D6.4: Initial Requirements of the Interoperability Testbeds

Author(s)	Doina Cristina Duma (INFN)
Status	first version for internal review
Version	V1.0
Date	11/02/2018

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 objectives of this deliverable are to present the initial requirements of the Interoperability Testbeds as a result of the analysis of the gap analysis on technical and political barriers that prevent the interconnection of e-infrastructures and the application of the FAIR principles, the first architectural design of the interoperation of various types of infrastructures, which could participate to the future EOSC, the requirements regarding interoperability aspects expressed by the projects' fifteen Science Demonstrator and the four proposed Data Interoperability Demonstrator, taking into consideration also interoperability principles and description of different interoperability layers present in the New European Interoperability Framework.

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 -WP6-D6.4	
Deliverable lead	**INFN**
Related work package	**WP6**
Author(s)	D. C. Duma
Contributor(s)	T6.3 team
Due date	**31/12/2017**
Actual submission date	**28/02/2018**
Reviewed by	**Giuseppe La Rocca (EGi Foundation)**
Approved by	**Brian Matthews**
Start date of Project	**01/01/2017**
Duration	**24 months**

Versioning and contribution history

Version	Date	Authors	Notes
0.1	05/12/2017	Doina Cristina Duma (INFN)	ToC available
0.2	12/12/2017	Doina Cristina Duma (INFN)	Extended, final ToC
0.3	17/12/2017	Xavier Jeannin (RENATER)	Added Networking Req. , PiCo2 description
0.4	18/12/2017	Michael Schuh (DESY)	Added info on Photon & Neutron SD, and VisIVO
0.5	01/01/2018	Doina Cristina Duma (INFN), Daniele Spiga (INFN)	Added sections Execution, Introduction, Grid-Cloud Interoperability
0.6	10/01/2018	Doina Cristina Duma (INFN)	Improvements in various sections
0.7	12/01/2018	Doina Cristina Duma (INFN), Xavier Jeannin (RENATER), Alain Franc (INRA), Violaine Louvet (Univ. Grenoble ALPES)	Improved Networking section, added Data Interoperability section
0.8	16/01/2018	Doina Cristina Duma (INFN), Giuseppe La Rocca (EGI.eu)	Added info on SDs & ESOCpilot-AARC collaboration

1.0	19/01/2018	Doina Cristina Duma (INFN), Giuseppe La Rocca (EGI.eu), Thomas Zastrow (MPG), Erik van den Bergh (EMBL-EBI), Dario Vianello (EMBL-EBI), Giovanni Morelli (CINECA)	Added info on SDs
1.02	11/02/2018	Doina Cristina Duma (INFN), Giuseppe La Rocca (EGI Foundation)	Improvements following suggestion from reviewer (Giuseppe La Rocca)

TABLE OF CONTENT

EXECUTIVE SUMMARY

In the context of the EOSCpilot project, the Interoperability Work Package (WP6) aims to develop and demonstrate the interoperability requirements between e-Infrastructures, domain research infrastructures and other service providers needed in the European Open Science Cloud. It deals with interoperability in general leading to the ability to "plug and play" the services of the EOSC in the future. While the *Science Demonstrators* will show how some particular services can work together, we also need to work on a more general framework for interoperability between data and services. This WP works closely with WP4 (Science Demonstrators) and WP5 (Services), taking inputs from the science demonstrators to ensure the relevance of the interoperability framework, and from the Services.

The interoperability was mapped in two tracks:

- Research and Data Interoperability:
    - The Research and Data Interoperability Track – that provides the research infrastructure and domain expert view in the work programme with focus on data interoperability. The definition of a Data Interoperability framework in EOSC is based on the FAIR principles - data and services need to be Findable, Accessible, Interoperable and Reusable.
- Infrastructure Interoperability:
    - The complementary usage of Cloud, Grid, HTC and HPC infrastructures, including large datastores, through high speed networks and performant data transfer protocols and tools. The high level objective is to facilitate the most adequate infrastructures for the treatment of extensive amounts of data, generated by new generations of instruments, observatories, satellites, sensors, sequencers, imaging facilities and numerical simulations, and produced by well-known data intensive communities but also by the long tail of science.
    - In the Infrastructure Interoperability track the provider view is in the centre of the work programme.
    - The federated infrastructure pilots that have to be set up with the resources provided by other partners involved in this WP and by the selected Science Demonstrators will enable the analyses of the existing interoperation mechanisms for software components, services, workflows, users and resource access within existing RI systems.

The above objectives are envisioned to be implemented through the instantiation of multi-infrastructure, multi-community pilots. Services and the Science Demonstrators defined in WP4 and WP5 have been deployed and validated in these pilots from the standpoint of maturity, scalability, and usability for a future EOSC.

The main objective of this deliverable is to describe the initial requirements of the interoperability testbeds

that have to be set up in order to meet the needs of the Science Demonstrators and Data Interoperability Demonstrators, to show the interoperability among infrastructures, solutions and services proposed.

In this first report we focus on the analysis of the interoperability aspects and requirements from the point of view of e-infrastructures and data interoperability. Broadly speaking we start from the aspects identified by the gap analysis (detailed in D6.1) like networking, authorization and authentication, use of grid and cloud technologies. We take into consideration the interoperability principles and description of different interoperability layers present in the New European Interoperability Framework [reference]. We analyze also the workplans of the first two sets of selected Science Demonstrators, and finish with a short summary of the last 5 Science Demonstrators, yet to start their activities. This report provides as a result an initial list of requirements that guided and will continue to guide the setup of the first interoperability pilots, and the validation of the services and proposed solutions.

The results of the use and the validation of the different pilots, enriched with eventual new requirements and changes to be applied, together with the feedback received from the different stakeholders involved will be described in the second deliverable, D6.5 – "Interim Interoperability Testbed report". The activity will conclude with the last report, the deliverable D6.10 – "Final Interoperability Testbed report", with results and final feedback on the interoperability testbeds.

INTRODUCTION

The EOSCpilot project has been funded to support the first phase in the development of the European Open Science Cloud (EOSC). One of the main three objectives of the project is to develop a number of demonstrators functioning as high-profile pilots that integrate services and infrastructures to show interoperability and its benefits in a number of scientific domains such as: Earth Sciences, High-Energy Physics, Social Sciences, Life Sciences, Physics and Astronomy. The activities in this direction will leverage on already existing resources and capabilities from different research infrastructure and e-infrastructure organizations to maximize their use across the research community.

In the context of the project the Interoperability Work Package, WP6, has specific objectives that are guiding the activity of task T6.3 - Interoperability pilots (service implementation, integration, validation, provisioning for Science Demonstrators) :

- Providing the architecture, validated technical solutions and best practices for enabling interoperability across multiple federated e-infrastructures, overcoming current gaps expressed by user communities and resource providers.
- Validating the compliance of services provided by WP5 with specifications and requirements defined by the Science Demonstrators in WP4.
- Defining and setting up distributed Interoperability Pilots, involving multiple infrastructures, providers and scientific communities, with the purpose of validating the WP5 service portfolio.
- Assessing the maturity level of solutions in close cooperation with the Science Demonstrators, taking into account factors such as TRL, openness, scalability, user community adoption and sustainability.

The purpose of this document is to provide a first analysis of the requirements on interoperability testbed coming from different sources/stakeholders.

The document is organized as follows:

The document starts with an introduction including information about the EOSC, the EOSCpilot project position regarding the interoperability in the EOSC context, the recommendations coming from the European Commission communications on European Cloud Initiative (ECI) - Building a competitive data and knowledge economy in Europe[1] and New European Interoperability Framework (EIF) [2].

The section is completed with details regarding the work started on the first pilots on connecting computing centers (networking layer), distributed authorization and authentication (the collaboration with AARC project), data interoperability demonstrators and first integration tests of grid and cloud services through a CMS - HEP community complex workflow. The activities described were guided by:

- the gap analysis results presented in the Deliverable D6.1,
- the description of the framework that needs to be set up to allow the interoperability between the e-infrastructures and RI involved in the EOSC project contained in the deliverable D6.2 and
- the first report on Data Interoperability: Findability and Interoperability – D6.3.

The second section contains the analysis of the workflows of the use cases of the three sets of Science Demonstrators, even if we focused more on the first set of Science Demonstrators that have already

[1] http://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1472045853498&uri=CELEX:52016DC0178

[2] https://ec.europa.eu/isa2/eif_en

started their implementation activities. In this analysis we have highlighted their requirements and needs in terms of service, data and infrastructures interoperability.

The document closes with some conclusions and future work on how to improve the interoperability pilots, based on the feedback that will be collected in the following months to better address the Science Demonstrators use cases and the results of the data demonstrators.

1. INTEROPERABILITY IN THE EOSC CONTEXT

The idea of an European Open Science Cloud (EOSC)[3] took shape in 2015, as a vision of the European Commission of a large infrastructure to support and develop open science and open innovation in Europe and beyond. The EOSC is projected to become a reality by 2020 and will be Europe's virtual environment for all researchers to store, manage, analyse and re-use data for research, innovation and educational purposes.

In April 2016, the Commission presented its blueprint "European Cloud Initiative - Building a competitive data and knowledge economy in Europe"

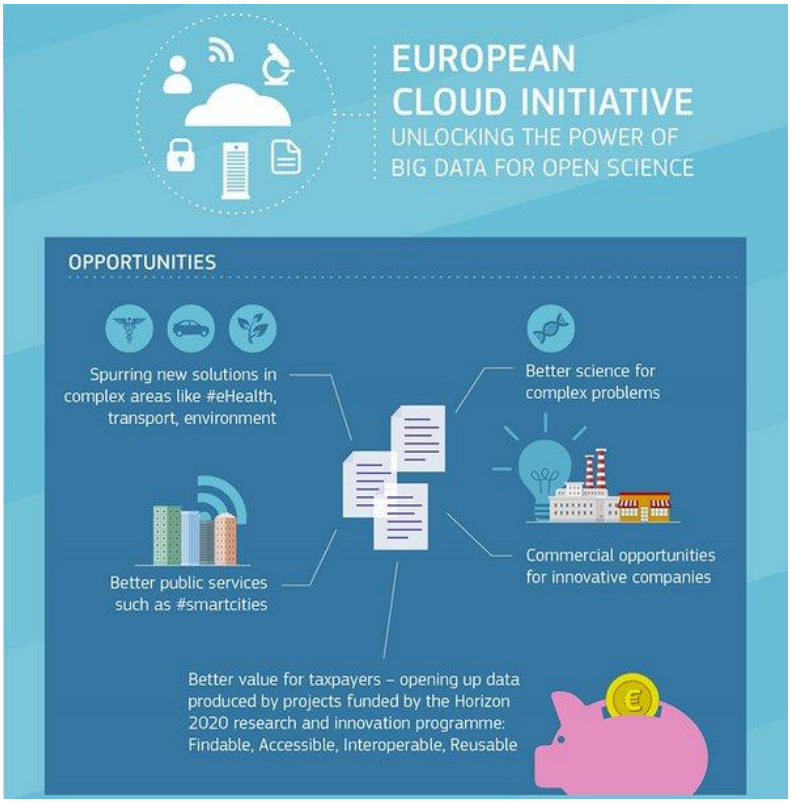


Figure 1: European Cloud Initiative

for cloud-based services and world-class data infrastructure to ensure science, business and public services reap benefits of big data revolution. As Europe is the largest producer of scientific knowledge in the world, it is well placed to take the global lead in the developing of a science cloud. In order to fully exploit the potential of data as a key driver of Open Science and the 4th industrial revolution [Office3] , several specific questions need to find their answer:

[3] https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud

- How to maximize the incentives for sharing data and to increase the capacity to exploit them?
- How to ensure that data can be used as widely as possible, across scientific disciplines and between the public and the private sector?
- How better to interconnect the existing and the new data infrastructures across Europe?
- How best to coordinate the support available to European data infrastructures as they move towards exascale computing?

It is mentioned in the blueprint that by bolstering and interconnecting existing research infrastructure, the Commission plans to create a new **European Open Science Cloud** that will offer 1.7 million European researchers and 70 million professionals in science and technology a virtual environment with open and seamless services for storage, management, analysis and re-use of research data, across borders and scientific disciplines by federating existing scientific data infrastructures, today scattered across disciplines and Member States. This will be underpinned by the **European Data Infrastructure**, deploying the high-bandwidth networks, large scale storage facilities and super-computer capacity necessary to effectively access and process large datasets stored in the cloud. It is well evidenced in the blueprint that integrated world-class HPC capability, high-speed connectivity and leading-edge data and software services, including, or building on, existing services by OpenAIRE, EUDAT, EGI, INDIGO - DataCloud, HelixNebula, PRACE and GÉANT, are needed by the European scientists and other lead users from industry (including SMEs) and the public sector.

We already see from the blueprint the first interoperability requirements – they are the answers to the questions above, and in the next sections we will show the progress of the work done in the direction mentioned in the blueprint, and which services with what characteristics were envisaged.

The ECI report underlines the fact that even if initially the focus is on the scientific community, "the user base will be expanded to the public sector and to industry, creating solutions and technologies that will benefit all areas of the economy and society." Taking this into account as well as other new EU initiatives, like the EU eGovernment Action Plan 2016-2020[4] , or new EU policies, such as the revised "Directive on the reuse of Public Sector Information"[5], the "INSPIRE Directive"[6], and the "eIDAS Regulation"[7] has led the EU Commission to adopt on March 2017 the new **European Interoperability Framework** (EIF) as part of its Communication (COM(2017)134)[8], giving specific guidance on how to set up interoperable digital public services. Although this framework is mainly aimed at European public administrations, as stated above it has the roots also in the EU Commission initiatives aimed at the scientific research world, and as such it was designed to be a generic framework – "a commonly agreed approach to the delivery of European public services in an interoperable manner. It defines basic interoperability guidelines in the form of **common principles, models and recommendations"**[9]. Using **the EIF** to steer European interoperability initiatives will **contribute to a coherent European interoperable environment, and facilitates the delivery of services that work together, within and across organisations or domains**. The new framework includes interoperability principles and models (Figure 2) to be implemented by an updated set of recommendations. It puts special emphasis on how these recommendations will apply in practice with the help of concrete existing solutions. New recommendations have a stronger focus on openness and information management, data portability, interoperability governance, and integrated service delivery.

[4] https://ec.europa.eu/digital-single-market/en/european-egovernment-action-plan-2016-2020

[5] https://ec.europa.eu/digital-single-market/en/european-legislation-reuse-public-sector-information

[6] http://inspire.ec.europa.eu/about-inspire/563

[7] https://ec.europa.eu/futurium/en/content/eidas-regulation-regulation-eu-ndeg9102014

[8] http://eur-lex.europa.eu/resource.html?uri=cellar:2c2f2554-0faf-11e7-8a35-01aa75ed71a1.0017.02/DOC_1&format=PDF

[9] http://ec.europa.eu/isa2/sites/isa/files/eif_brochure_final.pdf

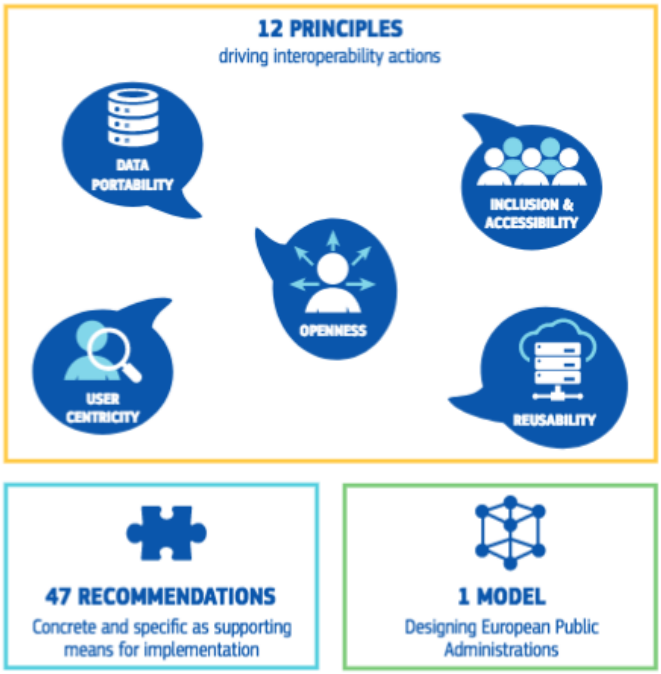


Figure 2: EIF Insights

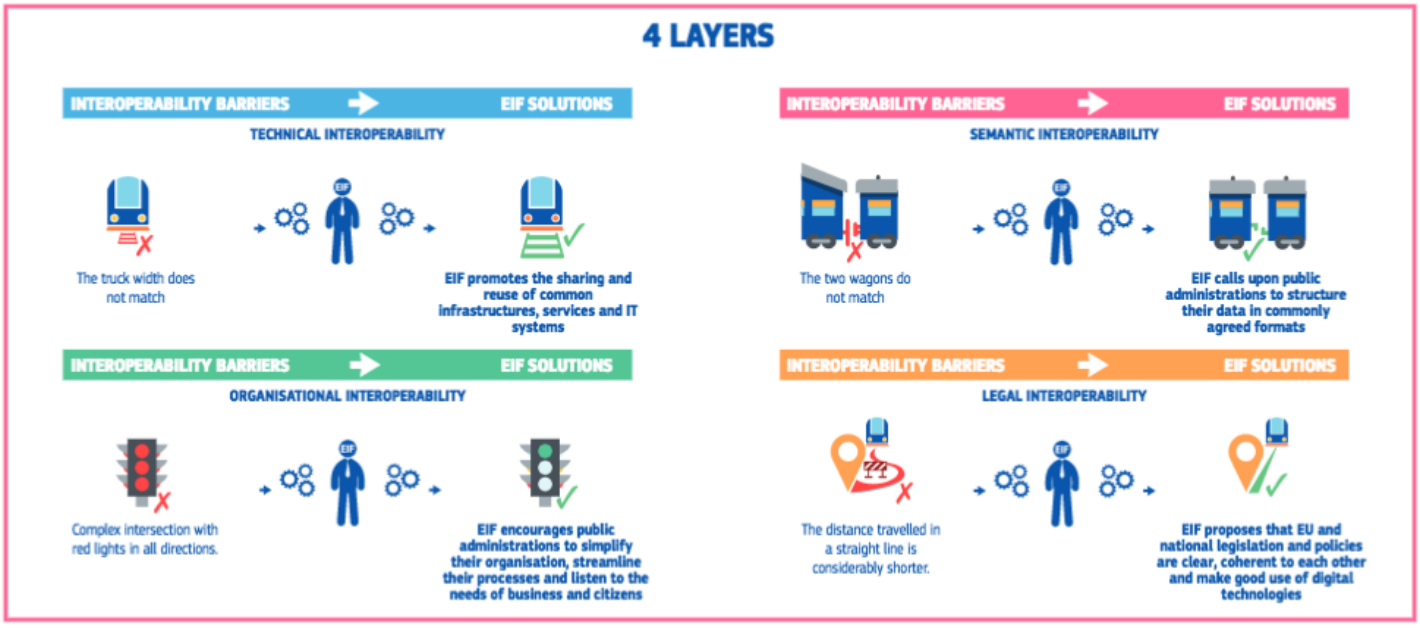


Figure 3: EIF layers of interoperability to be considered when implementing European public services

