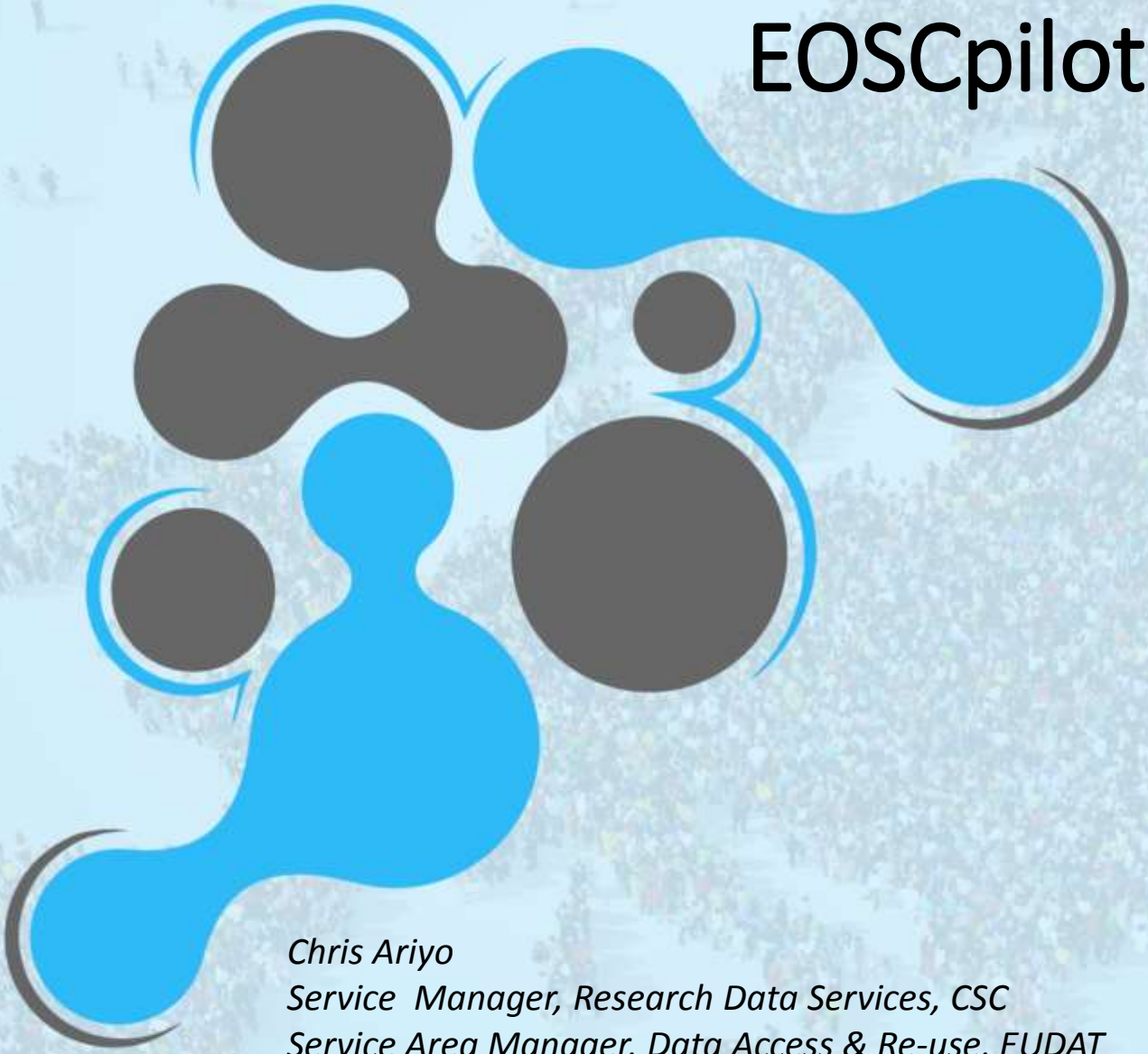
EOSCpilot Updates



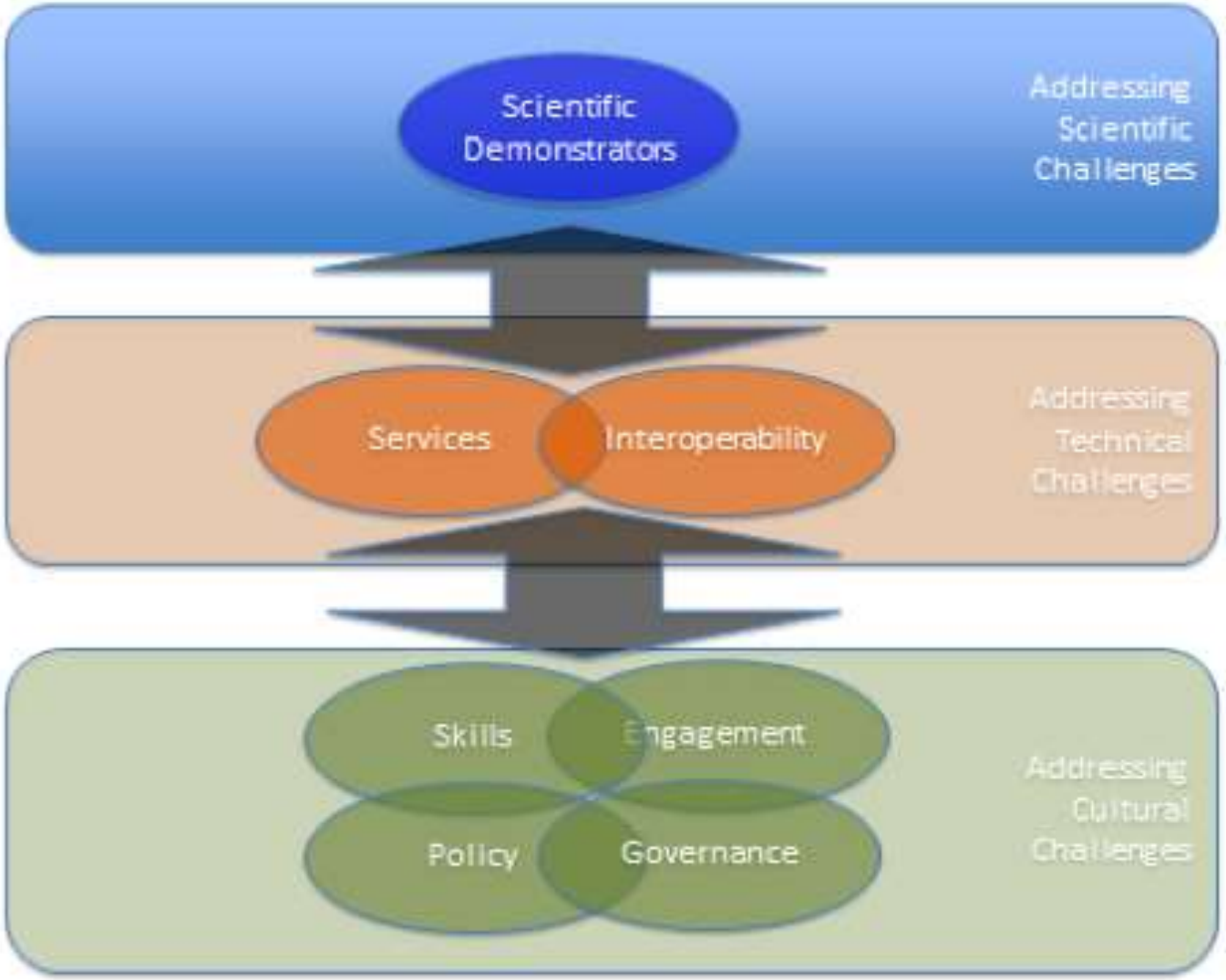
Chris Ariyo
Service Manager, Research Data Services, CSC
Service Area Manager, Data Access & Re-use, EUDAT

