D6.7: Revised Requirements of the Interoperability Testbeds

Author(s)	Doina Cristina Duma (INFN)
Status	Final version
Version	V1.04
Date	28/09/2018

Dissemination Level

☒ 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 revised requirements for the interoperability testbeds, and more generally for the interoperability at the infrastructural/networking level and of the various services envisaged to be part of the EOSC environment. It also considers the new requirements as reported by the partners participating to the WP6 activities, the Science Demonstrators representatives and their shepherds, and also how the initial requirements were addressed.

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-D6.7	
Deliverable lead	**INFN**
Related work package	**WP6**
Author(s)	D. C. Duma
Contributor(s)	**T6.3 team & SD shepherds**
Due date	**30/06/2018**
Actual submission date	**28/09/2018**
Reviewed by	**Giuseppe La Rocca & Nuno Fereira**
Approved by	**Mark Thorley (UKRI)**
Start date of Project	**01/01/2017**
Duration	**24 months**

Versioning and contribution history

Version	Date	Authors	Notes
0.01	15/05/2018	Doina Cristina Duma (INFN)	ToC available
0.02	25/07/2018	Doina Cristina Duma (INFN), Michal Schuh (DESY), Andrea Ceccanti (INFN), Xavier Jeannin (RENATER), Violaine Louvet (U. Grenoble), Alain Franc (INRA), Giuseppe la Rocca (EGI Foundation), John Kennedy (MPCDF), Kathrin Beck (MPCDF), Michael Schuh (DESY), Thomas Zastrow (MPCDF), Nick Juty (UMAN), Giorgos Papanikos (Athena RC), Roberto Scopigno (CNR)	Version including contributions received from some of the partners
1.00	15/08/2018	Doina Cristina Duma (INFN)	Version ready for the internal review
1.01	03/09/2019	Giuseppe La Rocca (EGI Foundation)	Internal Review
1.02	17/09/2018	Nuno Fereira (SURFsara)	Internal Review
1.03	24/09/2018	Doina Cristina Duma (INFN), Alessandro Costantini (INFN)	Final version addressing reviewers recommendations
1.04	28/09/2018	Volker Beckman (CNRS) & Mark Thorley (UKRI)	Minor typographic edits prior to submission.

***Disclaimer**:* The content of the document herein is the sole responsibility of the publishers and it does not necessarily represent the views expressed by the European Commission or its services.

While the information contained in the document is believed to be accurate, the author(s) or any other participant in the EOSCpilot Consortium make no warranty of any kind with regard to this material including, but not limited to the implied warranties of merchantability and fitness for a particular purpose.

Neither the EOSCpilot Consortium nor any of its members, their officers, employees or agents shall be responsible or liable in negligence or otherwise howsoever in respect of any inaccuracy or omission herein.

Without derogating from the generality of the foregoing neither the EOSCpilot Consortium nor any of its members, their officers, employees or agents shall be liable for any direct or indirect or consequential loss or damage caused by or arising from any information advice or inaccuracy or omission herein.