Between the interoperability principles and related recommendations that we could cite, of interest for the activities described in this report, from the point of view of technical interoperability and interoperability testbeds/demonstrators providing interoperable services, in particular from the technical and semantic interoperability layers (Figure 3) are:

- **Openness** (principle 2) - mainly relates to data, specifications and so ware.
  - ***Recommendation 2***: Publish the data you own as open data unless certain restrictions

apply
    - ***Recommendation 3***: Ensure a level playing field for open source software and demonstrate active and fair consideration of using open source software, taking into account the total cost of ownership of the solution.
    - ***Recommendation 4***: Give preference to open specifications, taking due account of the coverage of functional needs, maturity and market support and innovation
- **Reusability** (principle 4) - of IT solutions (e.g. software components, Application Programming Interfaces, standards), information and data, is an enabler of interoperability.
    - ***Recommendation 6:*** Reuse and share solutions, and cooperate in the development of joint solutions when implementing European public services.
    - ***Recommendation 7***: Reuse and share information and data when implementing European public services, unless certain privacy or confidentiality restrictions apply
- **Technological neutrality and data portability** (principle 5):
    - ***Recommendation 8***: Do not impose any technological solutions on citizens, businesses and other administrations that are technology-specific or disproportionate to their real needs
    - ***Recommendation 9:*** Ensure data portability, namely that data is easily transferable between systems and applications supporting the implementation and evolution of European public services without unjustified restrictions, if legally possible.
- **User centricity** (principle 6) - Users' needs should be considered when determining which public services should be provided and how they should be delivered.
    - ***Recommendation 12***: Put in place mechanisms to involve users in analysis, design, assessment and further development of European public services
- **Security and privacy** (principle 8)
    - ***Recommendation 15***: Define a common security and privacy framework and establish processes for public services to ensure secure and trustworthy data exchange between public administrations and in interactions with citizens and businesses.
- **Preservation of information** (principle 11) - records and information in electronic form […] must be preserved and be converted, where necessary, to new media
    - ***Recommendation 18:*** Formulate a long-term preservation policy for information related to European public services and especially for information that is exchanged across borders.
- **Assessment of effectiveness and efficiency** (principle 12) - Various technological solutions (e.g. cloud computing, Internet of Things, big data, and software-as-a-service) should be evaluated when striving to ensure the effectiveness and efficiency of a European public service
    - ***Recommendation 19***: Evaluate the effectiveness and efficiency of different interoperability solutions and technological options considering user needs, proportionality and balance between costs and benefits
- **Semantic Interoperability layer** - Semantic interoperability ensures that the precise format and meaning of exchanged data and information is preserved and understood throughout exchanges between parties
    - ***Recommendation 30***: Perceive data and information as a public asset that should be appropriately generated, collected, managed, shared, protected and preserved
    - ***Recommendation 31***: Put in place an information management strategy at the highest possible level to avoid fragmentation and duplication. Management of metadata, master data and reference data should be prioritised.
    - ***Recommendation 32***: Support the establishment of sector-specific and cross-sectoral communities that aim to create open information specifications and encourage relevant communities to share their results on national and European platforms.
- **Technical Interoperability layer** – covering applications and infrastructures linking systems and services. Aspects of technical interoperability include interface specifications, interconnection services, data integration services, data presentation and exchange, and secure communication protocols

- ***Recommendation 33***: Use open specifications, where available, to ensure technical interoperability when establishing European public services.

- **Conceptual model** for integrated public services provision – promoting the idea of **interoperability by design** and **reusability as a driver for interoperability**
  - ***Recommendation 36:*** Develop a shared infrastructure of reusable services and information sources that can be used by all public administrations
  - ***Recommendation 37:*** Make authoritative sources of information available to others while implementing access and control mechanisms to ensure security and privacy in accordance with the relevant legislation
  - ***Recommendation 38:*** Develop interfaces with base registries and authoritative sources of information, publish the semantic and technical means and documentation needed for others to connect and reuse available information.
  - ***Recommendation 41:*** Establish procedures and processes to integrate the opening of data in your common business processes, working routines, and in the development of new information systems.
  - ***Recommendation 42:*** Publish open data in machine-readable, non-proprietary formats. Ensure that open data is accompanied by high quality, machine-readable metadata in non-proprietary formats, including a description of their content, the way data is collected and its level of quality and the license terms under which it is made available. The use of common vocabularies for expressing metadata is recommended
  - ***Recommendation 44:*** Put in place catalogues of public services, public data, and interoperability solutions and use common models for describing them

According to the "Report on the governance and financial schemes for the European Open Science Cloud"[10] the European Open Science Cloud is envisioned by the European Commission as a supporting landscape to foster open science and open innovation: a network of organisations and infrastructures from various countries and communities that supports the open creation and dissemination of knowledge and scientific data.

The creation of EOSC is aimed at removing technical, policy and human barriers, leading to knowledge creation and economic prosperity in Europe. The European Commission's "European Cloud initiative" publication[11], issued in April 2016, set an ambitious vision for the European Open Science Cloud: "to give Europe a global lead in scientific data infrastructures and to ensure that European scientists reap the full benefits of data-driven science."

As mentioned in the "Realising the European Open Science Cloud"- the first report and recommendations of the Commission High Level Expert Group on the European Open Science Cloud, the EOSC is intended to set off the ground by federating existing scientific data infrastructures that are now spread across disciplines and EU member states. This will make access to scientific data easier and more efficient.

In this complex context the EOSCpilot project[12] is supporting the first phase in the development of the EOSC. The project brings together stakeholders from research infrastructures and e-Infrastructure providers and will engage with funders and policy makers to propose and trial EOSC's governance framework. It has has selected a number of Science Demonstrators[13] functioning as high-profile pilots that integrate services and infrastructures to show interoperability and its benefits in a number of scientific domains: Earth Sciences, High-Energy Physics, Social Sciences, Life Sciences, Physics and Astronomy.

EOSC services will need to be distributed - a decentralised System-of-Systems (SoS) based on collaboration and coordination, with components independently provided and managed by local, regional, national and international organizations. The EOSCpilot project will demonstrate the usage of existing digital

[10] https://ec.europa.eu/research/openscience/pdf/ospp_euro_open_science_cloud_report-.pdf

[11] http://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX:52016DC0178

[12] https://eoscpilot.eu/

[13] https://eoscpilot.eu/science-demonstrators

infrastructures in a combined manner, across disciplines and borders to realise the both the goals of the FAIR principles for sharing data, and equivalent principles to sharing software, methodologies and all aspects of the research lifecycle. It will determine the interoperability needed for an efficient use of IT services and equipment, and of our precious scientific data, which obstacles we are facing, and what solutions are already available.

In the framework of the EOSCpilot, WP6 aims to develop and demonstrate the interoperability requirements between e-Infrastructures, domain research infrastructures and other service providers needed in the European Open Science Cloud. It provides solutions, based on analysis of existing and planned assets and techniques, to the challenge of interoperability. Two aspects of interoperability will be considered: **Data interoperability**, ensuring that data can be discovered accessed and used according to the FAIR principles, and **Service interoperability,** ensuring that services operated within different infrastructures can interchange and interwork. For continued development of the EOSC, and future more advanced federated solutions, the services "internal layers" or the service portfolios need to be standardised allowing for Discoverable and Interoperable service components. High-level **organisational interoperability** leading to Interoperable providers is also needed (cf. Enterprise Architecture (EA)) to meet all expectations on interoperability; lock-in effects are avoided by allowing service transparency and service mobility.

The main activity of the task T6.3 - Interoperability pilots (service implementation, integration, validation, provisioning for Science Demonstrators) is to set up demonstrators to show interoperability among infrastructures and to foster the adoption of the solutions according to the FAIR principles, in close collaboration with the Services and Science Demonstrator WPs (WP4 & WP5).

The work of the Interoperability WP is aligned with the expected impact of the INFRADEV-4-2016 call, requiring the project to "facilitate access of researchers across all scientific disciplines to the broadest possible set of data and to other resources needed for data driven science to flourish". In order to run this infrastructure and to enable usability, a rich set of interoperable infrastructure services (IaaS) ranging from AAI, reliable storage endpoints, cloud management frameworks, SDN endpoints, to infrastructure monitoring, billing and accounting is required. This WP will gather infrastructures, operating computational facilities, and user communities from "the large" to "the long tail of science", avoiding unsustainable fragmentation.

1.1. E-Infrastructures Level Interoperability

At the e-infrastructure level the analysis of possible initial requirements on potential interoperability testbed, starts from the work done in Task 6.1, "e-infrastructure gap analysis & interoperability architecture", in order to:

- perform a gap analysis on technical and political barriers that prevent the interconnection of e-infrastructures and the application of the FAIR principles, reported in deliverable "D6.1: e-Infrastructure Gap Analysis"[14], and represented in Figures 4 & 5
- define the architectural design of the interoperation of various types of infrastructures, which could participate to the future EOSC, presented in deliverable "D6.2: EOSC Architecture Design and Validation Procedure"[15], containing also technical solutions and their validation methods.

[14] http://bit.ly/eoscpilot_D_6_1

[15] http://bit.ly/eoscpilot_D_6_2

Challenges:

- Connect diverse infrastructures, i.e. with different technologies, access policies, …
- Provide infrastructures to diverse communities, with different requirements, culture, expectations, etc.
- Having fast and reliable network connections between the e-infrastructures
- Making communities and colleagues aware of existing solutions
- Having experts able to support the interconnection of e-infrastructures

Figure 4: E-infrastructures Interoperability Gaps and Challenges

Bridges:

- Global Authentication and Authorization Infrastructure(s) (AAI)
- Working on a European level to ensure fast, reliable, and affordable network connections
- Provide easy-access services to open e-infrastructures to diverse communities and to be able to include diverse e-infrastructures
- Mutualise e-infrastructures across disciplines and technologies
- Bring communities together in terms of common vocabulary, global services, etc.
- Build human networks, ensuring education and training on relevant issues, sharing expertise across communities, borders, and e-infrastructures.

Figure 5: E-infrastructures Interoperability possible solutions

We summarize the main finding of the gap analysis and architecture design work in the form of a list of requirements, with information on possible solutions, services that could meet those requirements, ongoing or future activities, taking in consideration only the technical aspects relevant for this report:

- Authentication and Authorization Infrastructure (AAI)
  - Requirements
    - Common access policies – to be set up, to define a global efficient Authentication and Authorization Infrastructure that federates existing AAI
  - A joint effort between EOSCpilot and AARC project is undergoing to closely align the AARC blueprint architecture and the related policies with the recommendations for the underlying authentication and authorisation design within the EOSC – presented in Section 1.1.2 bellow
  - Potential solutions coming from different projects:
    - INDIGO IAM from INDIGO-DataCloud
    - EGI Checkin from EGI
    - eduTeams from eduGAIN
    - B2ACCESS from EUDAT
  - Potential integration between the above services, based on the standards they use, are described in Annex A3, “Description of AAI technical solutions deployed in the main e-

infrastructures", of deliverable D6.1

- Data
  - Requirements
    - harmonization of protocols, tools and solutions already used in the different infrastructures and interoperability in services that rely on these tools and solutions
  - Interoperability approach
    - Standardization of the interfaces to services.
    - Development of new open interfaces
  - Possible solutions coming from different projects and infrastructures, that could be tested and validated from the point of view of interoperability, as part of the Science Demonstrators pilots
    - iRODs, from CNRS
    - Onedata, FTS, dCache, from INDIGO
    - B2FIND, B2SHARE, B2SAFE & B2STAGE, from EUDAT
    - GridFTP, GPFS and gtransfer, used by PRACE
- Computing
  - Requirements
    - Interoperability of different infrastructure, computing resources (Grid, Cloud, HTC, HPC) and services they host, in order to to allow users to use all types of resources they need in their scientific workflows in a seamless manner.
  - Possible solutions coming from different projects and infrastructures, that could be tested and validated as part of the Science Demonstrators pilots
    - INDIGO-DataCloud solutions for cloud & container management, like PaaS Orchestrator, Infrastructure Manager, Mesos & Kubernetes Clusters technologies, udocker
- Network
  - Requirements
    - Stable and robust network services – already very well addressed by NRENs and GÉANT
    - High quality of the last mile connectivity (end user to EOSC e-Infrastructure)
    - Improved and guaranteed network capability
    - Enforced level of network security
- Core infrastructures:
  - Requirements
    - Data and Network Accounting
    - Guaranteed traceability for all actions and services
  - Possible solutions coming from different projects and infrastructures, that could be tested and validated as part of the Science Demonstrators pilots
    - EGI APEL – for accounting of usage of heterogeneous resources of different types (computing, data)
    - Work on the alignment of EOSC requirements on this area with the AARC2 project[16] within Policy and Best Practices Harmonization – Accounting and Data Protection
- Portal & user-friendly services

[16] https://aarc-project.eu/

1.1.1. Pilot for Connecting Computing Centres

In the context of EOSC, the infrastructure interoperability naturally requires high-speed communication among the data centres and between data centres and user sites. Since cloud services and infrastructures are distributed all over Europe, network connectivity between the providers and consumers of the cloud services is therefore a pillar on which EOSC has to be built. The stability, robustness and high performance of the network infrastructure are crucial elements for the EOSC to success, in order to provide the highest level of user satisfaction in accessing EOSC's services.

In this perspective, the objective of this pilot (named PiCo2) is to promote a possibility of data and codes flow along the edges of a network, between data centers (data storage and computation, Tier 1 and Tier 2), which are agnostic as far as a thematic application is concerned, i.e. which can serve a as large as possible diversity of scientific communities.

1.1.1.1 Sites involved

All sites involved are currently located in France. It is expected that, through WP6 and WP6.3, it evolves towards a European wide project. Current participants are (contacts are mentioned):

- French National centers: IDRIS, Orsay, CNRS (Denis Girou) and CC IN2P3, Lyon, CNRS (Eric Fède)
- French National Grid Initiative (France-Grille) and CEA (Geneviève Romier, Jérôme Pansanel & Jean-Pierre Meyer)
- Network: GÉANT (Xavier Jeannin)
- Mesocentres (Tiers 2) : MCIA, Université de Bordeaux (Pierre Gay) and GRICAD, CNRS & Université de Grenoble-Alpes, Grenoble (Violaine Louvet)
- INRA/INRIA, Bordeaux, for testbeds, data and programs (Alain Franc, Olivier Coulaud.

1.1.1.2 Organization of the work

A first choice has been to prioritize data flow to start pico2, and install a flow for codes between computing infrastructures as a further step. A first technical choice for data flow has been to connect data storage infrastructures which were already organized and available within a iRODS zone (see {https://irods.org/) and to build a federation between them.

Proof of concept and full scale test: A proof of concept of this possibility has been set up between a Tier 1 infrastructure (IDRIS) and two Tier 2 centers (MCIA, Mésocentre de Bordeaux and GRICAD, Mésocentre de Grenoble). MCIA has an iRODS server for connection between all laboratories which are member of the mesocentre, with about 300 To of data storage. The technical solution has been simple: to install an iRODS client of these zones on a post-treatment machine (ADA) at IDRIS. Provided a user can pull from IDRIS to iRODS zone, the security could be guaranteed quite easily. As a second step, the possibility to push on ADA has been provided too. A full scale test of data flow, making use of the MCIA iRODS zone, between IDRIS located at Orsay and PlaFrim (a development platform of size Tier 2) located at IMB/INRIA/LaBRI at Bordeaux has been done with success: several tens of data files, of about 10-40 Go each, have been comfortably transferred from ADA to PlaFrim, with high security.

Technical groups: Upon reporting of this full scale test between two sites, it has been decided to organize the project with technical groups, each devoted to a given task. Currently, two groups have been set up

- a technical group on federation of several existing iRODS zone, led by Oliver Henriot, GRICAD
- a technical group on the connection between sites, through GÉANT, led by Xavier Jeannin, RENATER.

a. Technical group for iRODS zones federation

Access to data across different HPC clusters is crucial to allow scientists to scale out the testing of hypotheses developed in the laboratory. Traditional transfer methods, such as physical supports exchange,

ftp or scp are often prone to shortcomings. They can be cumbersome, slow, insecure, complex to set up and operate or even a source of error. Seamless access to data from all computational platforms is therefore a desirable goal. We aim to leverage the federation capability offered by the iRODS open source data management system, present within many mesocenters and research institutions, in order to offer this capability. To do this, this group is set up in order to:

- Create a group of experts to coordinate and exchange on the the operations to setup federations between iRODS zones
- Carry out POC federations between pilot sites
- Assess the impact, performance and adoption of the set up federations
- Create a working manual to assist the setup of new site federations in order to generalize the solution

b. Technical group for connection between sites

The European NRENs and GÉANT offer the best possible connectivity between European academic data centres, enforcing an over provisioning capacity policy aimed to host R&E user data at their peak rate, whereas commercial Internet connectivity is far from being optimal in this context, being engineered based on different drivers (profitability, sustainability, etc.).

Beyond the standard IP connectivity, the GÉANT+NRENs network is able to provide additional advanced network services, like point-to-point L2 circuits and Virtual Private Network (VPN) that could improve data exchange performance. The HEP and HPC communities, for example, chose to use a L3VPN service (called LHCONE and respectively PRACE) which allows the fast transfer of terabytes of data between HEP and HPC data centre all over the world, on a daily basis. Commercial cloud providers also recommend this type of connection for their clients:

- https://aws.amazon.com/answers/networking/aws-multiple-region-multi-vpc-connectivity/
- https://cloud.google.com/interconnect/
- https://azure.microsoft.com/fr-fr/services/expressroute/

A VPN allows connection to data centres with the assurance that the site sending data is an authorized member of EOSC and thus this traffic could use a path with relaxed security policy, avoiding having to go through an expensive perimeter security device (e.g. a firewall system), which also tends to deteriorate the throughput speed of large data flows.

Network reliability requirement

EOSC is going to be a federated environment for scientific data sharing and computing services. The confidence that the user/application will have in this federated environment will be impacted by the availability and reliability of the services provided. This means that a reliable and resilient network underpinning these services is an absolute requirement.

EOSC infrastructure will be a complex infrastructure involving various elements (storage element, computing resource, network, …) contributing to deliver the service to the end users. This could lead to a situation where it will be difficult to identify the origin of problem. For instance, users encounter a bad experience as their applications are very slow. What is the origin of the problem? Latency increased in the WAN, User's LAN, EOSC data center LAN, application, user's PC, other cause?

