a new release candidate for the Common Workflow Language (CWL) was published. The processing of the Prefactor pipeline using the CWL standard is shown in Figure 11. It includes the support for manipulation of large files without making copies on the file system, as needed for LOFAR data sets.

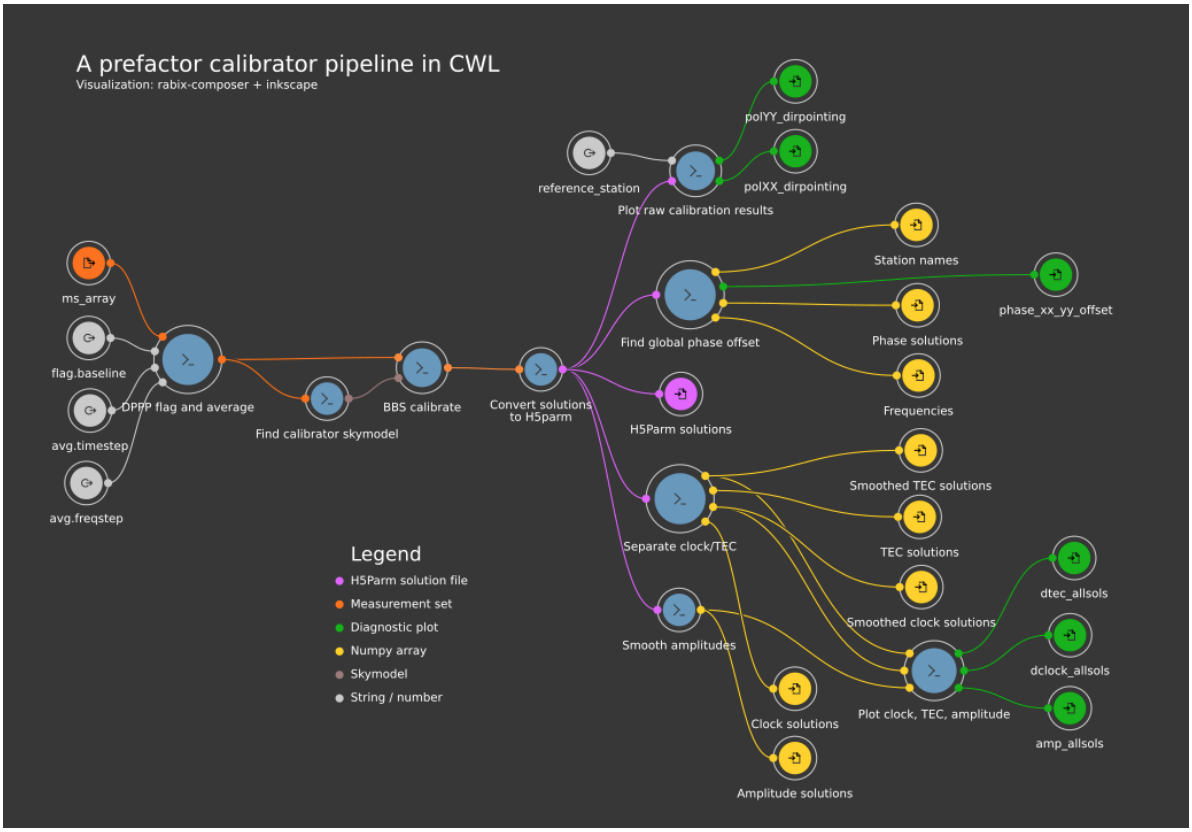

Figure 11 - The Prefactor CWL pipeline, visualised with Rabix and coloured by hand.

As further recommendations provided by the Science Demonstrator, it would be beneficial if EOSC can:

- Provide a central **EOSC inked data platform** as a low-threshold entry point for communities that are new to the technologies.
- Provide support for Workflow Languages for describing analysis workflows and tools in a way that makes them portable and scalable across a variety of software and hardware environments, from workstations to cluster, cloud, and high-performance computing (HPC) environments.
- Broad support for VM/containers would be extremely useful to deal with application deployment complexity (for LOFAR, but likely for many communities). General purpose support for the CWL standards would be useful as well.
- Strengthen the synergies between service and e-Infrastructure providers.

2.3. Third Set of Science Demonstrators

The second, and last, EOSCpilot Open Call for Science Demonstrators in August/September 2017 resulted in five new Science Demonstrators with execution from December 1, 2017, to November 2018:

- Generic Technologies: Frictionless Data Exchange Across Research Data, Software and Scientific Paper Repositories.
- Astro Sciences: VisIVO - Data Knowledge Visual Analytics Framework for Astrophysics.
- Social Sciences and Humanities (SSH): VisualMedia: a service for sharing and visualizing visual media files on the web.
- Life Sciences and Health Research – Genome research - Bioimaging: Mining a large image repository to extract new biological knowledge about human gene function.
- Earth Sciences-Hydrology: Switching on the EOSC for Reproducible Computational Hydrology by FAIRifying eWaterCycle and SWITCH-ON.