TABLE OF CONTENT

EXECUTIVE SUMMARY 3

INTRODUCTION 4

1. Evolution of interoperability aspects in the EOSC context 5

1.1. E-Infrastructures Interoperability 7

1.1.1. Pilot for Connecting Computing Centres 7

Network requirement for PiCO2 8

1.1.2. WLCG/AARCv2/EOSCpilot AAI pilot for distributed authorization and authentication 8

1.1.3. Grid-Cloud interoperability demonstrator for HEP community 11

1.2. Data Level interoperability 12

2. Science demonstrators & Interoperability aspects 15

2.1. First set of Science Demonstrators 15

2.1.1. ENVRI Radiative Forcing Integration 15

2.1.2. Pan-Cancer Analysis in EOSC 16

2.1.3. Research with Photons & Neutrons 16

2.1.4. Collaborative semantic enrichment of text - based datasets (TEXTCROWD) 18

2.1.5. WLCG Open Science Demonstrator - Data Preservation and Re-Use through Open Data Portal (DPHEP) 19

2.2. Second Set of Science Demonstrators 19

2.2.1. HPCaaS for Fusion (PROMINENCE) 20

2.2.2. EGA Life Science Datasets Leveraging EOSC 20

2.2.3. Virtual Earthquake and Computational Earth Science e-science environment in Europe (EPOS/VERCE) 21

2.2.4. CryoEM workflows 21

2.2.5. Astronomy Open Science Cloud access to LOFAR data 22

2.3. Third Set of Science Demonstrators 22

2.3.1. Frictionless Data Exchange Across Research Data, Software and Scientific Paper Repositories 22

2.3.2. Mining a large image repository to extract new biological knowledge about human gene function 23

2.3.3. VisIVO: Data Knowledge Visual Analytics Framework for Astrophysics 23

2.3.4. Switching on the EOSC for Reproducible Computational Hydrology by FAIR-ifying eWaterCycle and SWITCH-ON 23

2.3.5. VisualMedia: a service for sharing and visualizing visual media files on the web 23

3. Conclusions and Future Work 25

Annex A.1 - GLOSSARY 26

Annex A.2 - EOSC-hub – Technology readiness levels 28

Annex A.3 - EOSC SDs testbed commonalities – v1 29

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. Broadly speaking, the overall plans of the selected EOSCPilot Science Demonstrators are to showcase how EOSC services, and resources, can be combined each other to produce, as result, a pilot services as stated in their submitted work-plans. The WP6 overall plan is to provide a more general framework for interoperability between data and services. To cope with these goals, WP6 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:
    - It that provides the research infrastructure and domain expert’s 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 data stores, 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 and also by the long tail of science.
    - In the Infrastructure Interoperability track the provider’s 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 objective of this deliverable is to present an updated panel of the requirements, not only on interoperability testbeds but in general on the interoperability, at the infrastructural/networking level and of the various services envisaged to be part of the EOSC environment, the new requirements and how initial ones were or not addressed, as reported by the partners participating to the WP6 activities, the Science Demonstrators representatives and their shepherds.

The final results of the validation of the interoperability testbeds, of the services requested and deployed on those testbeds, the architecture defined, in deliverables D5.1[1], D5.4[2], D6.2[3] and in the future D6.8[4], and implemented, will be reported in the final WP6 deliverable, the D6.10 “Final Interoperability Testbed report”.

[1] D5.1: Initial EOSC Service Architecture the high-level EOSC (reference) service architecture.

[2] D5.4: Final EOSC Service Architecture.

[3] D6.2: EOSC architecture design and validation procedure.

[4] D6.8: Final EOSC architecture.

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 infrastructures 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.

This document is providing an update on the (new) requirements on the interoperability testbeds, and on the services deployed, from all the players involved in the project activity: Science Demonstrators, Data Interoperability demonstrators and research e-infrastructures. In some cases, instead of requirements, recommendations that emerged as a result of the activity carried out are reported.

The document is structured in two main sections. The first one will look into the *Interoperability in the EOSC context*, with subsections delving into the interoperability requirements at both the e-infrastructure and data research e-infrastructures levels. The second one regarding the Science Demonstrators, with three subsections grouping the demonstrators according with their start in relation to the project period.

The document closes with a summary of the requirements, comments and suggestions on the infrastructures and services interoperability improvements and on the next steps, including their validation from the point of view of maturity level of solutions for what regards TRL, openness, scalability, user community adoption and sustainability.

1. EVOLUTION OF INTEROPERABILITY ASPECTS IN THE EOSC CONTEXT

The first report that gathered the initial requirements, the D6.4 – Initial requirements of the Interoperability Testbeds, started by analysing the EOSC environment, the various reports that defined the European Commission vision of a large infrastructure to support and develop open science and open innovation in Europe and beyond:

- The blueprint "European Cloud Initiative - Building a competitive data and knowledge economy in Europe"[5] – presenting the aims to develop a trusted, open environment for the scientific community for storing, sharing and re- using scientific data and results, the **European Open Science Cloud**, underpinned by the deployment of a super-computing capacity, the fast connectivity and the high-capacity cloud solutions they need via a **European Data Infrastructure.**
- The new **European Interoperability Framework** (EIF) as part of its Communication (COM(2017)134)[6], giving specific guidance on how to set up interoperable digital public services.
- The first report and recommendations of the Commission HLEG on EOSC "**Realising the European Open Science Cloud**"[7] – containing implementation recommendations regarding EOSC, one of them been the development of a concrete plan for the architecture of data interoperability of the EOSC.