The diagnostic process is crucial for a good user experience. Each element contributing to EOSC must be able to provide a relevant monitoring of this domain. In this context, a proper network monitoring system is necessary. PerfSONAR is a monitoring tool widely popular in the NREN community, and can contribute to identify easily the root of potential problems or performance issues. GÉANT is working on several solutions (PerfSONAR, SQM project, e.g.) that can be validated in EOSCPilot.

Initial requirement for computing centres connectivity

In the context of this "interoperability testbed", it will be a very heavy workload to test all the possible network scenarios for EOSC. The project aims to deploy a specific monitored connectivity service between Pico2 data centres, a layer 3 Virtual Private Network. This project aims to verify the following points:

- Are all type of sites able to be connected with this solution (Europe, regional network)?
- Monitoring solution
    - reliability and performance monitoring
- What is the impact on the Pico2 sites?
    - Organization, internal network configuration
    - Do data centres have the technical knowledge and culture for implementing this connectivity solution?
- L3VPN governance:
    - Organisation,
    - Connectivity between L3VPN and Internet
    - Security agreement
    - Deployment easiness

Another important aspect that is beyond the scope of Pico2 is the cost model of such solution.

1.1.2. Initial Requirements on Pilots for distributed authorization and authentication

The EOSCpilot and the AARC project [reference], have started a collaboration activity in the field of authorization and authentication, policies and recommendations regarding their design. In this section we present the first steps done by this collaboration and the plans for the setup of an interoperability pilot in this area[17].

The EOSCpilot is putting forward a governance framework for the EOSC and contributing to the development of European open science policy; supporting a demonstrators that integrate services and infrastructures to show interoperability and its benefits; and engage with a broad range of stakeholders, across borders and communities, to build the trust and skills required for adoption of an open approach to research. EOSCpilot recognises the importance of building on existing results that are already endorsed by research and e-Infrastructures, and adopting best practice in the delivery of open science.

The AARC project started in May 2015 with the aims of championing federated authentication and authorization to support research collaborations, libraries and e-infrastructures. For two years, AARC worked with research communities and e-Infrastructures to jointly design, test and promote technical and policy solutions to address well-known challenges that prevented the adoption of federated access among scientific collaborations. AARC has made significant progresses to this end by delivering a blueprint architecture [reference] and a set of related policies to support research collaborations to deploy federated access.

The EOSCpilot project management and the AARC project management are working together to closely align the AARC blueprint architecture and the related policies with the recommendations for the underlying authentication and authorisation design within the EOSC.

The EOSCpilot project has a number of activities which share a common interest with AARC, including:

- Making recommendations on the policy drivers for open science;
- Developing a framework for the service architecture of the EOSC, including core services; and
- Making recommendations for standards to interoperability between infrastructures and their component data and services.

[17] http://bit.ly/eoscpilot_aarc

AAI is recognised as a key issue for interoperability in the EOSC, and an interoperability pilot on this area will be developed.

AARC is working on a number of different aspects that relates to identity technology as well as on policies aspects that could well support EOSCpilot needs, namely:

- Account linking and evaluating combined assurance (e.g. to link ORCID id);
- Assurance aspects;
- Command line (as opposed to web-based/portal) access;
- Multi-factor authentication;
- Scalable authorisation to access data and relevant roles in communities;
- Best practices for operating components that are foreseen in the AARC blueprint architect;
- A number of different pilots with different research communities to support them to deploy the AARC blueprint architecture and the related set of policies.

The new funding for AARC (May 2017 - April 2019) includes support for cross e-infrastructures interoperability activities along with support for the establishment of AEGIS, a forum for e-Infrastructure operators to coordinate the deployment of interoperable federated AAIs based on the AARC results.

There is a mutual interest in agreeing on a collaboration to work on the following areas:

- A first joint workshop (to take place in the first months of 2018) to cover main principles of the AARC blueprint and key policy aspects as well as EOSCpilot goals and planned developments
- Adoption of the AARC Blueprint Architecture and its accompanying implementation guidelines and support policy frameworks in the EOSCpilot architecture
- Scope a pilot within the EOSCpilot as a interoperability demonstrator
- Participation of EOSCpilot as observer in AEGIS

In the context of EOSCpilot, Interoperability Pilots, partners already deployed some of the available AAI solutions, like INDIGO IAM and WaTTS for token translation, from INDIGO DataCloud project, B2ACCESS, from EUDAT, as part of the Science Demonstrators testbeds. These services are the first step towards the interoperability pilot to be built in the framework of the collaboration with AARC. The activity will continue in the coming months with the creation and definition of the AAI demonstrator, leveraging the AARC experience, following the example of its piloted solutions[18], in order to assess whether the solutions envisaged by EOSCpilot meet the functional and technical integration requirements of research communities and e-infrastructures.

1.1.3. Grid-Cloud interoperability demonstrator for HEP community

The Compact Muon Solenoid (CMS) is one of the two general purpose experiments at the Large Hadron Collider (LHC) at CERN in Geneva. CMS relies on the distributed computing capacities of the WLCG in order to process and analyze the collision data taken during LHC live time. While WLCG provides an unprecedented processing capability in scientific computing, the expectation for the next decade predicts an increase in the computing needs which will be difficult to support with standard Grid facilities at research institutions.

Solutions having the potential to provide additional (e.g donated, hired, temporary, etc.) computing capacity to the LHC experiments and hence to CMS are of extreme interest for the collaboration itself, even more if the technical solutions would allow for a seamlessly integration with existing experiment computing infrastructures and the exploitation of innovative Big Data analytics platforms

[18] https://aarc-project.eu/pilots/piloted-solutions/

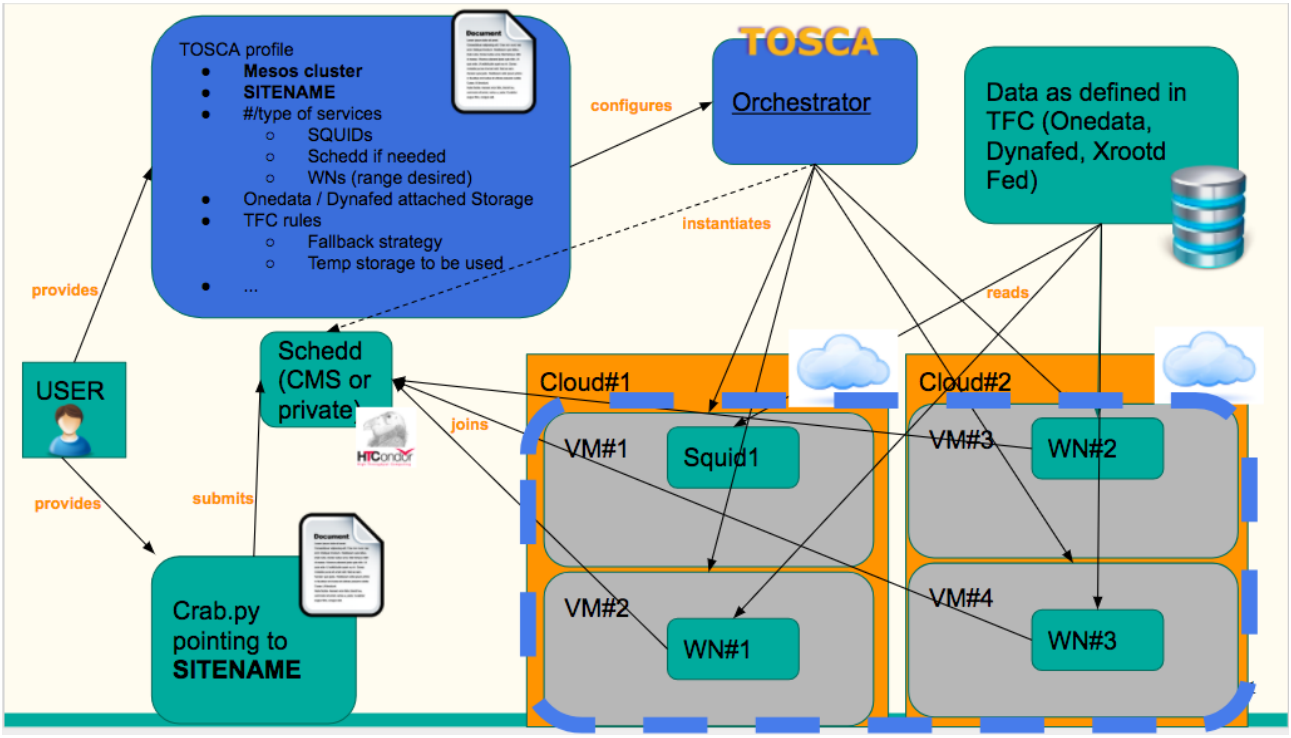


Figure 6: CMS use-case

Through the pilot described here it was deployed and tested software used by the large community of the CMS experiment at LHC on a private OpenStack cloud platform at INFN. The complex and intricate workflow, which includes data IO and data processing, has been used also to evaluate grid-cloud interoperability in the contest of CMS computing. We tested the data analysis access pattern which includes both reading data from grid storage elements and to write processed data back to the grid storages.

The whole runtime environment of CMS reconstruction has been containerized. Also the GRID Middleware dependencies are embedded at the level of Dockers.

HTCondor is the solution adopted by CMS to manage a worldwide Global Pool and thus we have containerized the HTCondor services such as the condor_master daemon. All the additional components, including squid proxy, security cache, are containerized as Dockers. CVMFS is mounted on Virtual Machine and is used to distribute the experiment software among grid and cloud stack. All the end user specific software libraries, configurations and any user-specific files, is shipped through HTCondor itself.

Finally, regarding data I/O Xrootd is used as protocol for grid-data ingestion into clouds, both for data reading and data writing.

The described flow has been enabled using technological products developed in the context of the INDIGO-DataCloud project, because of their natural match with most of the technical requirements derived from the above principles.

More in detail, INDIGO PaaS Orchestrator[19] together with the Infrastructure Manager[20] is used to guarantee not only a high coverage of supported platforms (such as OpenStack, OpenNebula, Microsoft Azure, Amazon etc.), but also to offer brokering capabilities between IaaS platforms, depending on SLA or, possibly, data locality etc.

TOSCA templates and Ansible roles are used to describe the overall cluster, including its applications and related interactions. The actual deployment of the TOSCA templates is then delegated to Apache Mesos clusters. Marathon is adopted as application framework to manage the above described docker application. Marathon offers two major advantages: the self-healing feature and the resources auto-scaling. Through this TOSCA based flow it is enabled also the propagation to downstream services of all the configuration parameters including tokens. The latter is a key point for the grid-cloud interoperability in the context of the CMS computing infrastructure.

Because of what just stated, the glue of the described stack is the INDIGO Identity Access Management (IAM), responsible for the authentication of the user both with PaaS layer and with IaaS providers and with third party services. A key features is provided by a IAM satellite component: the Token Translation Service[21] (TTS). TTS translates the incoming OpenID-Connect token into possibly several type of credentials, in this case it provides X.509 certificates. X509 is the standard authentication protocol used in the Grid environment, such mechanism, combined with the high level of abstractions described before facilitate, or better, make really possible the migration towards the cloud environment.

Tests were run in order not only to validate the setup but also to evaluate the performance comparing the results between grid and cloud solutions:

- physics validation: running a real analysis workflow - D-meson invariant mass reconstruction with Kpipi decays - results in Figure 7
- measuring CPU Time/Job Time with DODAS (Dynamic On-Demand Analysis Service) ephemeral Site performance VS oficial Grid Sites - results presents in Figure 8

The results show how the designed workflow, and solutions adopted perform at the same level as when using traditional grid services.

[19] https://www.gitbook.com/book/indigo-dc/orchestrator/details
[20] https://www.gitbook.com/book/indigo-dc/im/details
[21] https://indigo-dc.gitbooks.io/token-translation-service/content/

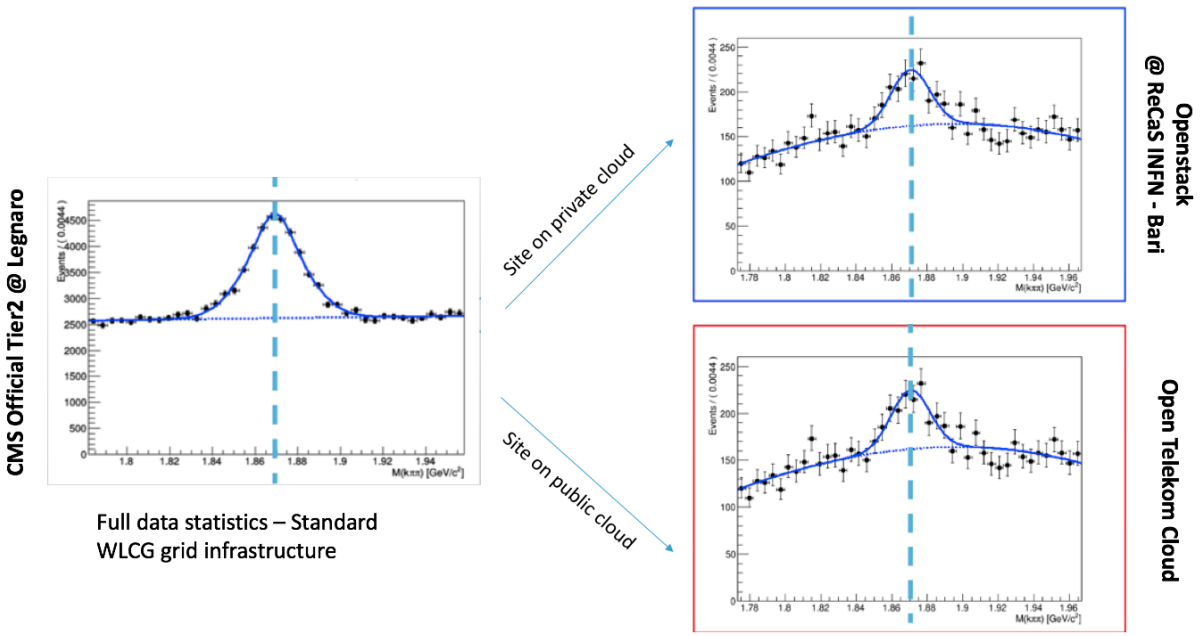


Figure 7: Validation tests results

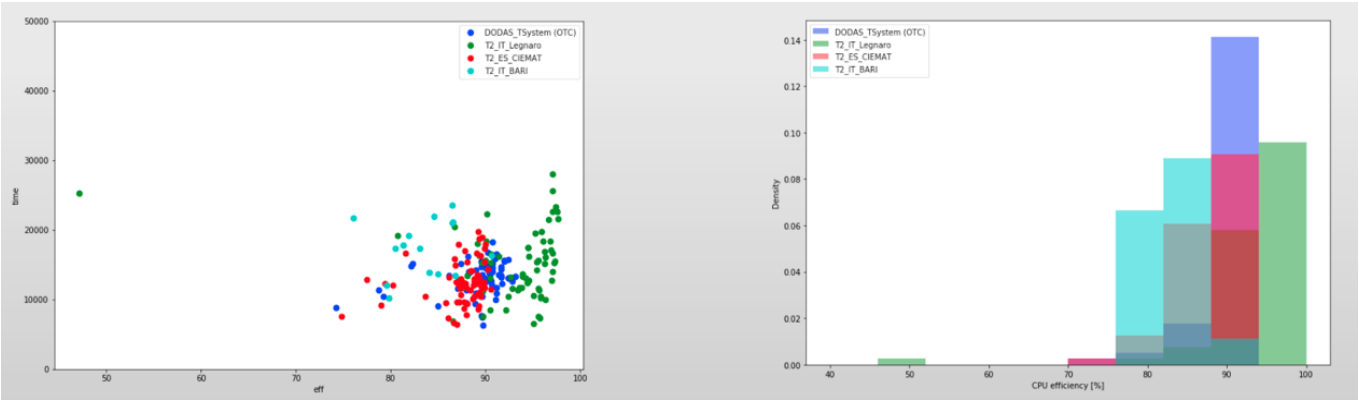


Figure 8: Performance Evaluation results

1.2. Data Level interoperability

The EOSCpilot project aims to improve interoperability between European data infrastructures, enabling efficient means to discover, access and share data in a sustainable manner, which is also easy to implement. As mentioned in the introductory part the interoperability in the EOSCpilot is mapped into two tracks: the research and data interoperability and infrastructure interoperability, reflected in the WorkPackage 6 structure: Task 6.1 - e-infrastructure gap analysis & interoperability architecture; Task 6.2 EOSC Research and Data interoperability and Task 6.3: Interoperability pilots (service implementation, integration, validation, provisioning for Science Demonstrators).

The main objective of the EOSCpilot data interoperability task (task 6.2) is to establish principles and develop mechanisms that enable the EOSC to provide research and data interoperability across the diversity of existing (and potential future) research communities, RIs, and other research assets.

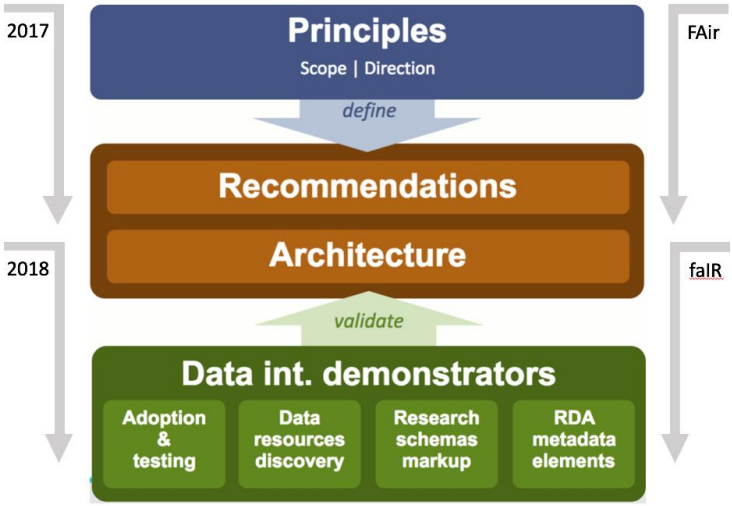


Figure 9: Research & Data Interoperability Work plan

The first deliverable regarding Data Interoperability, the D6.3 – "1st Report on Data Interoperability: Findability and Interoperability"[22] reports on the development of a set of guidelines to facilitate findability and reuse through a simple set of metadata properties that can be easily adopted, and that are applicable across scientific domains. It describes the **process** used in the identification of these properties, the **properties** themselves, and analyse how they relate to those properties proposed for use by other, comparable, international efforts. It also describes the **means** by which these guidelines can be adopted, and list the **demonstrators** that are under consideration to validate them. As mentioned in the report in order to **"balance the impact and effort of the data-interoperability group, it was decided to define the data interoperability goal focusing on one and probably the most important requirement described by the EOSCpilot project: demonstrate how to facilitate the availability of scientific data in EOSC.**

In order to better understand the data-interoperability demonstrators, described later in this section, we provide a short summary of the **guiding principles** of the task activity, the **process** and **strategy** envisaged in order to define dynamic guidelines and a data interoperability architecture proposal that will ease the efficient finding, access and use of public, scientific data. The dynamic aspect refers to the evolution of both guidelines and architecture as a consequence of the feedback received after their testing and validation by the data providers and consumers.