EOSCpilot
The European Open Science Cloud for Research Pilot Project
www.eoscpilot.eu

SeaDataCloud, Annual Meeting, Athens, Greece, October 18-19, 2017

EOSC pilot
The European Open Science Cloud for Research Pilot Project
www.eoscpilot.eu

Project overview



WP4 – WP5

PHASE 1 (WP4)

- Science demonstrator objectives, specifications, requirements

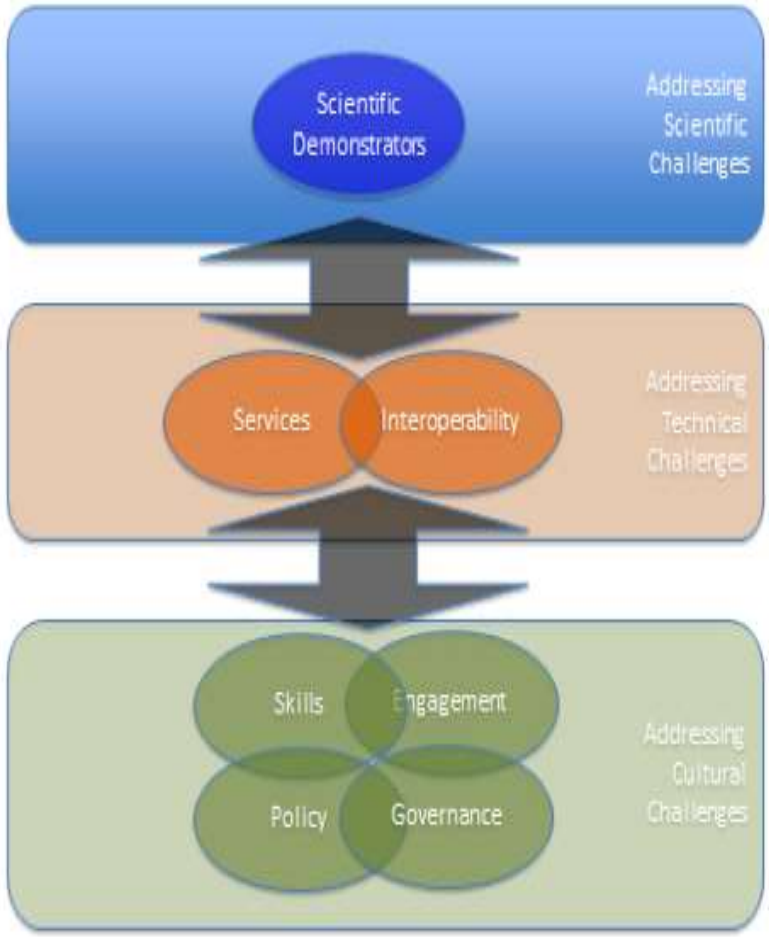
PHASE 2 (WP5.4)

- Analysis of requirements, gap analysis
- Service pilot design
  - → both assisted by WP4 shepherd
- Development of technical bridges
- Testbed implementation
  - Access
  - Coordinated service provisioning
  - Support

(WP6.3: provisioning of generic e-Infrastructure components, validation of functional/non functional requirements, test suites)

PHASE 3 (WP4)

- Validation reports



Science Demonstrators in EOSCpilot

Challenging projects helping to define the infrastructure needed by European researchers, while showing the scientific excellence and societal impact that could be achieved by EOSC (European Open Science Cloud).