The final evaluation status of the last set of selected Science Demonstrators is described in the following sections.

2.3.1. Frictionless Data Exchange Across Scientific Repositories

Objective of the pilot

The OAI-PHM protocol is currently the world-wide standard adopted to exchange metadata and content across scientific repositories, nevertheless this protocol is affected by the following weak points:

- It is unsuitable when there is a need to exchange large quantities of metadata;
- It was mainly designed for metadata transfer, the support for content exchange is lacking, and
- There is an inconsistent implementation across providers.

The main objective of the Frictionless Data Exchange Across Research Data, Software and Scientific Paper Repositories pilot was to **promote alternative solutions to interact with research repositories**.

Implementation

In the context of the EOSCpilot project, **the Frictionless Data Exchange Across Research Data pilot promoted a fast and highly scalable** exchange **of data across repositories** based on the ResourceSynch[57] protocol developed by the CORE team at the Open University (OU) in collaboration with Los Alamos National Laboratory (LANL) and the Data Archiving and Networked Services (DANS).

During the first part of the funding period, in collaboration with the CORE team, a ResourceSynch endpoint was configured to target the Science Demonstrator requirements. A first implementation of a new scalable client solution for accessing CORE metadata and large amount of data with ResourceSynch protocol was designed early 2018. Several tests were conducted, with different configurations, to perform a quantitative evaluation of ResourceSynch against the widely used OAI-PMH protocol, especially focusing on the use case of aggregating millions of resources from scientific repositories to the CORE aggregator and enabling others to keep in sync with relevant parts of this dataset.

The original set-up was further refined during the second part of the project, and a discussion with the ResourceSynch community, regarding a possible standardization of the approach, has started. Starting from July 2018 a connection with the OpenAIRE developers was established to use the ResourceSynch endpoint in their ecosystem. Additionally, initial engagement with the other Science Demonstrators (e.g. TEXTCROWD and DPHEP) was established to investigate a further development of the solution in different domains. This interaction with the OpenAIRE developers and was facilitated by WP5.4.

[57] http://www.openarchives.org/rs/toc

Achievements

- The pilot proved that ResourceSync is a valid protocol to improve the interoperability. The protocol can be applied to any data repositories. The main requirements are about disk space and processing power, in case of big volumes of data this could be challenging, and correct trade-off needs to be found.
- OpenAIRE developers have been supported in implementing ResourceSync in their ecosystem.

Recommendations

- All the data hosted in EOSC should have a standard metadata definition to enable an easier distribution and discoverability.
- Promote ResourceSynch in EOSC for helping the integration and interoperation of distributed systems.

2.3.2. VisIVO: Data Knowledge Visual Analytics Framework for Astrophysics

Objective of the pilot

The Astrophysical community has set up a new suite of cutting-edge Milky Way surveys that provides a homogenous coverage of the entire Galactic Plane and that have already started to transform the view of our Galaxy as a global star formation engine. Over the past decade, new instruments have made possible to generate vast amounts of data which will create enormous challenges for capturing, managing and processing of this data. For this specific scientific domain, the volume, the complexity of datasets and the scientific challenges to be tackled require a radical re-evaluation of the current science and data analysis techniques. The main objective of this Science Demonstrator focused on investigating the use of the EOSC technologies for the archive services and intensive analysis employing the connection with a science cloud gateway.

Technical capabilities

For this pilot, the following technical capabilities were identified:

- Compute - Allocated sufficient cloud capacity in the EGI cloud compute Infrastructure to execute intensive analysis.
- Security & Operations - Enhanced the AAI framework of the VisIVO portal to support federated authentication mechanism.

Implementation

During the funding period, the VisIVO Science Demonstrator planned to integrate in EOSC the ViaLactea Visual Analytics (VLVA) tool, a visual analytics environment based on the VisIVO framework developed by INAF Catania. This tool was offered to the scientific community through the ViaLactea science gateway[58] based on the WS-PGRADE/gUSE technology framework.

Since the ViSIVO science based is based on the WS-PGRADE/gUSE technology framework, to facilitate the integration of the ViaLactea Visual Analytics tool in EOSC, WP5.4 organized a technical meeting with the VisIVO and the EPOS/VERCE teams. The aim of this meeting was to share the good practices and experience in interacting with EOSC.

Achievements

- The ViaLactea science gateway was successfully connected with the EGI FedCloud Infrastructure. Sample test workflows have been submitted to the DCI-Bridge VA. The implementation of the workflow running SED analysis[59] was finalized (including the code porting from IDL to GDL) and submitted to the VisIVO VA.