There were identified 13 principles grouped in 3 categories:

- Reuse: Leverage the rich legacy of RIs
  - rely on metadata from RI metadata catalogues
  - support an ecosystem of catalogues and metadata flow
  - provide quality recommendations to feedback to RIs
  - Making data FAIR is the responsibility of the RIs and their data repositories
- Least: Minimal metadata for maximal benefit
  - Findability first
  - Common and minimum metadata
  - Focus on common data types: datasets and data repositories
  - Flexible metadata models to embrace domain specifics
  - Service requirements and operational metadata as first-class citizens
- Practical: Sustainable and pragmatic delivery
  - Engage existing data repositories from EOSC science demonstrators
  - Reuse methods to expose dataset metadata through metadata catalogues
  - Simple to implement, easy to sustain

[22] http://bit.ly/eoscpilot_D_6_3

- Deliver guidelines and demonstrators

Phases of the guidelines and a data interoperability architecture proposal definition process:

- Review – existing metadata catalogues
- Find use cases
- Map important properties – of use case requirements to metadata models and standards
- Provide guidelines – for minimal properties, suitable data models and technologies
- Evaluate – technologies following proposed guidelines
- Demonstrate – the proposed strategy and recommendations, collect feedback

To drive and support the guidelines and the strategy proposal, partners and external stakeholders were engaged directly through workshops and surveys:

- BlueBRIDGE workshop: "FAIR friendly research data catalogues: How far are we? - April 3, 2017. 9th RDA Plenary Meeting. Barcelona, Spain
- How FAIR Friendly is your data catalogue? Exposing FAIR data in EOSC - September 8, 2017. Open Science Fair 2017. Athens, Greece
- EOSCpilot data interoperability technical workshop: Data catalogues and datasets in the European Open Science Cloud - 4-5 October, 2017. Genome Campus, Hinxton, UK

During these events a set of 39 collective recommendations, agreed by the majority of participants, were identified. For extensive details please see the D6.3. For the purpose of this deliverable, requirements on interoperability pilots addressing various demonstrators/use cases, there were identified at least 3 recommendations, mainly during the data interoperability F2F technical workshop, were there was also a dedicated topic/session regarding the "Requirements from e-infrastructure services":

- **E5.11 - Steps within the demonstrator**
  - Registration of EOSCpilot dataset metadata in a domain specific data catalogue.
  - Metadata sharing with other data catalogues
  - Use of metadata by a service to find and access the data
- **E5.12 - Indexing datasets from data catalogues**
  - use EUDAT-B2FIND as a test case to access and index metadata, focusing on minimum metadata provided by OAI-PMH interfaces and schema.org markup
- **E5.13 - Registries of data resources**
  - Use FAIRsharing for the discovery of catalogues and resources compliant with EOSCpilot data interoperability recommendations, of standards and data policies adopted by data resources

Working in conjunction with stakeholder data resource owners and catalogue providers, there were identified a number of **scientific demonstrators** for the testing and validation of the EDMI metadata, as well as of the comprehensiveness and adoptability of the guidelines generated by the work of the data-interoperability task. The feedback and learnings from these demonstrators will be used to make the necessary changes and improvements in order to have a practical data interoperability strategy by the end of the EOSCpilot project.

1.2.1. Findability and accessibility of datasets via operational metadata

- Rationale
  - Challenges for current services in finding the right operational metadata that can help them to manage data:
    - many services in science rely on data maintained in third party data resources
    - not easy to find, access, transfer and keep updated copies of the data
    - data resources employ different data models, a diversity of interfaces
    - highly distributed nature of the data
  - EOSCpilot project aims to provide metadata recommendations and a strategy to make data from third party resources more findable and accessible for services.

- This demonstrator will test how the proposed functional and operational metadata, EDMI, will help services to find, access, transfer and replicate data available in third party data resources

- Objectives
  - Involve at least two **data repositories** to adopt the recommendations of this project to expose dataset functional and operational metadata.
  - Involve one **catalogue of datasets** to index and expose minimum functional and operational metadata from at least one data repository
  - Involve at least **one service** to test the benefits of using the metadata proposed in this project.
- Participants
  - The PRIDE database, which hosts Proteomics datasets
  - The OMICsDI catalogue of datasets, which hosts metadata about omics datasets
  - The EUDAT-B2Find metadata catalogue, which indexes the metadata of scientific records

1.2.2. Discovery of compliant data resources and data catalogues

- Rationale
  - Challenges – even if catalogues of datasets offer an easier overview of what data is available from data resources, and where it can be found, it is more difficult to ascertain which data resources have been indexed by a particular metadata catalogue and which resources are compliant with proposed recommendations
  - This demonstrator:
    - will provide a better overview of existing catalogues and data resources indexed by these catalogues.
    - will help to recognise which catalogues and data resources comply with the recommendations of this project.
- Objectives
  - Involve at least an **existing catalogue** indexing data resources to help users and services to find dataset catalogues and data resources compliant with the project recommendations
    - **Create** a collection of data resources per dataset catalogue
    - **Associate** the EOSCpilot recommendations to those catalogues and data resources complaint with the recommendations
  - Involve a **catalogue of datasets** to register in a catalogue of data resources the list of data resources indexed and their compliance with the recommendations
- Participants
  - The FAIRsharing catalogue of databases, repositories, standards and data policies
  - The OMICsDI catalogue of datasets, which hosts metadata about omics datasets

1.2.3. Research schemas for exposing dataset metadata

- Rationale
  - Challenges
    - metadata catalogues do not provide a programmatic interface to ease the findability and accessibility of data
    - metadata catalogues do not provide the minimum properties contained in the proposed EDMI recommendation
    - there is a need to need to provide a simple and quick way to implement a solution which allows metadata catalogues to expose this structured metadata
  - Schema.org [reference] provides a simple mechanism to expose structured metadata using the existing web interfaces of metadata catalogues and data resources
  - This demonstrator:

        - Will explore the use Schema.org in a manner akin to that used by Bioschemas to facilitate exposing minimum scientific metadata, naming it “Research Schemas”
- Objectives
    - Community
        - **Start and support** the Research Schemas community effort
        - **Organise the first community meeting** to engage the community and plan future activities
    - Technical
        - **Use** Research Schemas as a vehicle to expose the minimum metadata properties proposed in the recommendations.
        - **Recycle** Bioschemas ideas to come up with a prototype of how to expose dataset metadata based on the recommendations.
        - Start to **define profiles** (metadata specification on top of existing schema.org types) based on the recommendations for scientific dataset and data catalogue
        - Come up with several examples to facilitate adoption
        - Test Research Schemas with one data resource to expose metadata
        - Test with **one catalogue of datasets** to expose metadata
        - Test with **one catalogue of datasets** to index schema.org metadata from a data resource.
- Participants
    - The PRIDE database, which hosts Proteomics datasets
    - The OMICsDI catalogue of datasets, which hosts metadata about omics datasets
    - The EUDAT-B2Find metadata catalogue, which indexes the metadata of scientific records
    - The Bioschemas.org community

1.2.4. Description and guidelines per metadata property

- Rationale
    - Work is going on in the RDA Metadata Interest Group (MIG) [reference] to provide detailed descriptions and recommendations for dataset metadata properties
    - EOSCpilot wants to contribute to and reuse these descriptions and recommendations, with particular focus on those properties identified as part of the recommended set of minimum properties (EDMI).
    - This demonstrator – intends to ensure the collaboration with the RDA MIG group to describe dataset metadata properties
- Objectives
    - Select a set of **dataset properties of interest** to start with (e.g. identifiers)
    - **Propose new properties** if any EDMI property is missing in the RDA MIG dataset proposal
    - **Propose structure and template** for properties to capture and harmonise feedback from the community.
    - **Contribute** to the definition of the property, linking with existing guidelines (especially domain specific) and **summarise recommendations**
- Participants
    - RDA MIG group
    - Identifier.org
    - Nick Juty (ELIXIR/CORBEL UMAN)

For the demonstrators presented in this section the preparation activity just started at the end of the year, after their identification after the discussions that took place during the meeting organized by the Data Interoperability team, T6.2, and their definition in the D6.3 deliverable. The objectives presented above for each demonstrator constitute the requirements for setting up the pilots. The result of the period was the

identification of the participants to the pilots committed to fulfill those requirements.

It will follow now the period of the setup, through group meetings between the interested participants. Face-to-Face, hackathons, will be organized in order to better define the characteristics of the pilots to be setup and of tests to be undergone. First results and status of the pilots will be reported in the updated report , D6.4 - "Interim Interoperability Testbed report"

2. SCIENCE DEMONSTRATORS & INTEROPERABILITY ASPECTS

One of the objectives of the EOSCpilot project is to understand and address the barriers that stop European research from fully tapping into the potential of data. To improve interoperability between data infrastructures, the project has to engage different scientific and economic domains, countries and governance models and will demonstrate how resources can be shared even when they are very large and complex. The aim of the **EOSCpilot Science Demonstrators** is to show the relevance and usefulness of the **EOSC Services** and their enabling of **data reuse**, to drive the EOSC development.

2.1. First set of Science Demonstrators

EOSCpilot started with five Science Demonstrators pre-selected from a call in 2016, with project execution from January to December 2017:

- **Environmental & Earth Sciences** - ENVRI Radiative Forcing Integration to enable comparable data access across multiple research communities by working on data integration and harmonised access
- **High Energy Physics** - WLCG: large-scale, long-term data preservation and re-use of physics data through the deployment of HEP data in the EOSC open to other research communities
- **Social Sciences** – TEXTCROWD: Collaborative semantic enrichment of text-based datasets by developing new software to enable a semantic enrichment of text sources and make it available on the EOSC.
- **Life Sciences** - Pan-Cancer Analyses & Cloud Computing within the EOSC to accelerate genomic analysis on the EOSC and reuse solutions in other areas (e.g. for cardiovascular & neuro-degenerative diseases)
- **Physics** - The photon-neutron community to improve the community's computing facilities by creating a virtual platform for all users (e.g., for users with no storage facilities at their home institutes)

2.1.1. ENVRI Radiative Forcing Integration

The scientific area of this Science Demonstrator lies in the Environmental & Earth Sciences - Environment, Greenhouse Gases and Climate Change and its scientific focus are:

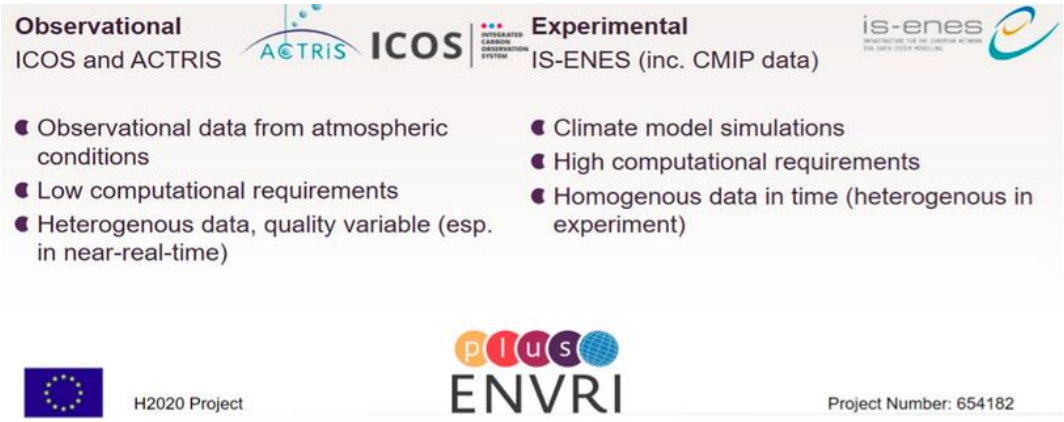
- to demonstrate dynamics of greenhouse gases, aerosols and clouds and their role in radiative forcing
- the interoperability between observations and climate modeling
- the cooperation between environmental research infrastructures

From a technical perspective, the pilot aims to build a data integration and metadata integration prototype to allow:

1. Climate data model users to get relevant (climatological or specific-time) observations, and
2. Data users to access relevant climate data sets and support scientists in their analysis.

The core objectives of this pilot is to build a **data integration demonstrator** and a **metadata integration demonstrator** which can be used to uniformly address and access IS-ENES (**I**nfra**S**tructure for the **E**uropean

Network for **E**arth **S**ystem Modelling) and ICOS (**I**ntegrated **C**arbon **O**bservation **S**ystem) & ACTRIS (**A**erosols, **C**louds, and **T**race gases **R**esearch **I**nfrastructure) data - this prototype is the base layer of the analysis demonstrator showing the usage of ENES and ICOS data together (Figure 10).



Services, technical and interoperability requirements to support this SD pilot are:

- for the **data integration framework,** that aims to make (initially parts of) multi-petabyte climate model data archives hosted at DKRZ and IPSL accessible for EGI/ICOS based on a common cloud services:
    - The Onedata[23] solution, developed in the context of the INDIGO-DataCloud H2020 project, to provide a transparent access to ICOS and IS-ENES datasets.
    - A reliable and scalable cloud Infrastructure where the IS-ENES data download and synchronization software (synda[24]) can be installed.
- for the **metadata integration framework,** that aims to make data searchable based on common APIs/middleware services or also external metadata catalogues:
    - A translator component able to query the IS=ENES metadata search API and to export the metadata for the (initially) agreed data subset to Onedata
    - To export data based on IS-ENES metadata search API[25]
- technical requirements:
    - On-demand allocation of VMs based on common Linux distribution (Debian/Ubuntu, RedHat) with at least 2 vCPU cores, 8 GB of RAM and 40GB of disk storage
    - A running installation of the Onedata framework to provide transparent access to datasets.
- interoperability:
    - An agreement between the IS-ENES and ICOS on the target metadata profile to be used in the metadata integration framework

In Figure 9 we can see the simplified data workflow in ICOS and the **required service** to operate the system, mainly from the **EUDAT** CDI (**C**ollaborative **D**ata **I**nfrastructure) services suite[26] like: B2FIND (for data discovery and exploration), B2SAFE (for long-term data persistency), B2STAGE (for data ingestion and staging)

[23] http://ec.europa.eu/isa2/sites/isa/files/eif_brochure_final.pdf
[24] https://github.com/Prodiguer/synda
[25] https://esgf-data.dkrz.de/search/esgf-dkrz/
[26] https://eudat.eu/catalogue

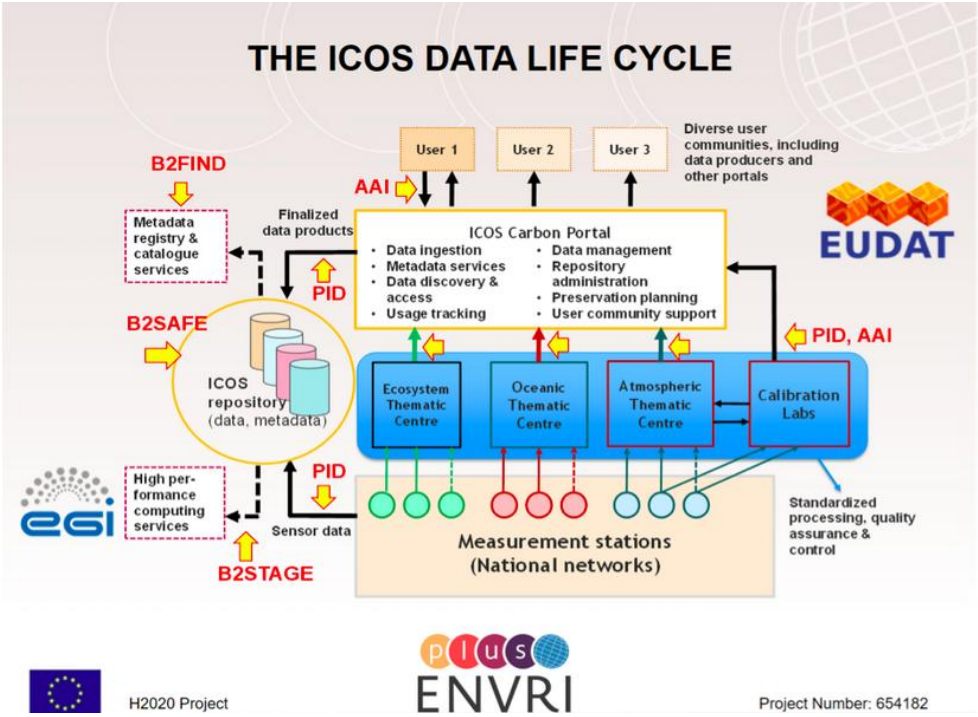


Figure 10: Simplified ICOS data workflow & required services

In order to make the multi-petabyte of **climate data hosted at DKRZ**[27]**,** offering also the **B2FIND** and **B2ACCESS** services**, and IPSL**[28] accessible from ICOS, the existing **Onedata** installation hosted at **CYFRONET** has been used to implement a **generic data export/synchronization pipeline**. This virtual volume mounted in one of the VM running in the **EGI Federated Cloud** infrastructure will be used to download and synchronize IS-ENES datasets with the synda library[29].

For the metadata integration, which is one of the main interoperability issue raised by this Science Demonstrator, IS-ENES will further develop the translator component to query the IS-ENES metadata search API and export them in Onedata.

2.1.2. Pan-Cancer Analysis in the EOSC

The Pan-Cancer Analysis of Whole Genomes project (PCAWG) is analysing large cohorts of cancer genomes, and pursuing so-called pan-cancer studies to identify factors that may be involved in tumor formation and disease progression across multiple cancer types. PCAWG is currently analyzing >2800 cancer whole genomes, largely on academic and public clouds, and is also developing approaches for data integration with transcriptome & clinical data to address specific hypotheses (Figure 11).