- Supporting 15 Science Demonstrators (SDs) in EOSCpilot, 12 months each
- First five SDs: Preselected from 29 proposals prior to EOSCpilot begin, start: Jan 2017
- Second five SDs: Selected after 1st Open Call in April 2017 from 30 proposals, start: July 2017
- Third five SDs: To be selected after 2nd Open Call in Aug/Sep 2017 from 26 proposals, start: Dec 2017

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

Science Demonstrator PanCancer

SD ID: Pan-Cancer Analyses & Cloud Computing within the EOSC

ORGANISATION: Genome Biology Unit, European Molecular Biology Laboratory (EMBL)

CONTACT: Sergei Iakhnin

Email: llevar(at)gmail.com

OVERVIEW:

There are a multitude of challenges being faced in the life sciences, health, food, fishery and agriculture sectors. In cancer research, Europe has taken a technical lead within international consortia around cloud-based pan-cancer genomic analysis. This global competitive advantage can be maintained by leveraging open science analysis models around controlled access data sets developed in collaboration with researchers elsewhere in the world. These analysis frameworks could also be re-used to analyse cardiovascular and neuro-degenerative diseases as well as stimulating biotech/pharmaceutical industries to use public cancer genomic data in R&D.

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

Science Demonstrator PanCancer

Major European Cloud Use-Case in Genomics:

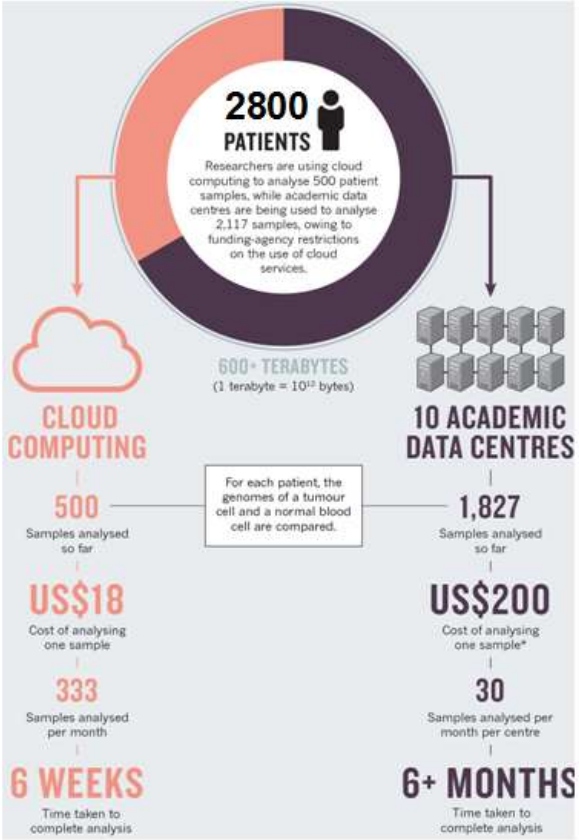
The Pan-Cancer Analysis of Whole Genomes Project (PCAWG)

- **Mission of PCAWG phase I**: joint reanalysis of **2,800 cancer genomes**, 1 Pb of DNA data, using hybrid clouds.
- **Global network** (includes USA, Canada, Japan, Korea, Europe)
- **European leadership** (until recently co-funded by EC / Eurocan-Platform).

- Our initial focus has been on **standardizing** genomic **data processing**, and data redistribution, **on the cloud**.
- We **aim to make these processes inter-operable for the EOSC**.

- Outcomes of PCAWG phase I are intended to be published in **2017**.

The International Cancer Genome Consortium (ICGC)



Stein, Knoppers, Campbell, Getz & Korbel, *Nature* 2015

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

Science Demonstrator TEXTCROWD

SD ID: TEXTCROWD - Collaborative semantic enrichment of text-based datasets

ORGANISATION: PIN srl

CONTACT: Franco Niccolucci

Email: franco.niccolucci(at)gmail.com

OVERVIEW: The Social Sciences and Humanities research communities face a fragmented research landscape that can be supported by EOSC. The EOSC would help overcome such fragmentation, by building on structuring and integrating initiatives such as the CLARIN, DARIAH and E-RIHS ERICs, and Digital Humanities Organizations (e.g. their Association ADHO) to offer advanced text-based services addressing common research needs (see recent survey by PARTHENOS). One example is enabling the semantic enrichment of text sources through cooperative, supervised crowdsourcing, based on shared semantics, and then to make this work available to others via EOSC. This would benefit many scientists in the long-tail even if delivering such a service presents real challenges around interoperability and multilingualism.