[58] https://vialactea-sg.oact.inaf.it/

[59] https://www.ict.inaf.it/gitlab/VisIVO/ViaLacteaVisualAnalytics/tree/EOSC-VisIVOScienceDemostrator

- The ViaLactea science gateway was registered as Service Provider (SP) of the EGI AAI Check-In service to allow users to access the EOSC infrastructure with federated credentials. The set-up was enabled by WP5.4.
- A custom DCI_Bridge Virtual Appliance (VA)[60], to target the requirements of the scientific community, was created. The creation of this VA was facilitated by WP5.4 and members involved in the EPOS/VERCE Science Demonstrator.
- A dedicated DigiCert robot certificate[61] to access the cloud resources of the EGI Federation was generated for the VisIVO Science Demonstrator. The new robot was uploaded in the eToken server to generate Per-User sub-Proxy (PUSP)[62]. The access to the cloud providers of the EGI Federation was also enabled by WP5.4.
- First data releases of the Galactic Plane was deployed on the EGI FedCloud resources (~400GB).
- Designed a simple RESTful Cloud Gateway (RCG) API to allow the execution of workflows without requiring a personal user certificate.

The extended architecture of the Data Knowledge Visual Analytics Framework to use the EOSC resources and services for hosting the knowledge base archiving service and perform intensive data analysis is shown in Figure 12.

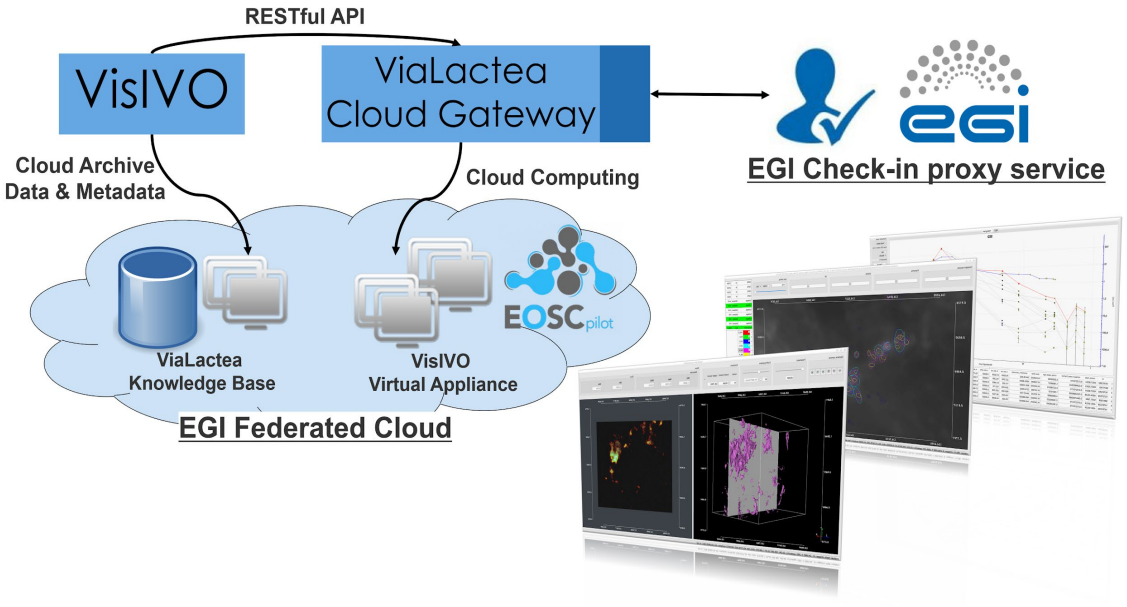

Figure 12 - The Data Knowledge Visual Analytics Framework.

Recommendations

As further recommendations provided by the Science Demonstrator, it would be beneficial if the EOSC can provide:

- A catalogue of services where users can evaluate each service based on user's experience, and documentation would be beneficial.
- Object Storage to save VLVA users' sessions data.

[60] https://appdb.egi.eu/store/vappliance/visivo.sd.va

[61] https://www.digicert.com/secure/requests

[62] https://wiki.egi.eu/wiki/Usage_of_the_per_user_sub_proxy_in_EGI

2.3.3. VisualMedia: a service for sharing and visualizing visual media files on the web

Objective of the pilot

High-resolution visual media, and 3D models, play a paramount role in Cultural Heritage (CH) research, as well as in other research communities. Images of artefacts, monuments and sites are the most widely used support to analysis and interpretation. The availability of an easy-to-use, yet powerful management service for images/3D models will improve data-based research. By enabling data sharing on the web, it will also improve collaborative research, both as regards team working on the same images and teams re-using images provided by others for new investigations.

In this scenario, the VisualMedia Science Demonstrator aims to bridge this gap by **piloting an easy visualization service of complex visual media assets**. From a technical standpoint, the service will initially upload the visual media files, available in different formats, in the EOSC cloud infrastructure, convert these media contents into an efficient web format, and finally make the resulted converted media contents ready for web-based visualization.