[27] https://www.dkrz.de/?set_language=en&cl=en

[28] https://www.ipsl.fr/en

[29] https://github.com/Prodiguer/synda

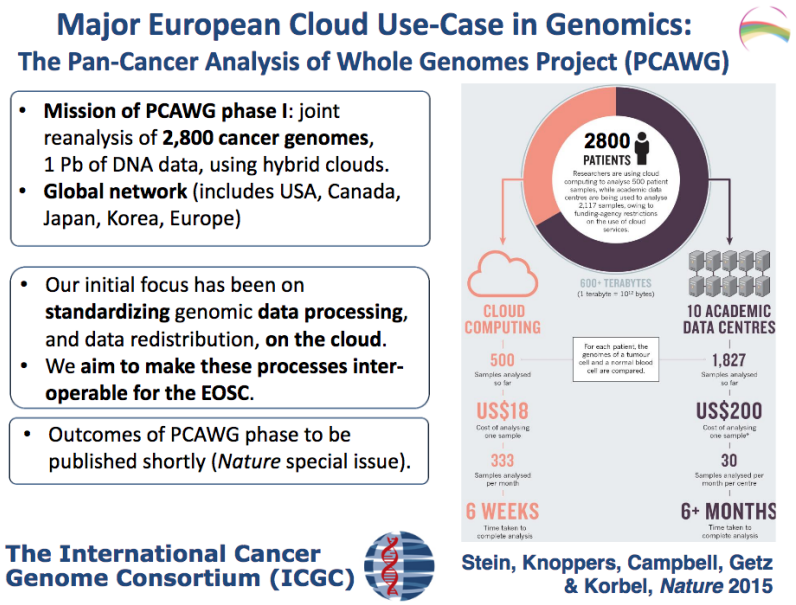


Figure 11: PCAWG Overview

The PCAWG project, in order to overcome the challenges of orchestrating analyses of thousands of human genomes on the cloud, has developed a computational framework, Butler[30], able to operate both on public and academic clouds. This highly flexible framework facilitates management of virtual cloud infrastructure, software configuration, genomics workflow development, and provides unique capabilities in workflow execution management.

Butler aims to be a comprehensive toolkit for analysing scientific data on clouds. To achieve this goal it provides functionality in four broad areas:

- Provisioning: Creation and teardown of clusters of Virtual Machines on various clouds.
- Configuration Management: Installation and configuration of software on Virtual Machines.
- Workflow Management: Definition and execution of distributed scientific workflows at scale.
- Operations Management: A set of tools for maintaining operational control of the virtualized environment as it performs work.

One can use Butler to create and execute workflows of arbitrary complexity using Python, or you can quickly wrap and execute tools that ship as Docker containers, or are described with the Common Workflow Language[31] (CWL).

In the context of the EOSCpilot project the aim of the SD is to configure Butler for large - scale cloud - agnostic deployment and interoperability:

- Deliver **standardized pipelines** for processing patient genomes, distributed via **Docker & Common Workflow Language**.
- Employ PCAWG interoperable frameworks to process cancer whole genomes from **20 most common cancer types**
- Additionally integrate cancer exomes, gene expression and epigenetic data to facilitate discoveries of alterations affecting genes.
- **Establish a portable cloud-based federated solution for collaborative cancer genomics and associated health data management, and an environment accessible to European scientists for analysis**

[30] https://github.com/llevar/butler
[31] http://www.commonwl.org/

Currently the SD framework is deployed on EMBL-EBI Embassy Cloud (Figure 12), using 1000 compute cores; 1 PB of NFS storage, and it has been deployed on multiple EOSCpilot-provided resources such as Cyfronet and ComputeCanada. In the latter, Butler has been used to process 2 datasets of 50 TBs..

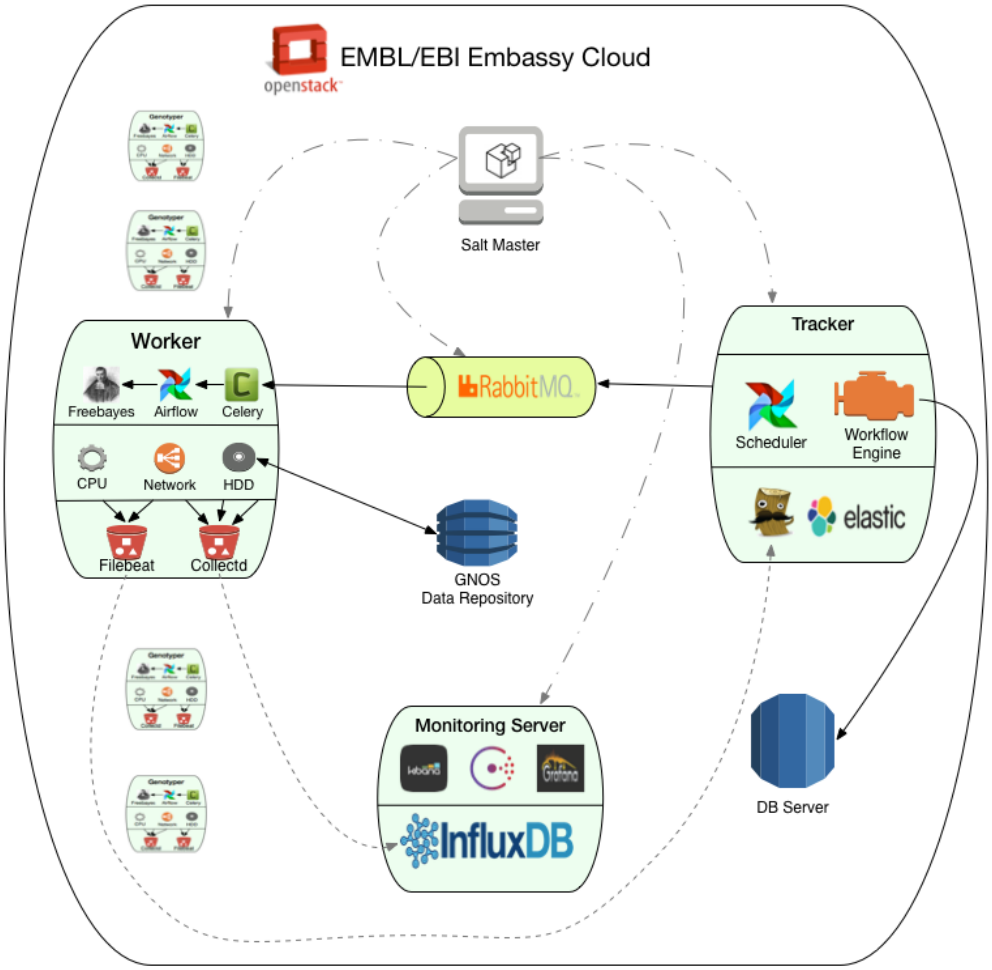


Figure 12: Butler deployment on EMBL-EBI Embassy Cloud resources

Services, technical and **interoperability requirements** to support this SD pilot are:

- Scalable IAAS cloud, offering:
  - Uniform APIs independently from the cloud provider erogating the service
  - Virtual machines with at least 16/32 vCPUs, with 1.875 GiB, 3.75 GiB and 7.5 GiB per vCPU
  - Ability to reliably sustain the creation of at least 200 VMs
    - More than 1.5K vCPUs
    - More than 1TB local SSD disk (or comparable performance, ~15000 IOPS) per VM
  - High bandwidth East-West connectivity, at least 10 Gb/s
  - High bandwidth North-South connectivity, 10Gb/s
  - Encryption at rest would be beneficial
  - Replicated Object storage. S3-compliance beneficial
  - Beneficial, but not strictly required at this stage:
    - Autoscaling
    - Spot-market like pricing
  - Out-of-the-box NFS-like clustered storage - Petabytes(s) scale

    - Able to reach 5MB/s per TB of allocated storage
    - Able to cope with >200 clients concurrently, retaining performance
    - Object storage
        - Petabyte(s) scale; More detailed requirements will follow as the Demonstrator progresses
    - Transparent data access
        - Petabyte scale
        - Able to securely transfer data across different datacenters
        - Support on-demand caching
        - POSIX compliant
        - ACL-style permissions
        - Able to reach 5MB/s per TB of allocated storage
- Technical services:
    - Data transfer mechanisms able to operate at the PB scale between data repositories and cloud providers. The ideal workflow would be as follows:
        1. Data gets pulled from central repositories and splitted among several cloud providers
        2. Cloud providers offers a system to enforce indexing, and access authorisation to the cached data
        3. User performs compute, with final dataset being much smaller than input data
        4. Final dataset is pulled back to central repositories for long term storage
    - Long term storage of data & software with bit preservation, to allow reproducibility of analyses in the long term (according to the FAIR principles)
    - Centrally hosted solution to manage data indexing, discoverability, authorization and access. Ideally, the central system should be accessible from each cloud provider infrastructure to propagate changes / update central replica catalog.
- Interoperability requirements:
    - Uniform APIs across different providers for all the exposed services
    - Compliance of services with the GA4GH APIs reference implementation (http://genomicsandhealth.org/)
    - Encryption of data at rest / in transit
    - Authorization mechanisms to allow user to access/release data

This demonstrator couples together two different but intrinsically inseparable use-cases:

- Data management lifecycle: data ingestion, access control setup, results fetching
- Analysis management lifecycle: analysis configuration, resources deployment (run and manage VMs/containers)

Cloud services are expected to support these in an as much as possible integrated way.

Massive computing projects as PanCancer will increasingly require to process huge datasets (>1PB) to fulfill their objective. As it is well know that moving data at this scale is highly inefficient, this use-case is advocating for a **shift in the way compute is performed** in the context of cloud: **compute should be brought to data, not the other way around**. Ideally, cloud providers will keep a copy of the datasets (or a subset of) near to their compute infrastructure, ready to be accessed by the workloads deployed by researchers. A **central EOSC catalog** (replica catalog) will allow researchers to discover in which provider a dataset is cached and request permission to access it. The authorisation - if granted - will be propagated back to the local storage system to allow access from the local compute resources.

After the deployment on the EMBL-EBI Embassy Cloud resources and successful integrated and tested

support for running

- Docker containers within Butler workflows, and
- tools described with the Common Workflow Language

within Butler, other deployment tests were performed onto Amazon AWS and Microsoft Azure commercial clouds, and additional resources have been made available by ComputeCanada and Cyfronet for project use. Deployments on EOSCpilot resources were successfully carried out in Q3 of the Science Demonstrator processing up to 50 TBs datasets.. Additional tests are planned and intensive tests are done on using the Onedata data management system into the framework.

2.1.3. Research with Photons & Neutrons

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 on Figure 13). The crystfel framework is increasingly used at various synchrotrons and FELs to analyze data 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.

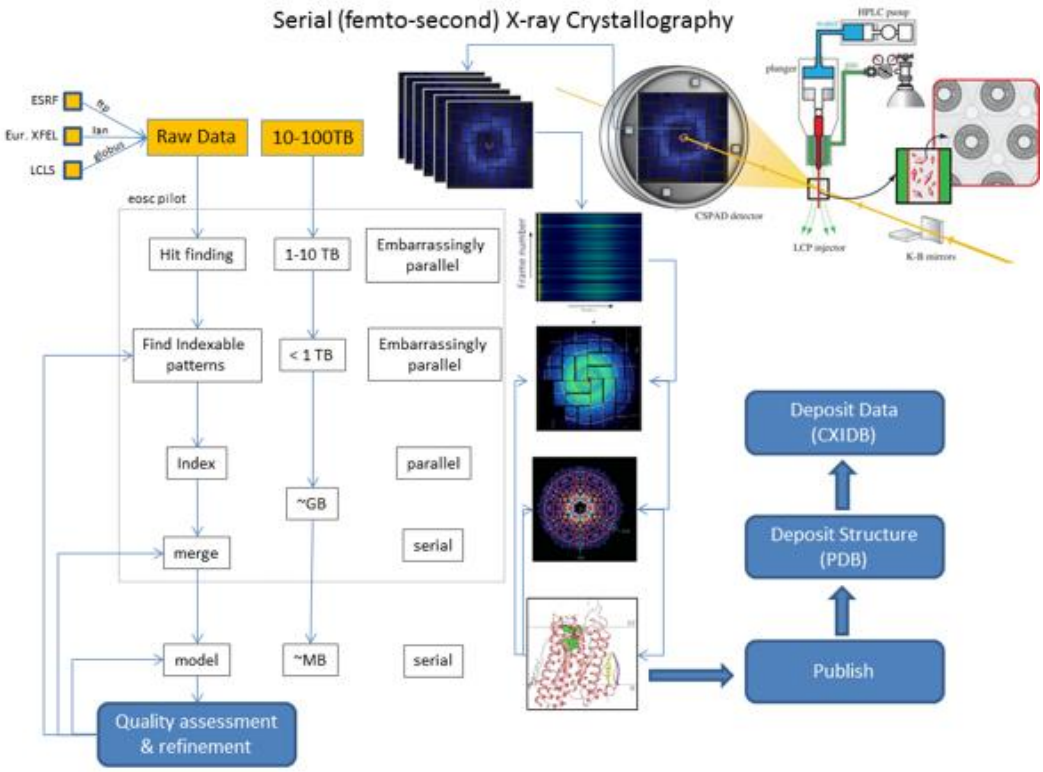


Figure 13: Photons & Neutrons Data Analysis Workflow

The SD deployed and tested software used by a large community in Structural Biology at Free Electron Lasers and Synchrotrons on a local OpenStack cloud platform and on local HPC clusters at DESY, Hamburg. The computationally challenging workflow was examined to identify and establish community specific cloud services and gain insight into technical, organizational, legal issues and interoperability requirements. We

tested two use cases:

- OnDA (online data analysis) is a modular and scalable utility designed for fast online feedback during serial X-ray diffraction and scattering experiments, based on open-source libraries and protocols (Figure 14).

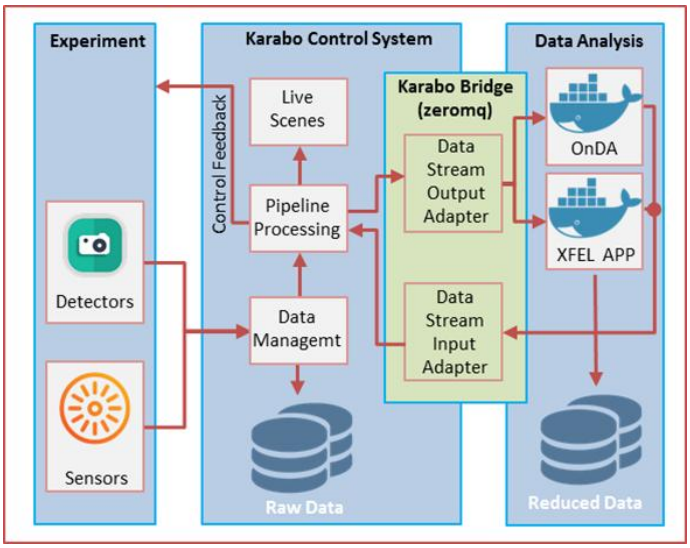


Figure 14: OnDA workflow coming from an Eur.XFEL application framework

- The Crystfel framework is used for the technique of Serial Femtosecond Crystallography (SFX) and comprises programs for data processing, simulation and visualization. It is a part of a complex, non-redistributable software stack, which is free to use by academia and non-profit organizations (Figure 15).

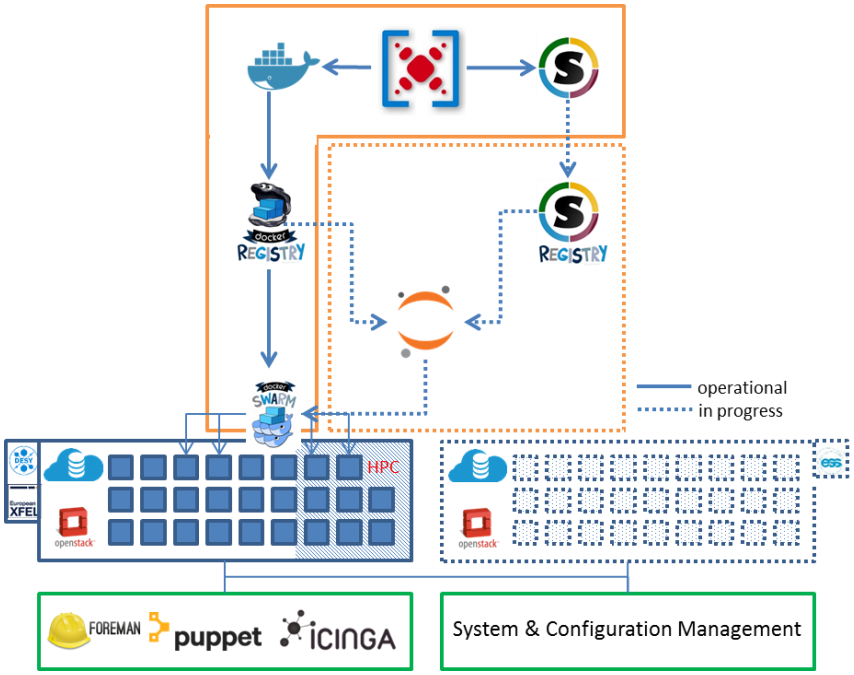


Figure 15: overview of the envisaged encapsulation of the SD frameworks for cloud deployment

The containerization of those applications as part of XFEL and CFEL data analysis services with interfaces to

data sources, additional software and a variety of other cloud services allowed to identify concrete interoperability requirements which are of importance for a successful integration of the analysed platforms and similar systems into the EOSC.

Tests on the local OpenStack infrastructure and on the HPC cluster have shown that the automation stack applied (Foreman, Puppet) is useful to spawn virtual machines on both types of infrastructures, reducing complexity and ensuring interoperability across platforms. However, this solution extends the range of external software and systems that are integrated into the provisioning process and therefore introduces interoperability challenges when solutions are migrated to other cloud platforms at partner institutes, where other automation platforms can occur.

To provide data exchange between VMs and external resources outside the OpenStack environment, we used AFS, CVMFS[32] and the local Owncloud instance. Configuration were easily integrated into the VM configuration and the deployment successfully automated as described above. Templating common data exchange configurations could help to foster interoperability in the EOSC. We will continue to work on this SD and further examine the integration of middleware solutions for mass storage systems like dCache and iRODS, which is a central interoperability aspect as they represent highly distributed systems and solutions to access data sources from different cloud providers. It appears the most suitable way to grant access to large datasets on a petabyte-scale (which is not too uncommon serial crystallography application).

In terms of network access, inter-network trust and speed and dynamically managed overlay networks for cloud VMs and container deployment in multi-cloud environments, we only have the DESY VPN to give access to resources and are interested in consuming services like the GÉANT Multi-Domain Virtual Private Network (MD-VPN). For container swarm communication this has to provide TCP and UDP connections. So far, we have only deployed container swarms on the local HPC cluster and cloud, but we are interested in an interoperability solution that demonstrates, how such swarms can be extended to operate in a multi-domain environment and how they can be migrated between clouds. Interoperable swarm and VM provisioning is an essential backend that enables Jupyter notebook users on Jupyter Hub servers in the cloud to run jobs on the underlying infrastructure.