CONTEXT: Cultural heritage and humanities datasets are largely based on texts:

Reports

Archaeology: excavations, surveys

Conservation: diagnosis, restoration – often mixed with numeric results

Grey literature

Literary/historical sources

Research articles

Monographs

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

Science Demonstrator TEXTCROWD (cont.)

Issues to solve

- Access to these data (partially) solved with online publication
  - Different repositories, access rules, fragmentation
- Discovery still problematic:
  - Resource registries: only the general scope of the text is described, according to a dataset model
  - Free-text search within specialized repositories: cumbersome, not always available, different repositories, incoherent results
  - In general, poor metadata for content
  - Need to add semantics of content
  - Indexing and metadata creation by hand is time-consuming and may be inconsistent/unreliable
  - Multilingualism
- Text mining with NLP and Information Extraction a solution

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

Science Demonstrator ENVRI/ERFI

- SD ID: ENVRI Radiative Forcing Integration
- Organisations & Contacts:
- Werner Kutsch, Alex Vermeulen, ICOS ERIC
- Ari Asmi, ENVRIplus, Paolo Laj, ACTRIS
- Stefan Kindermann, IS-ENES2 (DKRZ)
- Sylvie Joussaume, Sébastien Denvil, IS-ENES2 (IPSL)

SCIENTIFIC OBJECTIVES OF THE DEMONSTRATOR:

Focus on dynamics of greenhouse gases, aerosols and clouds and their role in radiative forcing,
Interoperability between observations and climate modeling,
Cooperation between environmental research infrastructures

TECHNICAL FOCUS:

Improvement of data integration services based on metadata ontologies,
Model-data integration by use of HPC,
Petascale data movement,
Innovative services to compile and compare model output from different sources, especially on semi-automatic spatiotemporal scale conversion

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

Science Demonstrator ENVRI/ERFI (cont.)

FAIR CHALLENGES:

Findability: Metadata ontologies matching between NETCDF-CF and in-situ metadata, data quality indicators.

Accessibility: Automated access routines between the RI repositories. For fully open data, this is not immediately problematic, but might require analysis on needed resources and APIs.

Interoperability: APIs, service integration, large data transfers, where to do processing (how to document?)

Reusability: Citing and persistently identifying scale-changed data-sets? How to transfer knowledge of data versions used.

ADVANTAGES:

MATURITY: All participating RIs have existing data systems, and their "act together"

INTERDICIPLINARITY: The case provides direct interdisciplinary activity (even though inside the environmental domain)

SYNERGY: This action directly uses the connections and needs of the existing RI actions within the cluster project ENVRIplus

ENGAGING COMMUNITY: The case is actually useful for the scientist (and RIs) in the field! It is not just a demonstrator!

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

Science Demonstrator High Energy Physics

SD ID: WLCG / DPHEP
ORGANISATIONS: CERN
CONTACT: Jamie Shiers, CERN
Email: Jamie.Shiers(at)cern.ch

OVERVIEW:

Funding agencies today require (FAIR) Data Management Plans, explaining how data acquired or produced will be preserved for re-use, sharing and verification of results. The preservation of data from CERN's Large Hadron Collider poses significant challenges: not least in terms of scale. The purpose of this demonstrator is to show how existing, fully generic services can be combined to meet these needs in a manner that is discipline agnostic, i.e. can be used by others without modification.

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

Science Demonstrator High Energy Physics (cont.)