Technical capabilities

For this pilot, the following technical capabilities were identified:

- Security & Operations - Enhanced the AAI framework of the pilot to support federated authentication mechanism.
- Processing & Analysis - D4Science, a platform for data validation, data enrichment and efficient data analysis.

Implementation

To support the implementation of this pilot service WP5.4 liaised with the CNR provider to facilitate the development of a dedicated Virtual Research Environment[63] (VRE) on top of the D4Science Infrastructure, including all the services needed by the Cultural Heritage research community. During the technical meeting with the CNR team, a series of activities were planned. These included: the authentication & authorization, storage and scalability (processing). The integration of the AAI framework started early in January 2018. Several authentication mechanisms were integrated in the pilot service, including the D4Science authentication framework, the support to social media, and password-less login authentication. Regarding the visual data uploading features, the original portal allowed to upload data from the remote user's machine. This option was also extended by enabling the possibility to load visual media data from the personal storage area of any specific user (i.e. from the D4Science user workspace). A new web front-page was also developed when the Visual Media service was consulted with mobile devices (tablets and smartphones). During the last part of the project, additional functionalities were integrated in the pilot services (e.g.: automatic production of thumbnails, introduced the implementation of image collections, and improved the file management just to name a few).

A view of the VRE set-up for the Cultural Heritage community is shown in Figure 13.

Achievements

During the past three months, WP5.4 sponsored the further extend of the AAI framework to provide seamless access to the VisualMedia VRE via the EGI AAI Check-in service. The additional work-plan was also complemented with the registration of the VisualMedia service in the EOSC Portal Marketplace and become a new EOSC service provider.

[63] https://services.d4science.org/web/visualmedia

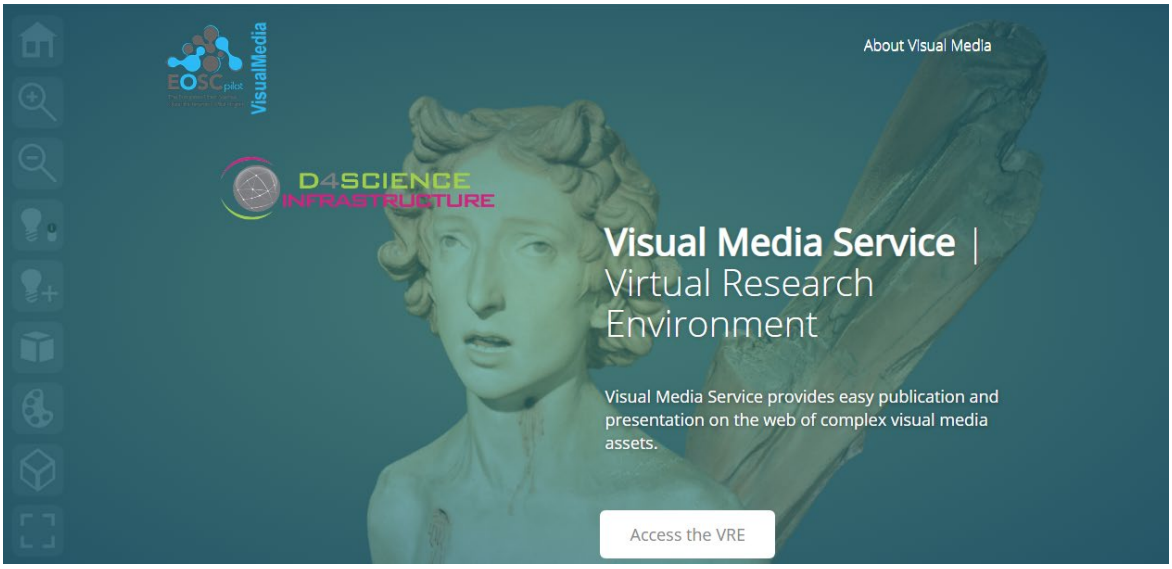

Figure 13 - The VisualMedia VRE homepage on the D4Science framework.

2.3.4. Mining a large image repository to extract new biological knowledge about human gene function

Objective of the pilot

The proposed Science Demonstrator is based on the Image Data Resource (IDR)[64], an online database of image datasets. IDR holds 41.5 TB of image data and includes all associated experimental (e.g., genes, RNAi, chemistry), analytic, and functional annotations. The main objectives of the Science Demonstrator were twofold:

- Prediction of new cellular functions for human genes using publicly available image data from genome-scale loss of function experiments.
- Assess how the EOSC could be used for image analysis tasks.

Technical capabilities

For this pilot, the following technical capabilities were identified:

- Compute - Allocated sufficient cloud capacity in the EMBL-EBI Embassy Cloud[65] to execute intensive analysis.

Implementation

To analyse publicly available image data from genome-scale RNAi screens and gain insights into cellular functions of human genes, an initial set of cloud resources were allocated at the EMBL-EBI Embassy Cloud. From a technical standpoint, the resources allocated to perform machine learning analyses of large image datasets included: up to 250 vCPU cores with up to 32-64 GB of RAM, and 1TB of NFS-like storage.