The practical implementation and scaling of the swarms introduces a strong requirement for interoperable registries, and repositories providing trusted services to distribute container solutions. They should apply access and authorization management, user-attribute and role management, thereby keeping track of acknowledged licenses. To guarantee interoperability of such systems, they should be derived from some EOSC standards, AAI, licenses and certificates.

The integration of the cloud infrastructure in the DESY network poses questions about firewall configurations, offering services in the DMZ, local DNS integration and IP address registration. For this discussion, input from the EOSC, if there will be EOSC proxy servers, which network communication will be used within EOSC, outcome of the AARC2 project etc. will be valuable. For interoperable trust between networks and cloud providers, also comparable metrics to control SLAs and QoS are needed.

An additional point, that received a lot of attention, is the graphical output of containerized, parallel software as well as input via such GUIs (Mouse, Keyboard, instruments). We need solutions for single users and also for multiple users at the same time, esp. for training and education but also to provide cloud

[32] https://cernvm.cern.ch/portal/filesystem

services for distributed teams. We have demonstrated VNC-servers to work, but posing speed and performances challenges and X11 forward in the VPN to work nicely, but introducing potentially severe security concerns. We see a demand for EOSC provided solutions that connect graphical frontends to cloud VMs, enabling interactive usage models, giving graphical input channels to running processes and graphical output channels to deliver e.g. intermediate results. This can facilitate the migration of workflows, but also users and developers from HPC Clusters to Cloud Services. No special interoperability needs are reported by this Science Demonstrator.

2.1.4. Collaborative semantic enrichment of text - based datasets (TEXTCROWD)

The scientific areas of this SD are Digital Humanities (DH) and Cultural Heritage (CH). The proposed pilot will support the Natural Language Processing (NLP) encoding/metadata enrichment of text documents that are the main part of datasets used in DH-CH research, and usually have poor metadata.

Machine learning technologies have suddenly acquired considerable importance, especially in disciplines such as archaeology, where the main information is contained in free text documents rather than in relational databases or other structured datasets. The Social Sciences and Humanities research communities face a fragmented research landscape as well, that can be supported by EOSC.

The EOSC would help overcome such fragmentation, by building on structuring and integrating initiatives such as the CLARIN[33], DARIAH[34] and E-RIHS[35] ERICs, and Digital Humanities Organizations (e.g. their Association ADHO[36]) to offer advanced text-based services addressing common research needs (see recent survey by PARTHENOS[37]). 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.

TEXTCROWD is an advanced cloud based tool developed within the framework of EOSCpilot project for processing textual archaeological reports. The tool has been boosted and made capable of browsing big online knowledge repositories, educating itself on demand and used for producing semantic metadata ready to be integrated with information coming from different domains, to establish an advanced machine learning scenario.

Services, technical and interoperability **requirements** to support this SD pilot are:

- Identity and Access Management: federated identity and authorization, especially for “orphan” researchers
- Global Data Access: provide access to distributed storage resources, for users who already have their own systems.
- Online Storage: access text datasets to be processed, reference datasets (vocabularies specific for each domain/language) and processed metadata. Maybe the service should be used to store and share small - scale research data, which probably are the majority of datasets for the relevant communities
- Cloud Compute: create a virtual environment where the NLP tool will be invoked by researchers, operating on datasets stored in the Online Storage facility.
- Technical & software:

[33] https://www.clarin.eu/
[34] http://www.dariah.eu/
[35] http://www.e-rihs.eu/
[36] http://adho.org/
[37] http://www.parthenos-project.eu/

- Storage: max 500 datasets, approx. 100K each = 50 M. Possibly storage and processing issue may affect a production system in the future, as there is a large number (> 2 millions) of small files to be processed. In other words, scalability might become an issue when moving to a production environment. However, there are workarounds also for this case, as the datasets do not need to be processed altogether and once processed they are almost completely static.
- The datasets should be stored in cloud storage with access control, sharing capabilities and RESTStyle API (e.g. OwnCloud)
- No HPC required.
- Virtual machine configured to run tools of the GATE family[38]. A VM with at least 4 CPUs, 8 GB RAM, 100 GB storage and a common Linux distribution (Debian/Ubuntu, RedHat, Suse etc.)

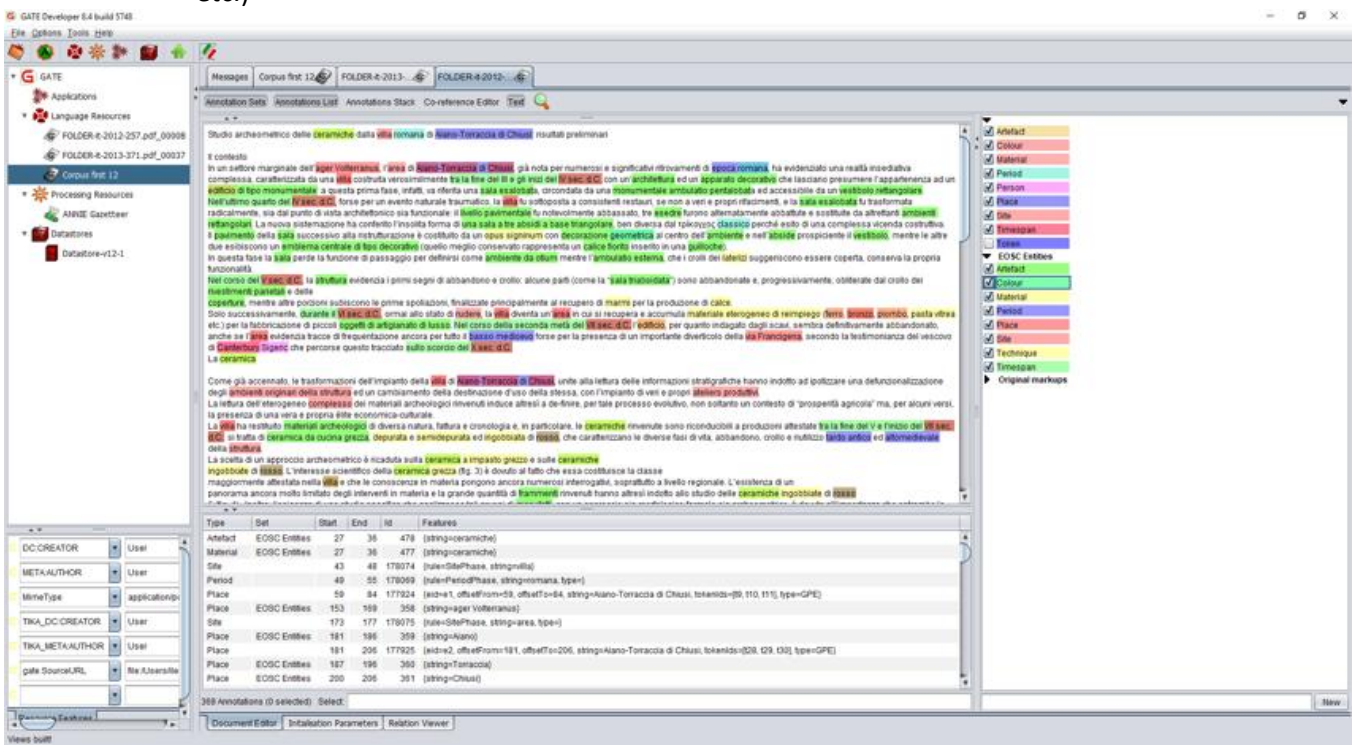


Figure 16: GATE tools (desktop view)

- Interoperability:
  - **all interoperability issues were already solved** as regards the common data model/underlying ontology. In the main phase after the pilot, more detailed access controls will be necessary. EUDATs **B2Access** could be a candidate for managing AAI

A virtual meeting was held with the D4Science team of CNR, in charge of providing **VRE cloud facilities** for the final deployment of TEXTCROWD, to develop a strategy for the migration of the tool in the cloud planned for next period. Discussion on technical aspects of this activity, aimed at clarifying and solving potential issues for final deployment, was going on within the technical teams.

At the time of writing of this report the implementation of the necessary components into the VRE has already completed and it offers:

- An workflow engine with GATE pipeline, operated as RESTstyle web services (running in Sheffield)
- Intuitive, web based user interface (Figure 17)
- User management

[38] https://gate.ac.uk

- Storage (private and shared files) (Figure 18)

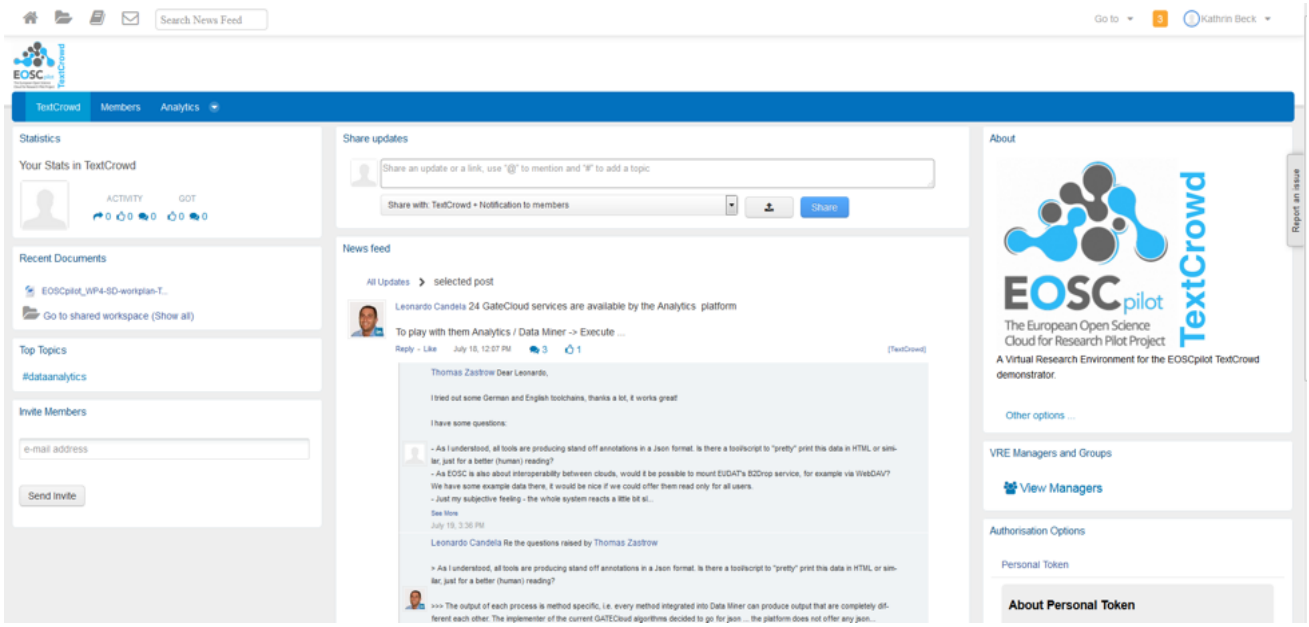


Figure 17: TEXTCROWD D4Science Dashboard

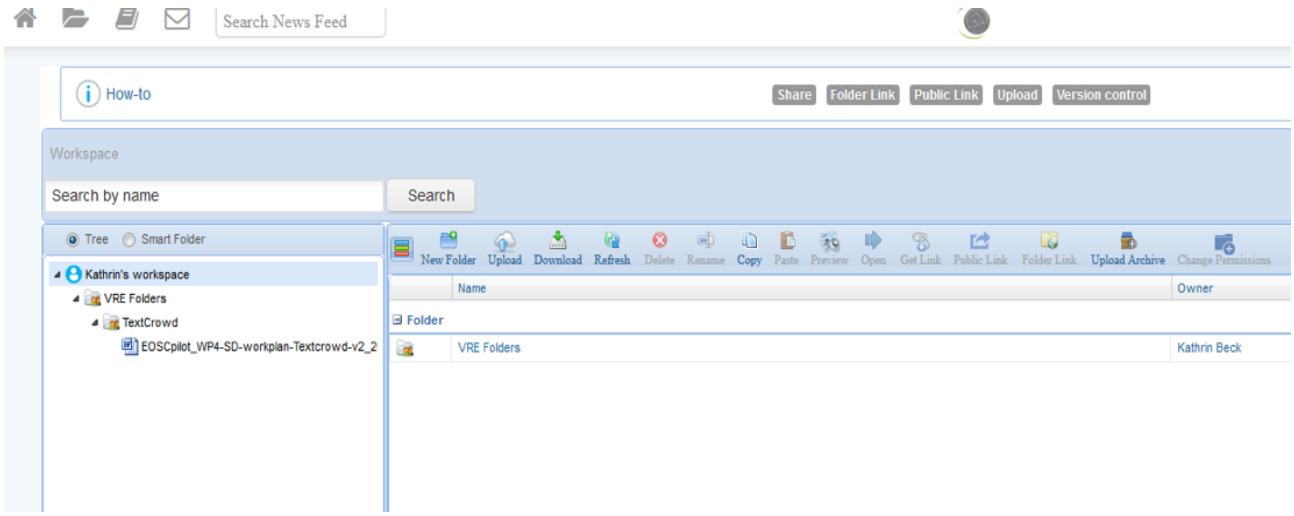


Figure 18: TEXTCROWD D4Science Storage

The TEXTCROWD service (https://textcrowd.d4science.org), empowered by the D4Science infrastructure, allows users to upload and store textual documents in a personal cloud folder, perform NLP and Named Entity Recognition (NER) operations, trigger the semantic enrichment process and get CIDOC CRM information in RDF. Results can be uploaded in a triple store or in another semantic enabled system and reused within the same context or in another VRE scenario on the same cloud. A schematic view of the pipeline is presented in Figure 19. The demonstrator was deployed inside an **EGI VM** with EUDAT **B2DROP** linking to **D4Science environment.**

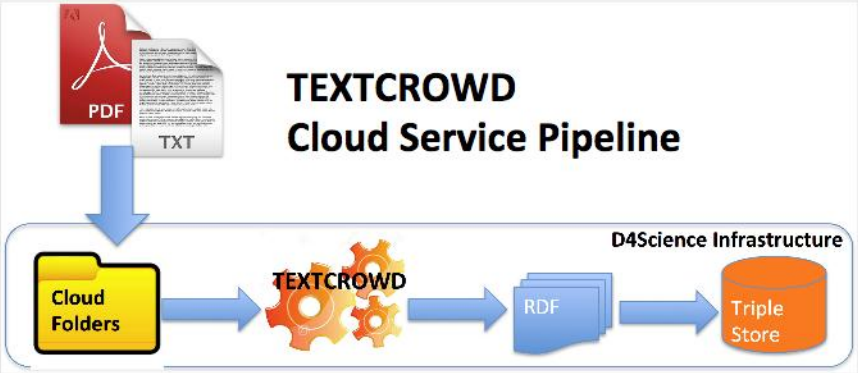


Figure 19: TEXTCROWD pipeline

2.1.5. WLCG Open Science Demonstrator - Data Preservation and Re-Use through Open Data Portal (DPHEP)

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. CERN is a practitioner of Open Science continuing on from the role played in terms of Open Access to publications. The LHC experiments all have data policies that call for their data to be preserved, for public releases of subsets of their data after a relatively short embargo period and for the ability to reproduce (at least key) analyses. (Data must be preserved before it can be re-used or shared). Experience from these Open Data releases has shown benefits not only in closely related disciplines (e.g. theoretical physics) but also in computer science in general. 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 high energy physics science demonstrator wants to deploy services that tackle the following functions:

1. Trusted / certified digital repositories where data is referenced by a Persistent Identifier (PID);
2. Scalable "digital library" services where documentation is referenced by a Digital Object Identifier (DOI);
3. A versioning file system to capture and preserve the associated software and needed environment;
4. A virtualised environment that allows the above to run in **Cloud, Grid and many other environments**

Services, technical and interoperability **requirements** to support this SD pilot are:

- Reliable "bit storage" in a trusted (certified) digital repository. Around 100TB of storage should be provided (eventually) to test at a realistic scale. Exposed via service such as **B2SAFE** preferably with integrated DOI generation on data ingest.
- Some "digital library" solution for storing documentation with DOIs (e.g. **B2SHARE** or some other Invenio[39] - based solution would seem most appropriate). (The CERNLIB long - writeups require 64MB of storage).
- **CernVMFS** for storing the software and associated configuration parameters. Again a minimal amount of storage would be required (few TBs). The software needed to process the data is specific to the experiments which produced the data and will be provided via CVMFS, which will act as a software distribution service.

[39] http://invenio-software.org/

- The above three services would allow the basic setup to be made and would need to be complemented by an additional service to allow the data to be "re-used" on multiple cloud environments, namely: **CernVM**
- The data processing step will require **efficient staging of the archived data sets** (recall from TAPE if needed) and the allocation of on - demand compute resources (computational resource requirements are likely to be low, handful of small virtual machines, not HPC). These resources are expected to come from a cloud providers, **CERN openstack, EGI Cloud resources, Commercial Cloud provider** (testing multiple cloud providers would be beneficial – allowing to show cross platform/infrastructure viability). The requirement of a low number of commodity level VMs will probably facilitate testing across providers.
    - data processing CernVM images will be used as a baseline image for data processing , these images will be deployed on the cloud resources
- **Interoperability requirements**:
    - for the e-infrastructure level interoperability - Will need to see which resources are used first , which e - infrastructure provider provides them
    - regarding data-interoperability:
        - Data (FAIR):
            - F - CERN open data portal (or equivalent) - Provides baseline service for data description (findability) - Metadata and DOIs. To be evaluated if/how this would need to be federated with other services in EOSC.
            - A - Resolvable datasets - DOIs assigned to data - need to evaluate how we resolve them. The AAI - may not be strong issue as data is openly available (public, anonymous access)
            - I - Need to address in later stages Q3 onwards Will be using standard services/protocols from e-infrastructure providers - so that's a good start
            - R - Need to address in later stages Q3 onwards.

The main goal of this demonstrator use-case is to demonstrate "best practices" regarding data management in the arena of LTDP, "open" data (sharing and re-use) - how can be realized on the EOSC. The pilot is equivalent to CERN Open Data Portal but using EOSC resources, thus allowing this solution to be opened to other communities. In Figure 20 we can see the expected interoperation between e-infrastructures, showing EGI and EUDAT services/resources required that are mapped to this use case.

Service	HEP	EOSC
Trustworthy Digital Repository (TDR)	CERN Castor+EOS	EUDAT TDR (part of CDI)
PID/DOI systems		EUDAT B2Handle
Digital Library	CERN Document Server	EUDAT B2Share (Zenodo)
Software + Environment	CVMFS + CernVM	CVMFS + CernVM Tested on EGI FedCloud

Figure 20: Mapping the WLCG/DPHEP use-case to services

The following figures present the frameworks solutions for the software and environment case (Figure 21) and data archive (Figure 22). The status of the pilots and the results of the tests will be further described in the report D6.5 - "Interim Interoperability Testbed report".