- ***OBJECTIVE***:
- The high energy physics science demonstrator wants to deploy services that tackle the following functions:
- Trusted / certified digital repositories where data is referenced by a Persistent Identifier (PID);
- Scalable “digital library” services where documentation is referenced by a Digital Object Identifer (DOI);
- A versioning file system to capture and preserve the associated software and needed environment;
- A virtualised environment that allows the above to run in Cloud, Grid and many other environments.

- ***TECHNICAL FOCUS:***
- The goal is to use non-discipline specific services combined in a simple and transparent manner (e.g. through PIDs) to build a system capable of storing and preserving Open Data at a scale of 100TB or more.

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

Photon-Neutron Science Demonstrator

SD ID: Photon-Neutron Science Demonstrator

ORGANISATIONS: DESY, EMBL, ESRF, EU-XFeL, ESS, ILL, STFC

CONTACT: Volker Guelzow, DESY

Email: volker.guelzow(at)desy.de

OVERVIEW:

The Photon Neutron Data Science Demonstrator will leverage on the photon-neutron community to improve computing facilities by creating a virtual platform for all users.
The Photon-Neutron science demo is based on the concept that:
Data sets become too large to take home;
The EOSC forms a platform for analysis and data storage
The EOSC allows for easy data sharing
Data rates require dedicated central IT infrastructure, way beyond previous requirements;
A wide variety of scientific users means significant number of data formats and analysis software.

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

Photon-Neutron Science Demonstrator (cont.)

OBJECTIVES:

Exploiting a community of more than 35,000 unique users (in 2011), the science demonstrator aims to enable cloud based storage and compute solutions, foster standardized data formats and allow transparent and secure remote access to scientific data. We will focus for this demonstrator on a particular data analysis framework outlined in the diagram. The crystfel framework is increasingly used at various synchrotrons and FELs to analyze date from serial (femto-second) x-ray crystallography. The nature of these experiments make a cloud-based distributed pipeline particularly appealing, since the framework can fully exploit large computational resources with tunable demands. The framework is well documented and vast amount of data are readily and openly available.

TECHNICAL FOCUS:

- Exploit and improve the crystfel framework for distributed computing.
- Provide compatible data analysis software
- Allow transparent and secure remote access to data
- Standardize data formats NeXus/HDF5 and annotation of data
- Test and establish (if feasible) web-services for easy consumption and visualization of the data
- Exploit existing authentication and authorization solutions
- Allow long term preservation of data
- Promote data policies in laboratories

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

Second five Science Demonstrators in EOSCpilot

- Leveraging EOSC to offload updating and standardizing life sciences datasets and to improve studies reproducibility, reusability and interoperability
  - *Principal Investigator: Dr. Jordi Rambla, CRG Center for Genomic Regulation (Spain)*
- PROMINENCE: High-Performance Computing as a Service (HPCaaS) for Fusion
  - *Principal Investigator: Dr. Rob Akers, Culham Science Centre (United Kingdom)*
- EPOS/VERCE: Virtual Earthquake and Computational Earth Science e-science environment in Europe
  - *Principal Investigator: Prof. Andreas Rietbrock, University of Liverpool, (United Kingdom)*
- CryoEM workflows: Linking distributed data and data analysis resources as workflows in Structural Biology with cryo Electron Microscopy: Interoperability and reuse
  - *Principal Investigator: Dr. Jose Maria Carazo, I2PC Instruct Image Processing Center (Spain)*
- Open Science Cloud access to Low-Frequency Array (LOFAR) data
  - *Principal Investigator: Dr. Rob van der Meer, ASTRON (The Netherlands)*

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

Leveraging EOSC to offload updating and standardizing life sciences datasets and to improve studies reproducibility, reusability and interoperability

Brief Summary of the Science Demonstrator:

- Services like the European Genome-phenome Archive (EGA, https://ega-archive.org) or big consortia portals offer life science 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. Probably, every new requestor of such datasets would start their analysis by re-mapping with a current reference genome and running current gold standard tools or pipelines like the Pan-Cancer one. It is quite probable that, when possible because of huge resources involved, the requestor would like to reproduce the results from the original project.
- Indeed, as most projects do not have a proper data management strategy, data are not made available according to the FAIR principles.
- The EGA would like to avoid such wasteful and redundant processing by providing refreshed versions of the original datasets, while increasing the homology and interoperability between data for crossed analyses, in a similar way that the Pan-Cancer project applies a common pipeline to datasets initially processed by each individual cancer project. When offloaded, the process should happen following high privacy preserving procedures like the ones used by EGA. This demonstrator will show how to minimize such wasting of time and resources by leveraging EOSC to offload the computing burden that such up-to-date process will put on EGA operational resources. It will foster reproducibility by packaging the original pipeline into standardized workflows, and using container technology to make such packages portable across different platforms. The new pipeline will be packaged in homologous ways, fostering reuse and reproducibility. As well, the process would be leveraged to apply FAIRification pipelines. The results will be send back to the EGA, thus every new user will benefit from those re-mastered datasets, while dataset discoverability will increase accordingly.

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

PROMINENCE: High-Performance Computing as a Service (HPCaaS) for Fusion

Brief Summary of the Science Demonstrator:

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).The requirements for HPC class machines is expected to only increase as the community preparesfor the next generation facility, ITER. However, access to HPC class infrastructure is quite restricted and obtaining time on this class of machine for algorithm development, testing and optimisation is already problematic. A few fusion centres have local access to smaller HPC class clusters but larger scale workflows and smaller fusion research centres are required to competitively bid for time on hardware such as PRACE Tier-0 and Tier-1 facilities, often also requiring visiting the centre.

Within this demonstrator we propose to make HPC class machines available as a cloud like service to the fusion community, in a similar way to those available through AWS (https://aws.amazon.com/hpc) and Microsoft Azure (https://azure.microsoft.com/en-gb/solutions/big-compute). This is currently not a feature offered by any of the current e-infrastructures (EUDAT, EGI, Indigo-DataCloud, PRACE, etc), but has been investigated in other H2020 and FW7 funded projects (cloudSME, HOST). Building and collaborating with these projects and the current infrastructure projects we aim to demonstrate that by using an industry standard Openstack instance at CCFE we can make HPC class nodes together with fast interconnects accessible to other members of the fusion community to exploit. The pilot infrastructure will be made available to the Eurofusion MST community for joint research (covering four large scale facilities: MAST-U at Culham, WEST at CEA, ASDEX Upgrade at IPP Garching and TCV and EPFL/CRPP Lausanne).

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

EPOS/VERCE: Virtual Earthquake and Computational Earth Science e-science environment in Europe

Brief Summary of the Science Demonstrator:

- Several seismological events, not last the 2016 destructive Amatrice earthquake sequence (Italy), that are still causing victims and damages to historical sites, have shown again the urgent need to understand the complexity of the underlying processes of an earthquake rupture/sequence. In the last decade, the availability of open access and high quality seismic observations has increased exponentially (100s of TBs). Open Source and community supported 3D seismic simulation tools in complex 3D media have become available (also supported by advances in HPC) and are now used by an increasing number of researchers. However, combining large amount of high quality data with complex 3D seismic simulations is still extremely challenging and has been accomplished only for a few regions on Earth. Furthermore, civil protection agencies have an increasing need of computing realistic scenarios of earthquake shaking to aid emergency planning and coordination of rescue efforts.
- The VERCE project has pioneered a VRE to support researchers using established simulation codes on high-performance computers in conjunction with multiple sources of observational data. This is accessed and organised via a science gateway that makes it convenient for seismologists to use these resources from any location via the Internet. Their data handling is made flexible and scalable by community-developed libraries, such as ObsPy (http://obspy.org), and data-intensive tools, such as dispel4py (https://github.com/dispel4py/dispel4py). It connects to federated data services of the FDSN (http://fdsn.org) to discover and ingest observational raw and parametric data delivered by worldwide seismological networks. Provenance driven tools (S-ProvFlow, https://github.com/KNMI/s-provenance) enable the rapid exploration of the results and of the relationships between data processes and users, which accelerates understanding and method improvement.