Achievements

The SD successfully deployed a cluster of 192 vCPUs with 4 GB of RAM in the OpenStack-based EMBASSY cloud to compute numerical features from 2M images made available in the EMBASSY cloud by the Image Data Resource.

The lessons learned from the project are:

- Cloud infrastructures did not offer a high performance shared file system. This requested additional work to refactor the application and meet the cloud in/out-bound connectivity.

[64] http://idr.openmicroscopy.org

[65] https://www.embassycloud.org/

As further recommendations provided by the Science Demonstrator, it would be beneficial if EOSC can:

- Have orchestration tools/services to facilitate the provisioning of resources and the execution of tasks.
- Offer generous resources allocations and a high-performance file system because image analysis pipelines require data to be in files and can have I/O intensive components.

2.3.5. Switching on the EOSC for Reproducible Computational Hydrology by FAIR-ifying eWaterCycle and SWITCH-ON

Objective of the pilot

The FAIR-ifying eWaterCycle and SWITCH-ON Science Demonstrator aimed to create a fully FAIR Hydrology forecasting system taking into account the best practices and lessons learned from the eWaterCycle[66] and SWITCH-ON[67] projects. The main input data for hydrological models is historical weather forecast data. Size of these weather forecast data range from 1 to 10 GB per day. There are two distinct modes of running hydrological models:

- In the “scenario mode”, a historic or future possibility is simulated. A typical scenario generates 2TB of output datasets.
- In the “forecasting mode”, the eWaterCycle system generates 500 GB of output datasets per day.

Technical capabilities

For this pilot, the following technical capabilities were identified:

- Data Management - Access to Onedata Data platform provided by EOSC to offer transparent data access to the datasets.

Implementation

To facilitate the store, discover and retrieve of the resulted data generated by hydrological models, and made them available for further analysis and visualization in a notebook-like environment, the OneData software technology, developed by CYFRONET, was adopted. To collect all the requirements, a technical meeting with the Cyfronet team was organized by WP5.4. During the meeting was agreed on the initial set-up for supporting the pilot service. From a technical standpoint, a dedicated oneprovider, with an initial storage space of 500GB was configured at INFN-CNAF. WP5.4, in collaboration with WP6.3, facilitated the identification of the service provider for hosting the resulted data produced by the models and the access to the resulted data with Onedata.

The high-level architecture used to implement a FAIR hydrology forecasting systems is shown in Figure 14.

As further recommendations provided by the Science Demonstrator, it would be beneficial if EOSC can:

- Actively engage dataset providers for creating a FAIR version of each dataset. Individual researchers are incapable of doing this themselves , mostly due to the lack of time. This can be done by fostering the creation of new intermediate roles such as: Data Stewards and Research Software Engineers.
- Provide a more coherent set of high level services, including solutions for: running standard-based workflows, file sharing of large data sets and use of DOIs.