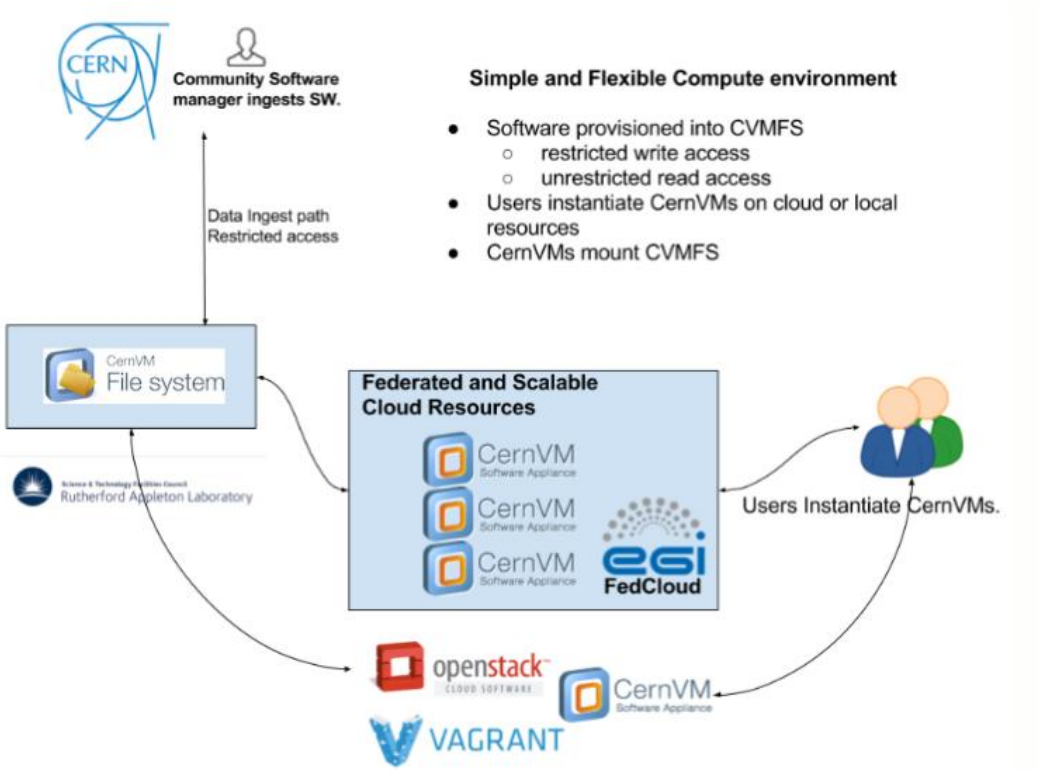


Figure 21: DP-HEP - Solution for Software & Environment Preservation

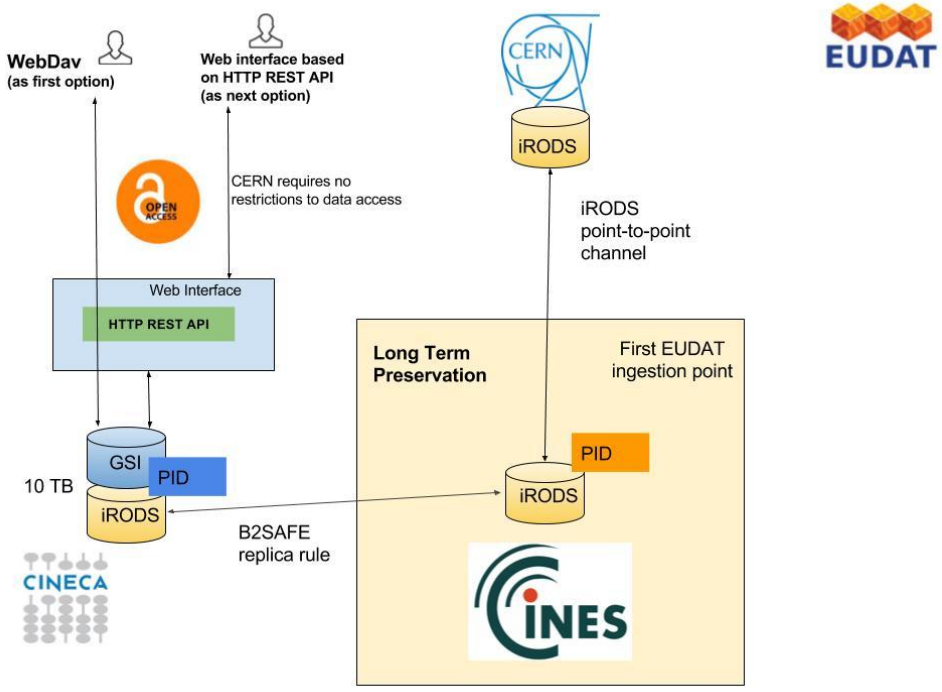


Figure 22: DPHEP - Data Archive Solution

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

2.2.1. HPCaaS for Fusion (PROMINENCE)

The scientific domain of this SD is the Fusion Energy Modelling, and the ultimate goal is to provide HPCaaS by making computing nodes available through an OpenStack API.

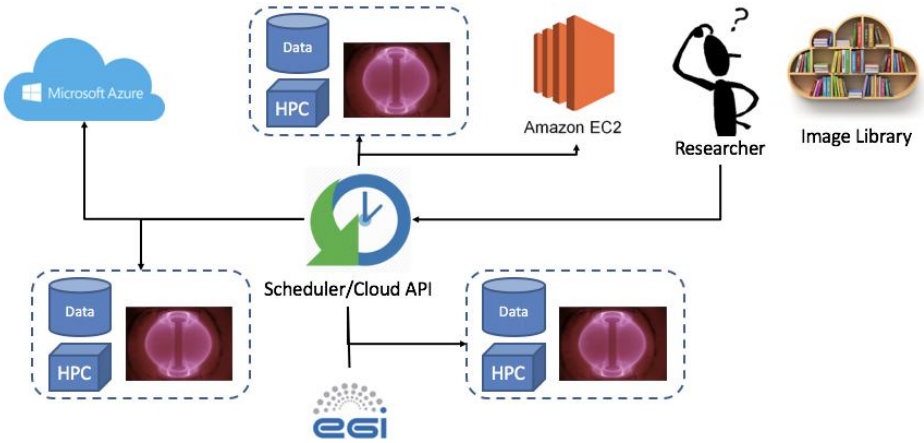


Figure 23: PROMINENCE - top level concept

The overall aim of the work on this pilot is to demonstrate the feasibility of launching MPI/OpenMP jobs on the EGI FedCloud (and other cloud instances) and allow external communities to assess whether HPC class nodes accessible through such a cloud interface are useful to them in their work.
Services, technical and interoperability **requirements** to support this SD pilot are:

- EGI FedCloud where we intend to not only act as a provider by making local resources available through an OpenStack API, but also as a consumer to demonstrate the principle of running MPI codes on other FedCloud sites
- EGI Cloud Container Compute for on-demand cluster provisioning (see later for resource requirements), plus sufficient scratch storage for the output of applications
- EGI 'generic' VO for demonstration
- INDIGO - DataCloud Synergy[40] solution for scheduling, but will need to understand how it works compared to Kubernetes
- Indigo Core PaaS for accounting and monitoring

[40] https://indigo-dc.gitbooks.io/synergy-doc/content/

If time allows, using Indigo Workflow plugin would be of interest (although not accounted for in the proposal, this would be interesting to investigate since many fusion problems require coupled multi-physics workflows.

2.2.2. EGA Life Science Datasets Leveraging EOSC

The scientific domains of the SD are the Computer and information Sciences and Biological sciences.

This demonstrator will leverage **EOSC resources** to refresh datasets from uploaded to the EGA[41] using newly available or updated reference data. Doing this, the new dataset will also be made available in a FAIR manner, adding metadata according to the attributes that have been chosen to contribute the strongest to the FAIR principles. Pipelines and security mechanisms will be developed as part of this demonstrator to automate this process. Figure 24 show the original and updated EGA pipelines, to be used to enable reproducibility, allow portability and Data re-analysis the main goals of this SD, by:

- Using a third part dataset (GoNL project) as use case
- Reproduction of the original pipeline
- Production of an updated pipeline
- Containerized versions of both pipelines
- Test both pipelines on the use case dataset

The **main requirements** to achieve these goals are:

- Data storage to handle and manipulate data
- Storage provider must have options for data security and privacy compliance
- Moderate amount of VM resources (5 or more 8-core VMs, 16GB RAM) for running automation and data management tasks
- Definition of data formats that maximize interoperability and FAIRness

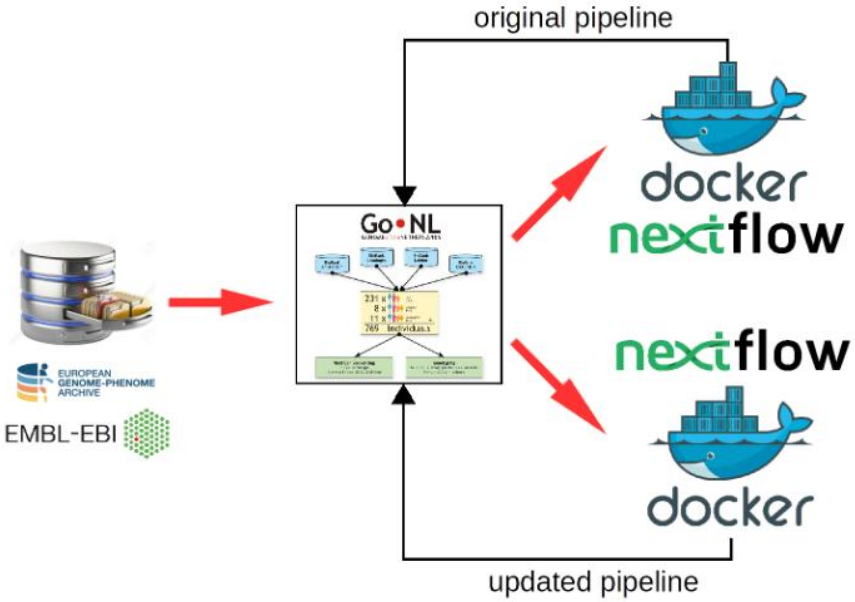


Figure 24: EGA original and updated pipeline

The **metadata management of research artifacts** (including persistent identifier) is the main interoperability issue raised by this Science Demonstrator.

[41] https://www.ebi.ac.uk/ega/home

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

This scientific demonstrator aims to demonstrate the interoperability of the EPOS/VERCE VRE, presented in Figure 25, with e-Infrastructures of the EOSC, and offer the possibility to compute realistic scenarios of earthquake shaking, visualise and compare the results to recorded strong motion records, to address the needs of civil protection agencies for more accurate earthquake scenario estimates.

To achieve this goal, a dedicated EGI Virtual Organisation (verce.eu) has been registered in the EGI Operations Portal[42], and some cloud providers of the EGI Federation have been identified to allocate the initial set of resources to execute earthquake simulations on the EGI Federated Cloud Infrastructure. There is no need for additional resources at the moment. Before to execute HPC simulations on cloud-based resources, some preliminary work has been carried out in order to enable the support of FedCloud Cloud infrastructure as backend resources for Workflows of the Portal, and extensions of the Science Gateway frontend and backend services to allow the implementation. This activity has also requested the update of the DCI_BRIDGE Virtual Appliance[43]used by the WS-PGRADE/gUSE portal to interact with the EGI Federated Cloud infrastructure.

In the last months the SD has also worked to extend the AAI mechanism used by the WS-PGRADE/gUSE portal in order to:

- enable federated access with the OIDC[44] module developed for Liferay by the INFN Catania
- add support to Per-User Sub-Proxy (PUSP)[45].

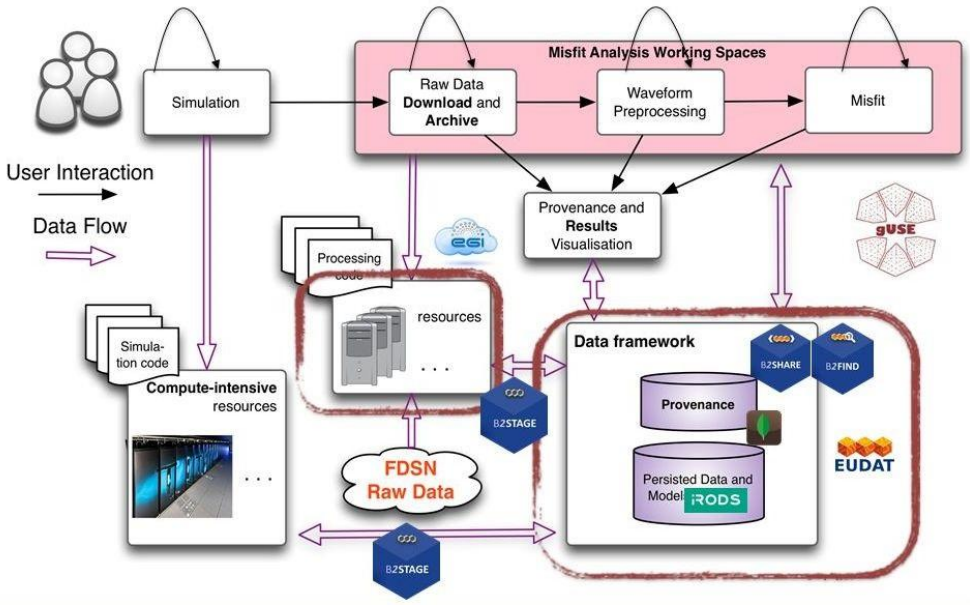


Figure 25: EPOS/VERCE architecture

The portal is now configured to make REST calls, and download, from the eToken server proxies certificates generated from a robot registered in the verce.eu VO. A portlet has been created to download the proxy and include it in the gUSE certificate store for use in workflows. The integration of the OIDC module and enable federated authentication mechanism is still work in progress.

[42] https://operations-portal.egi.eu/

[43] https://appdb.egi.eu/store/vappliance/fedcloud.slave.dci.bridge

[44] https://github.com/csgf/OpenIdConnectLiferay/tree/EGICheckIn

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

Download, preprocessing and Misfit workflows have been fixed for bugs and validated by domain scientists for their correct behavior. The frontend has been improved in multiple aspects for misfit calculation, e.g. on the Download Tab (for setting up the acquisition of measurement data from data centers), new query string parameters are automatically defined to limit the search and download of raw measurement data to the corresponding networks and stations of the simulation run. Also, following in the Processing Data Setup Tab, only raw-data download runs which correspond to the selected simulation will be shown.

The S-ProvFlow system has been improved in many aspects: Better Rest API methods, Provenance Repository performances, Frontend usability and modular "Dockerisation" of each component. Current development branch at https://github.com/aspinuso/s-provenance

For data resources, the pilot aims to utilise services and resources available through the EUDAT e-infrastructure.

2.2.4. CryoEM workflows

The scientific domain of the SD is the Structural Biology. The overall objective is to develop ways to share detailed information on cryoEM image processing workflows, concentrating on those processes usually run at the Facility level. This work should increase reproducibility in Science.

The idea is to write from Scipion[46] framework a workflow file that fully describes the image processing steps so that they can be re-executed resulting in exactly the same results (making the data more FAIR). This file should go with the raw data as acquired by large facilities in Europe (like Diamond and ESRF synchrotrons) as well as smaller EM facilities like (Necen, SciLife Lab, or CNB-CSIC). We foresee that in some facilities, this image processing workflow is performed on the cloud so that the technology employed must allow for this possibility.

Most of the work must be performed at the software currently used for executing these image processing workflows. It needs to be adapted so that:

- It writes a description file that accompanies the raw data and serves as a summary of the image processing steps.
- This file can be read and re-executed by the workflow manager resulting in exactly the same results.
- The file can act as a workflow template for processing new data.

The execution of the image processing workflow can be performed either locally or in the cloud, and the technology must allow both situations.

Once the software is adapted, it will be tested in one of the collaborating **EM facilities** and the possibility of using the **cloud for regular processing** in a daily basis for users will be checked. This test will comprise the last quarter of the project, and will require a machine with 16 CPU cores and 2 modern GPUs. Data is acquired at a rate of 1TB/day and is expected to stay at the server for at least 2 weeks.

As **interoperability requirement**, the Science Demonstrator needs to make sure that the analysis software to be run at the Facility complies with all **complex technical requirements** normally found in a large installation, like a synchrotron:

- the SD will work together with ELIXIR in terms of ontological development, centered in extending EDAM to cryoEM
- Most test Facilities will be Instruct-sites

[46] http://scipion.i2pc.es/

2.2.5. Astronomy Open Science Cloud access to LOFAR data

Science area of this SD is the Physical Sciences: - Research communities who benefit from to the science case are LOFAR[47] users, Square Kilometre Array (SKA)[48] user community, Radio Astronomy, Astronomy, Astrophysics, and other data-intensive research domains.

Objectives and goals:

- Existing LOFAR data will be **made readily available to a much larger and broader audience**, enabling novel scientific breakthroughs.
- **Data integration and data interoperability**: The open science enabled by this project, in combination with the EOSC ecosystem, will be a catalyst to make this happen with LOFAR data as well. This is of key importance, since LOFAR is two orders of magnitude more sensitive in its frequency range compared to previous instruments
- The LOFAR archive is already developed and demonstrated to be working at scale on existing infrastructures. The International LOFAR Telescope (ILT) consists of user consortia in 6 EU countries and three data centers (besides SURFsara, there are centers in DE and PL). The partners are committed to **publishing papers, datasets, tools, and workflows following the Open Science model** with the application of FAIR principles, and also as part of the Open Research Data Pilot in H2020
- Although this project deals with a concrete use case in radio astronomy, many of the existing tools we will **port** to the EOSC ecosystem (i.e., Xenon, CWL, Docker, Singularity) are broadly applicable. NLeSC already uses the tools in urban modelling, coupled climate simulations and virtual research environments in chemistry and archaeology

EOSC service requirements:

- computational platforms - one or more.
    - SURFsara HPC & HTC, LISA & Cartesius clusters
    - PSNC, FZJ
- integration with and/or use of federated AAI infrastructure offered by EUDAT & EGI FedCloud
- for registration and sharing of resulting data products - will evaluate and utilize FAIR services from e.g. EUDAT, EGI, and NLeSC
- bring processing to the data, by initiating jobs on processing facilities that are colocated to the LTA storage resource that hosts the data

Technical services needed:

- HPC/HTC cloud infrastructure
- LOFAR LTA storage
- LOFAR processing software
- EUDAT registration services, B2SHARE

2.3. Third Set of Science Demonstrators

The second 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

[47] http://www.lofar.org/

[48] https://www.skatelescope.org/

- Life Sciences and Health Research: Mining a large image repository to extract new biological knowledge about human gene function.
- Earth Sciences: Switching on the EOSC for Reproducible Computational Hydrology by FAIRifying eWaterCycle and SWITCH-ON

2.3.1. Frictionless Data Exchange Across Research Data, Software and Scientific Paper Repositories

The proposed SD will work to pilot a demonstrator service for fast and highly scalable exchange of data across repositories storing research datasets, manuscripts and scientific software. The data exchange in the demonstrator will be based on the ResourceSync protocol[49], that was designed and the first scalable implementation of which has been developed and deployed by the project team. This work will enable more efficient and effective information exchange between EOSC data providers and services, which is an essential step towards the realisation of the EOSC vision.