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

CryoEM workflows: Linking distributed data and data analysis resources as workflows in Structural Biology with cryo Electron Microscopy: Interoperability and reuse

Brief Summary of the Science Demonstrator:

- Structural Biology ("SB") aims at providing a detailed understanding of the 3D structure of macromolecular machines, many times reaching atomic resolution, as a fundamental step in the understanding of biological function. Among the structural biology (SB) techniques at the core the Research Infrastructure for SB, Instruct, microscopy under cryogenic conditions ("cryo-EM") is currently the fastest growing area, having been nominated "Method of the Year (2015)" by Nature.
- Typically, cryoEM starts with the acquisition of thousands of "movies", extremely noisy and large, at specialized facilities. Multiple image processing operations are then performed on these images by many different data analysis tools until a quantitative 3D electrostatic map is obtained, conceptually forming an "image processing workflow". Individual processing steps may be performed at the experimental facility, at supercomputer centres or in the scientist's home instutition, so that typically the workflow is geographically distributed.
- Public databases exist for the deposition of 3D maps (Electron Microscopy Data Bank (EMDB)), as well as for the deposition of key supporting evidence (EMPIAR). However, the way from the movies to the maps is full of case-dependent methodological image processing choices (the elements of the "image processing workflow"), linking different data sources at different sites and using a variaty of software with multiple versions. There is no standard method for recording these steps, and they are currently not deposited anywhere. Some details obviously can be found in publications, but they are far from being complete and, in any case, their description changes from author to author. Mining these heterogeneous data sets could bring new light to best practices and yet unknown data and analysis bottlenecks, but this currently impossible.
- In this Science-Demonstrator we want to address the proper reporting of cryoEM image processing workflows, ensuring provenance at the level of data and analysis tools, linking workflow information with raw data either at cryoEM facilities, individual laboratories or public repositories, so that reproducibility of scientific results were enhanced, data and analysis workflows could be reused and properly mined, allowing for a deeper level of interoperability among information sources.

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

Open Science Cloud access to Low-Frequency Array (LOFAR) data

Brief Summary of the Science Demonstrator:

The goal of this Demonstrator is to allow for the science community to locate, access, and extract science from the LOFAR archive without being an expert on data retrieval and data analysis tools. At the moment the LOFAR data archive is operational and mostly used by experts.

The pilot will develop services, based on existing tools such as Xenon, CWL, Docker, Virtuoso, that allow users to initiate processing on data stored in a distributed, large-scale archive. This implies the ability to define and run custom processing tasks on multiple heterogeneous hardware platforms using data in multiple storage locations. The system will be based on workflows standards (CWL) and Containers (Docker, Singularity) to ensure reproducibility, and increase FAIRness of the system as a whole. The resulting system will be available for any scientist with access to the EOSC, and usage of SURFsara and the other ILT data centers will ensure ample storage and compute resources.

As a result, users can easily create new scientific results based on archived data products. It provides users with large-scale compute resources not likely to be available at their home institutions. It produces an overall multiplication of the total science output of the LOFAR archive. It also demonstrates how to arrange and organize comparable data archives and analysis infrastructures in the context of the EOSC, as well as enabling FAIR access for the first time in this domain.

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

Contact Information

Thank you

Damien Lecarpentier Damien.Lecarpentier@csc.fi

Project Director, Research Infrastructures, CSC

EUDAT Project Director, EOSCpilot WP5 Services

Hermann Lederer lederer@rzg.mpg.de

Max Planck Computing & Data Facility (MPCDF)

EOSCpilot WP4 Science Demonstrators (co-leader)

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