[66] https://www.ewatercycle.org/

[67] http://www.water-switch-on.eu/

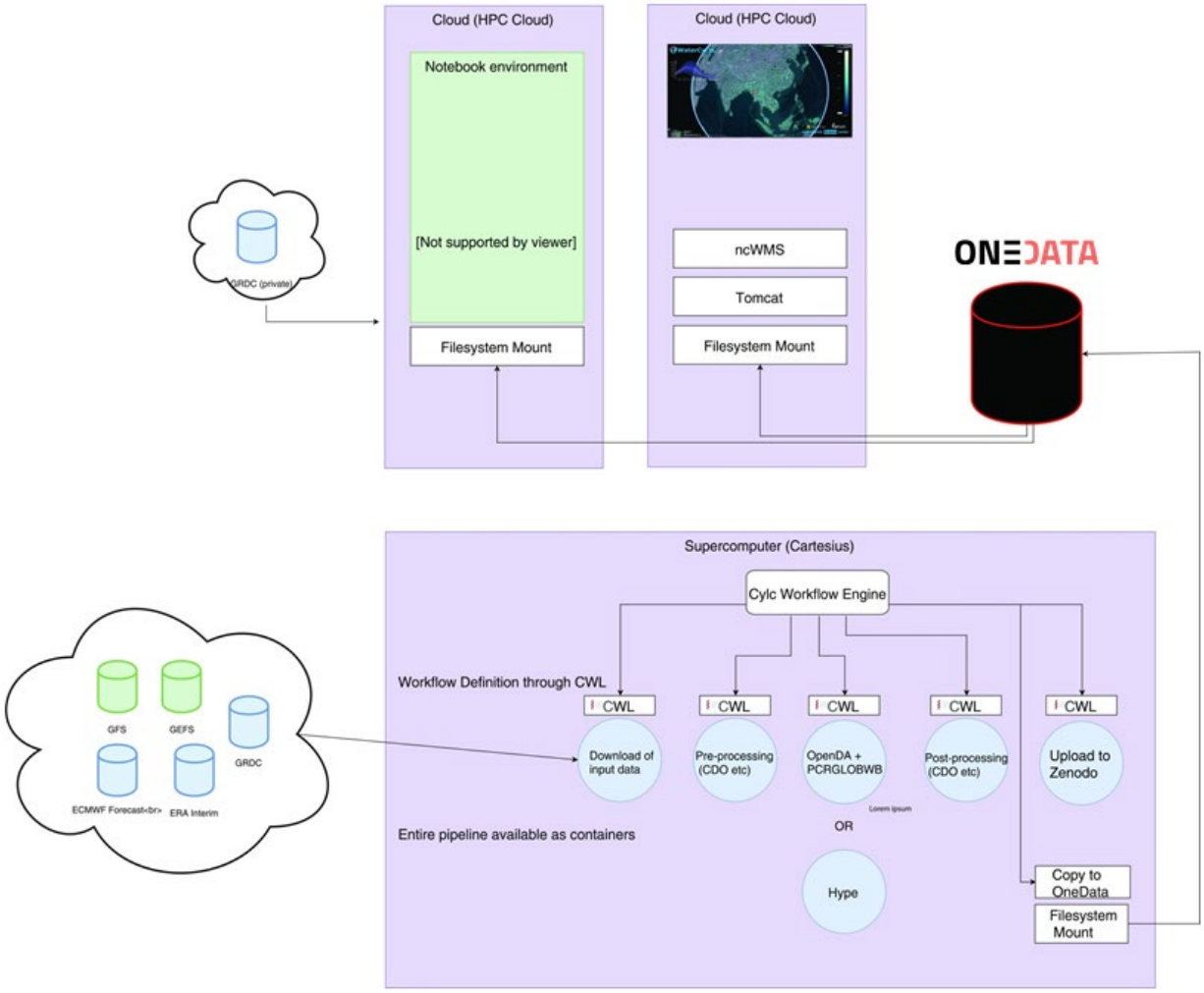

Figure 14 - Overview of the high-level architecture set-up for the FAIR-ifying eWaterCycle pilot.

3. ANALYSIS AND RECOMMENDATIONS

3.1. Analysis of Demand

Service, Data and Infrastructure providers contributed to the implementation of the pilot services providing computing, storage resources and high-level solutions to target the Science Demonstrators needs identified during the funding periods. Overall, the list of EOSC services and resources adopted by the selected Science Demonstrators during the implementation of the pilot services is shown in Figure 15. This classification takes into consideration the services and resources categorization available in the EOSC portal[68].

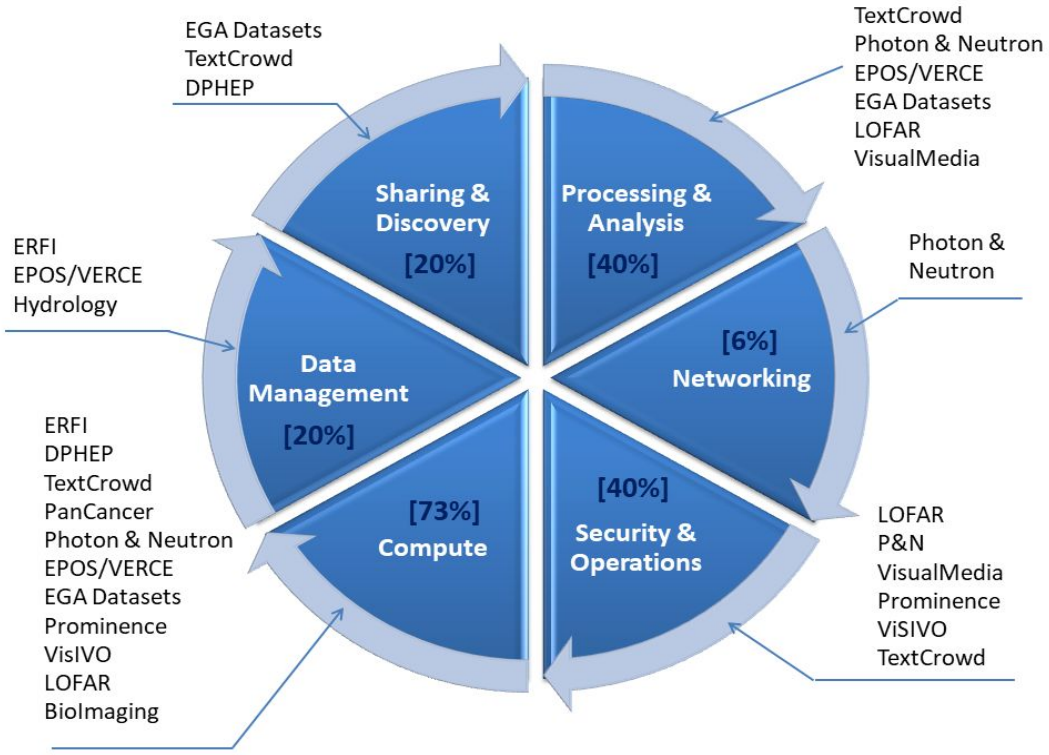

Figure 15 - Overview of the EOSC solutions adopted by the pilots.