The demonstrator will provide argument for modernising existing legacy communication mechanisms routinely used by thousands of research repositories. The code developed for the demonstrator will lower the barrier to adoption of ResourceSync across data providers irrespective of scientific discipline. In order to achieve this goal, the SD will:

- Quantitatively evaluate ResourceSync against OAI-PMH[50] on the use case of aggregating millions of resources from scientific repositories to the CORE[51] aggregator and enabling others to keep in sync with relevant parts of this dataset.

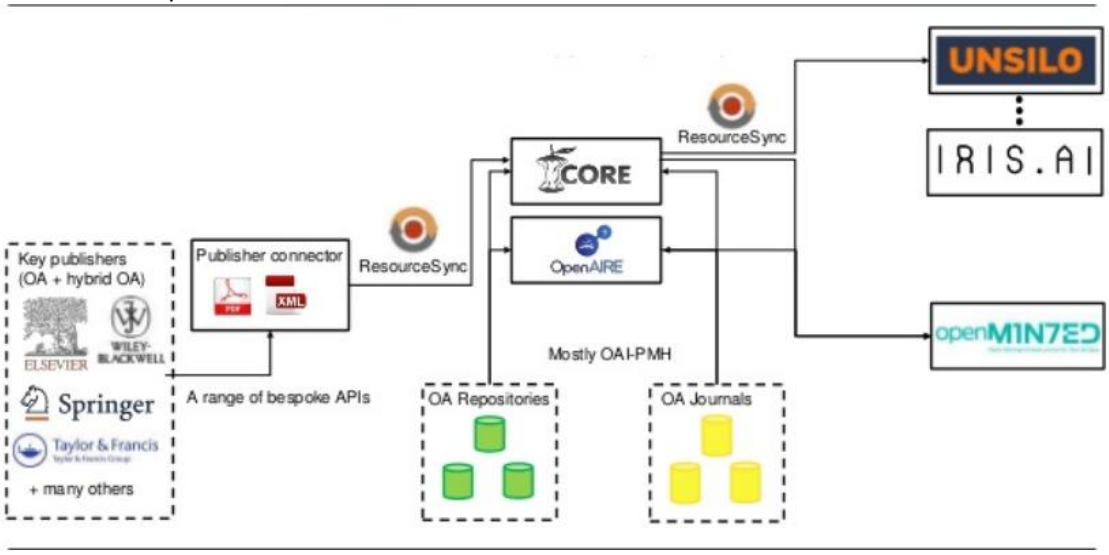


Figure 26: ResourceSync use-case - exposing enriched dta for TDM [52]

The demonstrator will showcase how scholarly communication resources, i.e. **research datasets, scientific manuscripts** (research papers, theses, monographs, etc.) and **scientific software**, can be effectively, regularly and reliably exchanged across systems using the ResourceSync protocol. It will apply ResourceSync on real-world use cases with millions of resources and a representative set of repository and service platforms reflecting the diversity of the EOSC ecosystem. The data synchronization will be shown:

1. across a **cross-disciplinary network of repositories** and
2. between **repositories and global value-added services**, such as those used in research evaluation, aggregation and workflow execution.

Initial requirements of services, technical and interoperability to support this SD pilot are:

- the demonstrator will be based on the **ResourceSync protocol**.

[49] http://www.openarchives.org/rs/toc
[50] https://www.openarchives.org/pmh/
[51] https://core.ac.uk/
[52] https://www.slideshare.net/petrknoth/seamless-access-to-the-worlds-open-access-research-papers-via-resourcesync

    - first implementation of ResourceSync that scales to millions of items has been developed and deployed by the CORE team at the Open University, UK (OU)
- the demonstrator will be used to evaluate the efficacy of this solution, benchmarking it against the current state-of-the-art (mainly OAI-PMH) in terms of:
    - speed (time)
    - complexity (steps required to complete)
    - reliability (recall)
    - freshness (e.g. average time gap between syncs)
- different platforms considered will include repository platforms, such as EPrints, DSpace, Fedora or CKAN
- different services will be considered, such as GitHub, CrossRef, Dryad, OpenAIRE, CORE or PMC
- deployment within the CORE infrastructure[53], providing a global aggregation service for repositories:
    - As of September 2017, CORE provides access to over 80M metadata records, 8M full texts and has over 1.5M monthly active users.
    - it consists of several small enterprise clusters for harvesting, indexing, processing and storing data
    - storage capacity is ~50 TB and is expected to double every 3 years.
    - it is also a primary user of an OU shared big data cluster facility, utilized for running Hadoop & Spark jobs of the CORE's ingestion pipeline.
    - This infrastructure is a good fit for meeting the **required level of processing and data storage for the development and deployment** of the proposed demonstrator and moving it to production following the end of the project.
- interoperability needs:
    - The key data challenges to be addressed are:
        - Improving **interoperability of scientific repositories**.
        - Addressing issues around the **low findability and accessibility of research artefacts by automated means**.
        - Increasing scalability of data exchange between scholarly communication systems
    - Availability of a diversified set of datasets to test on

The team is highly interested in working with the EOSC pilot and has already initiated a conversation with the TEXTCROWD and High Energy Physics EOSC demonstrators.

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

This work of this SD will leverage an existing strategic collaboration between the Euro-BioImaging - and BBSRC-funded Image Data Resource (IDR; http://idr.openmicroscopy.org) and the EMBL-EBI Embassy Cloud. IDR holds >40 systematic imaging datasets, >1 Mio experiments, and imaging data related to >19,600 human gene and >31,000 drugs or small molecule inhibitors. In this pilot project, we will establish the resources required to perform comprehensive machine learning analyses on these datasets, with the ultimate goal of identifying functional connections between genes and/or small molecules that target them based on image-based phenotypes. The pilot will test the validity of this approach and also demonstrate how a large cloud -based collection of published datasets can be re-used for novel discovery. Besides generating testable hypotheses about cellular functions, this study will also produce a reusable infrastructure and analyses for generating value from published image data.

Initial requirements of services, technical and interoperability to support this SD pilot are:

[53] https://core.ac.uk/dataproviders

- the project will analyse **publicly available image data** from genome - scale RNAi screens to gain insights into cellular functions of human genes.
    - This requires organizing **access to the large amount of image data** available at the IDR and performing machine learning analyses of these datasets.
- use the following EOSC resources:
    - access to compute resources:
        - vCPUs: initially - start with one VM for development and testing and will require scaling up to 250 vCPUs for the actual computation.
        - RAM: Initially each job will require 4 GB, however at later stages, jobs will require more (at least 32 GB, possibly up to 60 GB)
        - Storage: 1 TB of NFS - like storage shared between instances.
- access to the IDR database
    - currently deployed in the Embassy Cloud. For access - own tenancy to access the DB via the provided API
- access to ELIXIR resources
- software tools:
    - wnd-charm (https://github.com/wnd-charm/wnd-charm)
- R (https://www.r-project.org/), Python (https://www.python.org/), MariaDB (https://mariadb.org/)
- the SD implements a typical image analysis workflow of image processing for feature extraction followed by machine learning - based data analysis of the feature data. Moving the large amount of image data required by this project is highly impractical so compute has to be brought to the data

2.3.3. VisIVO: Data Knowledge Visual Analytics Framework for Astrophysics

This SD examines cloud based services for the astrophysics community. For this particularly data-intensive research, the VisIVO framework has been used for studies of the star forming regions accessing data from the Hi-GAL survey. The SD approach is the integration of VisIVO in the EOSCpilot e-infrastructure. The application is specifically designed to process and visualize multidimensional data, using advanced visualization algorithms and techniques including vector, tensor, scalar, texture, and advanced volumetric methods.

The work will be carried out on the own local OpenStack infrastructure at INAF. At a later stage in the project, in order to test interoperability, readiness for FedCloud integration will be assessed. Also interoperability with the B2STAGE EUDAT service and migration of workflows and software deployments between HPC environments and cloud environments will be addressed.

As a first task we foresee the implementation of a WS-PGRADE/gUSE portal frontend layer and communication with a DCI bridge, which is a standard interface that allows on demand provisioning of user services in Grid and cloud environments.

The interoperability with present solutions for GUI-based I/O will therefore be examined and guaranteed, for this purpose the work plan foresees the collection of best practices in establishing interaction with cloud instances by serving graphical user interfaces on the application layer.

Aiming to adhere to FAIR principles where possible, another task of this SD is to establish FAIR access to approx. 1TB of astrophysics survey data, including compact sources, filaments, bubbles, for which sustainable storage and sufficient upload is requested. This can be used to compare data repository services in the EOSC. Having restriction of use for some datasets, interoperable access control and authority management will be examined as a part of this task, too. In order to access the data, interoperability with Virtual Observatory Table Access Protocol will have to be supported so that that datasets can be searched by specifying the identifier of an object.

The main interoperability requirements raised by this Science Demonstrator is to use EUDAT services for supporting the FAIR principles.

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

This EOSC science demonstrator will combine lessons learned from SWITCH-ON[54] and eWaterCycle[55] to make the next step towards reproducible, reusable, open, data driven science in Hydrology, by showcasing a modern approach to large scale simulation - based science. It will create FAIR versions of data required for running involved hydrological models. It will also make these available along with the results of respective models for further usage by researchers in both hydrology and related fields. All data will be stored using the **EGI DataHub** system based on the open source **Onedata** technology, allowing seamless access to data from different compute infrastructures.

Initial requirements of services, technical and interoperability to support this SD pilot are:

- current **interoperability between various hydrology models** needs to be established and expanded.
    - **guidance** will be needed with regards to the FAIRification of these standards and the associated metadata as well as the **availability of resources to present these data** in a FAIR way. This could be in **existing repositories (Zenodo, OpenAire)** or in a new, hydrology specific repository.
- **computational and data storage resources** will need to be provisioned
    - Contact for this has already been established with various providers but this process will be recorded for the purposes of EOSCPilot.
- **a data sharing solution** needs to be established for which **OneData** is the main candidate.
    - **support** is needed from CyfroNet or other parties with experience in setting up OneData (EGI DataHub).
- interoperability needs:
    - As a common standard for coupling model the SD will make use of the **Basic Model Interface** (BMI) from the CSDMS project (http://csdms.colorado.edu) Additionally, the proposed tools within **containers** need to output data and the associated metadata that can be **uploaded to repositories such as Zenodo**.
- compute and data resources to be used:
    - The Dutch National Supercomputer Cartesius hosted by SURFsara.
    - The SURFsara HPC Cloud system.
    - The SuperMUC high-end supercomputer at the Leibniz-Rechenzentrum.
    - The EGI federated cloud through the CYFRONET service provider offering an OpenStack-based system.
    - Onedata technology from CYFRONET through the EGI DataHub to store, retrieve and discover data
- open science services envisaged:
    - Open Datasets used in Global Hydrological models such as NOAA and/or ECMWF weather forecast ensembles and reanalysis data, GRDC discharge data, HSAF satellite soil moisture products, etc.
    - Data in open standards such as NetCDF.
    - Open Source software models as part of the eWaterCycle and SWITCH-ON projects.

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

This Demonstrator aims at contributing to the following science area: Cultural Heritage (CH), i.e. museums and artworks, CH restoration, Archaeology - History and archaeology, and Art. However, the service may be

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

[55] http://www.ewatercycle.nl/

useful in any other scientific field producing and using images and 3D models, and thus it could be used by other infrastructure projects.

The availability of a generic public web service demonstrated with the ARIADNE Visual Media Service (http://visual.ariadne-infrastructure.eu/) has been welcomed by users, who have however pointed out possible improvements of the current service version. The Science Demonstrator will provide researchers with an integrated Virtual Research Environment where to publish on the web, visualize and analyze images and 3D models. The VRE will offer an effective common workspace enabling sharing, interoperability and re-use. The service supports publication on the web and browsing of three types of visual media:

- High-resolution 2D images (input converted in a multi-resolution format and browsed in real-time, zooming in and out);
- Reflection Transformation Images (RTI), also known as Polynomial Texture Maps (PTM) images, i.e. dynamically re-lightable images;
- 3D models (triangulated meshes, point clouds and textured models).

For each media type, the service supports automatic conversion to an efficient multi-resolution representation, offering data compression, progressive transmission and view -dependent rendering. Each media is presented visually with a standard browser (one for images, another for RTI images and a third one for 3D models). In the case of 3D models users may customize the behavior and graphical interface of the 3D browser, since alternative templates for 3D content are provided, to address specific sub-types of 3D scenes and to extend the flexibility of the overall system. As regards the metadata, the service allows a manual input using a simple form. The redesign of the system and its integration in a VRE will also improve discoverability, supported by the metadata enrichment component. A possible integration with TEXTCROWD will be evaluated for this purpose in this Demonstrator

Initial requirements of services, technical and interoperability to support this SD pilot are:

- Authentication
    - A first service component definitely missing in the current implementation of the Visual Media Server is **support for the user authentication**. The current service does not require a registration, but simply asks to the user some basic data (name, institution, email) every time a visual data file is uploaded, by means of a web form.
        - planning to exploit the **D4Science authentication, authorization, and accounting framework** delivered in each D4Science. It represents a more solid authentication component to allow a user to easily manage his own data, create and manage collections, insert groups of media in batches. The creation of “media groups” would make it possible to control access to proprietary data with better granularity, simplify sharing media with colleagues and collaborative publishing.
- Scalable Storage
    - The Visual Media Server (current implementation) runs on a single server; users' data (both raw input files and processed multiresolution files) are stored on the local hard disk. With a possible substantial increase of the users population this original model would not scale.
        - planning to exploit the **D4Science storage framework delivered in each D4Science VRE** (permanent storage of visual media files)
            - usual file size is 50-150 MB for RTI and 3D files, 10-50 MB for high resolution images. Users could be granted some cloud storage to upload large files more conveniently.
- Scalable Processing
    - The processing requirements are manageable on a single server, unless the user population increases substantially. Here the variable to consider is the average and peak numbers of contemporary users in the two main phases of use of the Visual Media Server:
        - Data upload (and subsequent file type conversion and processing);
            - Visual Media Service requires some significant processing over the input datasets (executed only once at data uploading time), since there’s a need

to convert the data in web-compliant formats (which should support easy progressive transmission of the data, adopting multiresolution encoding scheme and decomposition of the data to support view-dependant rendering functionalities). Data upload does scale quite well, since data conversion times are small (5 seconds for a 67M pixel RTI or 60 seconds for a triangle mesh with 1M triangles). Anyway, **if the number of users increases (uploads of several datasets in the same instant of time) this preprocessing phase could become a bottleneck.**

        - Interactive visualization.
            - In the current framework, i.e. with only a single server, it is estimated to be able to serve up to 100 simultaneous users, exploring the models/images in the same instant of time with high quality of service. But, in case the number of contemporary users will increase significantly (at the data access and visualization stage), it could also justify a more **sophisticated data access management** (i.e. data stored on multiple servers). Testing the demonstrator in a **cloud environment** should allow us to benchmark scalability, due to the larger number of users.
    - The SD will allow the empirical evaluation of the level of use and load of the system (using the D4Sicence monitoring system that keeps track of the resources used by the system).
        - planning to evaluate the possibility of using the D4Science load balancer enabling in this way a more sophisticated policy for the allocation of computing subtasks to multiple servers

3. CONCLUSIONS AND FUTURE WORK

In this report the first interoperability requirements from the different science demonstrators have been reported. The aim of these requirements gathering is to contribute to the definition of the EOSC Service Portfolio, so that the services envisaged are aligned with the needs and expectations of researchers.

In the next period the various Science Demonstrators presented above will continue their work on setting up the pilots foreseen by their work plans and will provide not only their first results but also possible new requirements.

In particular, the next period will see:

- for the PiCo2 - Pilot for Connecting Computing Centers project that during the first year has developed well within a group of several Tier1 and Tier 2 infrastructures. Data flows run fluently between centers, and the technical groups set up for solving the issues raised for those connections are working well.
    - For 2018, beyond carrying on the work of existing technical groups, our aim is threefold
        - expand the network to European scale, especially within WP6 and WP 6.3 partners (Italy, Germany), but not limited to
        - integrate some use cases, not as such (will not develop something specific for a specific community), but to have a portfolio of some life-size use cases in different disciplines for full scale tests (just to be in line with what is done in WP4 and WP5, and not interfere with science demonstrator procedure)
        - organize as well a flow of codes (compilation, runtime, notebooks) between the machines Tier 2 / Tier 1
- the definition of the scoped pilot between AARC2 and EOSCpilot as an interoperability pilot, aiming also to assess whether the AAI solutions envisaged by the EOSCpilot meet the functional and

technical integration requirements of research communities and e-infrastructures as described in the AARC Blueprint[56]

- the setup of the four identified data-demonstrators, and see how their requirements are met
- the last set of five Science Demonstrators defining their work plans, deploying and testing the services and technologies proposed, analyzing the results obtained.

The process of requirements gathering and analysis represents a continuous work, that will see updated reports on M16, with the D6.5 - "Interim Interoperability Testbed report", M18 on D6.7 "Revised requirements of the interoperability testbeds", and M24 with the D6.10 - "Final interoperability Testbed report". Through the activities of setting up the SDs pilots, while meeting their requirements, work will be done in order to assess also the maturity level of solutions for what regards TRL, openness, scalability, user community adoption and sustainability.

ANNEXES

A.1. GLOSSARY

Many definitions are taken from the EGI Glossary (https://wiki.egi.eu/wiki/Glossary). They are indicated by (EGI definition).

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.

[56] https://aarc-project.eu/architecture/

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 numbers 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 seconds. 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.
National Grid Initiative or National Grid Infrastructure (NGI)	(EGI definition)The national federation of shared computing, storage and data resources that delivers sustainable, integrated and secure distributed computing services to the national research communities and their international collaborators. The federation is coordinated by a National Coordinating Body providing a single point of contact at the national level and has official membership in the EGI Council through an NGI legal representative.
Virtual Organisation (VO)	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).

AAI	Authentication and Authorization Infrastructure
CMS	Content Management System
EDMI	EOSC Dataset Minimum Information
EOSC	The European Open Science Cloud
FAIR	Findable, Accessible, Interoperable and Reusable
RDA	Research Data Alliance
RIs	Research infrastructures