For each Science Demonstrator WP5.4 has also collected technical requirements and identified recommendations to tackle these requests. The full list of requirements and recommendations is reported in Table 3.

[68] https://eosc-portal.eu/

3.2. Recommendations

Requirement	Recommendation
Provide storage and data management solutions for scalable access to large data sets.	Implement a distributed data and compute infrastructure providing high performance access to distributed big data infrastructures and mechanisms for data mirroring and caching, supported by high-speed network connectivity.
Allow research communities to scale-up the in-house digital infrastructure and provide adequate capacity.	Procure and offer EOSC as a federated infrastructure that integrates existing community resources (data, applications, software, storage and computing) and provides additional adequate capacity to scale up existing in-house IT infrastructures. Procure EOSC as a high capacity system that meets the demands of data intensive science.
Simplify the access to heterogeneous services.	Provide easy to use environments like scientific gateways, VREs and data exploitation platforms as managed services that provide the required integration as turn-key solution; make service descriptions semantically rich and discoverable; offer ready to use integrated bundles of services with low-barrier procurement processes.
Interoperability between the different AAI solutions	Provide a managed federated AAI solution to allow users to access services and resources from different infrastructure providers.
Human support	Provide and sustain human networks through competency centres of experts working with scientific application developers in close cooperation.
Support for workflows	Provide support for running standard-based workflows.
Integration at the level of services and e-Infrastructure providers	Promote a tighter integration between service and Infrastructure providers for helping research communities to plug into their working environments, the available EOSC services and resources to support their day-by-day work.
Promote the development of FAIR services in EOSC	Extend the FAIR concepts currently applied to data also to IT services. Propose a set of recommendations for making services FAIR, or to further enable services to make data FAIR.
Enable the interoperability of cross-domain services	Include analysis of the underlying network requirements specifically when designing the interoperability of services across sites and organisations.

Table 3 - List of requirements and recommendations.

4. CONCLUSIONS

This deliverable provides a final evaluation report of the fifteen EOSCpilot Science Demonstrators selected through the open calls carried out by WP4.

Overall, the report shows that:

- The majority of the Science Demonstrators successfully developed their service pilots in the due time according to the agreed work-plans.
- Five Science Demonstrators were selected to further extend the pilot solutions and move into pre-production services during the three-month WP5.4 extension (from January to March 2019).
- Two Science Demonstrators experienced minor technical and organization/procedural issues that prevented the full implementation of their original work-plans.

For each Science Demonstrator WP5.4 has collected user experiences and lessons learnt from the execution of the pilot services, along with recommendations for shaping the EOSC architecture and driving the technological evolution of the EOSC services to meet the functional and non-functional needs of researchers. These recommendations complement the ones collected during the First Stakeholder Forum[69] and the Second Stakeholder Forum[70].

The final feedback collected by WP5.4 confirms that, even if the EOSCpilot Science Demonstrators had been selected from different scientific domains, the following set of cross-domain recommendations will still be relevant:

- In order to support a cross-discipline interoperability it is necessary for EOSC to promote a tightened integration between service and Infrastructure providers. Even if some integration activities have already started in the EOSC-hub project (e.g.: users can login to EUDAT services with the EGI AAI Check-In service) there is still a lot of work to be done. Interoperability between the different AAI infrastructures, in different organisations, is an important issue that needs to be solved in order to simplify access to both compute and storage resources. Without a fully working integration of services and resources, users have to deal with different protocols and access mechanisms. This is an additional overhead for the end-users that prevent the fully exploitation of the EOSC services for supporting advanced data-driven research.
- The analysis of the underlying network requirements is fundamental aspect that must be taken into account to make cross-domain services interoperable.
- Many EOSCpilot Science Demonstrators highlighted the need to have more support for Workflow Languages. The two solutions mostly adopted in this context are the Common Workflow Language (CWL) and NextFlow. During the execution of the project, WP5.4 promoted the adoption of the CWL community-based standard.
- Extend the FAIR concepts currently applied to data also to IT services. Propose a series of recommendations for making services FAIR, or to further enable services to make data FAIR.
- Availability of other Platform as a Service type services such as: databases and queues, would be beneficial, as well as a catalogue of datasets.

[69] https://eoscpilot.eu/events/eosc-stakeholder-forum-shaping-future-eosc

[70] https://eoscpilot.eu/events/second-eosc-stakeholders-forum

ANNEX 1 – TECHNICAL TALKS ORGANISED BY WP5.4

To address the functional gaps raised by the selected Science Demonstrators during the project, the following technical talks were organized by T5.4:

- **Introduction to OpenAIRE services**[71]
  - Paolo Magni, from the OpenAIRE project, introduced the OpenAIRE e-Infrastructure and the services to enable researchers, content providers, funders and research administrators to easily adopt open science.

- **Introduction about CVMFS**[72]
  - Catalin Coundarache from STFC reported about how the scalable, reliable and low-maintenance software distribution service can be used in the context of the pilot for software preservation.

- **The EGI Open Data Platform**[73]
  - Lukasz Dutka from Cyfronet provided a high-level overview about the Onedata software stack and showcase the solution with a live demo.

- **The eInfraCentral mission to align existing service catalogues**[74]
  - Jelena Angelis from the European Future Innovation System (EFIS) Centre, together with Jorge Sanchez and Thodoris Ntezes from JNP, provided a high-level overview of the project and explained how the project are helping research to facilitate advanced research making easier the discovery of e-Infrastructure services.

- **The EOSC-hub Service Catalogue and overview of the high-level services**[75]
  - Gergely Sipos from EGI Foundation introduced EOSC-hub and the EOSC-hub Service Catalogue with the initial set of existing mature services to support the entire research lifecycle.

- **The Common Workflow Language project**[76]
  - Michael Crusoe, the CWL co-founder and project leader, introduced this community-based standard.

- **Introduction to the EUDAT CDI and B2 services suite**[77]
  - Mark van Sanden from SURFsara introduced the EUDAT services suite.

[71] https://indico.egi.eu/indico/event/3393/
[72] https://indico.egi.eu/indico/event/3513/
[73] https://indico.egi.eu/indico/event/3588/
[74] https://indico.egi.eu/indico/event/3898/
[75] https://indico.egi.eu/indico/event/3898/
[76] https://indico.egi.eu/indico/event/3547/
[77] https://indico.egi.eu/indico/event/3966/

ANNEX 2 – GLOSSARY

The definitions below shall be considered for the purpose of this deliverable.

Term	Explanation
e-Infrastructures	(Definition of the Commission High Level Expert Group on the European Open Science Cloud in their report): this term is used to refer in a broader sense to all ICT-related infrastructures supporting ESFRIS (European Strategy Forum on Research Infrastructures) or research consortia or individual research groups, regardless of whether they are funded under the CONNECT scheme, nationally or locally.
High Performance Computing (HPC)	(EGI definition) A computing paradigm that focuses on the efficient execution of compute intensive, tightly-coupled tasks. Given the high parallel communication requirements, the tasks are typically executed on low latency interconnects which makes it possible to share data very rapidly between a large number of processors working on the same problem. HPC systems are delivered through low latency clusters and supercomputers and are typically optimised to maximise the number of operations per second. The typical metrics are FLOPS, tasks/s, I/O rates.
High Throughput Computing (HTC)	(EGI definition) A computing paradigm that focuses on the efficient execution of a large number of loosely-coupled tasks. Given the minimal parallel communication requirements, the tasks can be executed on clusters or physically distributed resources using grid technologies. HTC systems are typically optimised to maximise the throughput over a long period of time and a typical metric is jobs per month or year.
Virtual Organisation	A group of people (e.g. scientists, researchers) with common interests and requirements, who need to work collaboratively and/or share resources (e.g. data, software, expertise, CPU, storage space) regardless of geographical location. They join a VO in order to access resources to meet these needs, after agreeing to a set of rules and Policies that govern their access and security rights (to users, resources and data).
Cloud Computing	The practice of using a network of remote servers hosted on the Internet to store, manage, and process data, rather than a local server or a personal computer.
Data Analysis	Process of inspecting, cleaning, transforming, and modelling data with the goal of discovering useful information, suggesting conclusions, and supporting decision-making.
Open Science	The movement to make scientific research, data and dissemination accessible to all levels of an inquiring society, amateur or professional.
Science Demonstrators	High-profile pilots that integrate services and infrastructures to show the usefulness of the EOSC Services and will drive the further development of EOSC.

Term	Explanation
Science Demonstrator Contact	Contact person from a specific Science Demonstrator. They will work together with the Shepherd assigned to the Science Demonstrator in order to develop the proposed project.
Shepherd	Staff who supports the main contact of an approved Science Demonstrator in order to facilitate the engagement with the EOSCpilot project in establishing their technical use case, software tools, data models and scientific workflows going to be used.
AAI	Authentication and Authorization Infrastructure.
EOSC	The European Open Science Cloud.
FAIR	Findable, Accessible, Interoperable and Reusable.
RDA	Research Data Alliance.
RIs	Research Infrastructures.
Virtual Appliance	A pre-configured virtual machine image, ready to run on a hypervisor.