Subsequently, the following EOSC related reports regarding interoperability aspects have been released containing recommendations for data, services and infrastructures:

- the **EOSC Declaration**[8], followed by the **EOSC Declaration Action list**[9] - guiding the implementation of EOSC, mentions the interoperability aspects in the "*Data culture and FAIR data*" section, containing *recommendations* regarding the FAIR principles, that should apply not only to research data but also to data-related algorithms, tools, workflows, protocols, services and other kinds of digital research objects; *recommendations* regarding the Technical Implementation – describing the need of interoperable tools in order to make data FAIR – like Citation System, Common Catalogues, FAIR tools and services; in the "*Research data services and architecture*" section, talking about the EOSC architecture as a data infrastructure commons, federating existing resources across national data centres, European e-infrastructures and research infrastructures, the service provision and deployment, the co-ordination and progressive federation of open data infrastructures developed in specific thematic areas (e.g. health, environment, food, marine, social sciences, transport), the *recommendation* to implement a common reference scheme to ensure FAIR data uptake and compliance by national and European data providers in all disciplines
- the **Implementation Roadmap for the European Science Cloud** (Staff Working Document SWD(2018) 83)[10] – mentioning the In the Mid-Term Review of the implementation of the Digital Single Market Strategy[11], where the Commission confirmed its intention to come forward with an implementation Roadmap for the European Open Science Cloud, and referenced the *European Interoperability Framework (EIF)* and the *INSPIRE Directive*, mentioning *"Use of the ICT standards referenced in a European catalogue*[12] *would scale up the size of the market for digital products and*

[5] https://ec.europa.eu/digital-single-market/en/news/communication-european-cloud-initiative-building-competitive-data-and-knowledge-economy-europe

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

[7] http://ec.europa.eu/research/openscience/pdf/realising_the_european_open_science_cloud_2016.pdf#view=fit&pagemode=none

[8] http://ec.europa.eu/research/openscience/pdf/eosc_declaration.pdf#view=fit&pagemode=none

[9] http://ec.europa.eu/research/openscience/pdf/eosc_declaration-action_list.pdf#view=fit&pagemode=none

[10] https://ec.europa.eu/research/openscience/pdf/swd_2018_83_f1_staff_working_paper_en.pdf

[11] https://ec.europa.eu/transparency/regdoc/rep/1/2017/EN/COM-2017-228-F1-EN-MAIN-PART-1.PDF

[12] https://joinup.ec.europa.eu/solutions

services”. It is there (European Catalogue) where we find information on the **European Interoperability Reference Architecture (EIRA)**[13], part of the ISA^2 programme, Action 2016.32[14] - that introduces this reference architecture to guide public administrations in their work to provide interoperable European public services to businesses citizens. In the context of EOSCpilot, and WP6 in particular, the needs and recommendations contained in this action can be, and are, extrapolated to guide also our activities, the evaluation of services envisaged. For example:

- EIRA© is a reference architecture for delivering ***interoperable*** digital public services across borders and sectors. It defines the required capabilities for promoting interoperability as a set of architecture building blocks (ABBs). EIRA© has four main characteristics:
  - **Common terminology to achieve a minimum level of coordination**: *It provides a set of well-defined ABBs that provide a minimal common understanding of the most important building blocks needed to build interoperable public services*
    - this aspect in the EOSCpilot is followed by the WP5 work on ***“shared EOSC Terminology/Glossary”***
  - **Reference architecture for delivering digital public services**: *It offers a framework to categorise (re)usable solution building blocks (SBBs) of an e-Government solution. It allows portfolio managers to rationalise, manage and document their portfolio of solutions*.
    - This aspect is addressed in the EOSC services and infrastructure architecture design developed by the project – described in the D5.1 (“Initial EOSC Service Architecture the high-level EOSC (reference) service architecture”) and D6.2 (“EOSC architecture design and validation procedure”)
  - **Technology- and product-neutral and a service-oriented architecture (SOA) style**: The EIRA© adopts a service-oriented architecture style and promotes ArchiMate® as a modelling notation.
    - This aspect is taken in consideration by the *WP5 – Services, and WP6 – Interoperability*, in particular by the T5.4 (Service pilots) and T6.3 (Interoperability pilots) in their activities of identifying the services to meet Science Demonstrators needs, and to validate them.
  - **Alignment with EIF and TOGAF®:** The EIRA© is aligned with the European Interoperability Framework (EIF) and complies with the context given in the European Interoperability Framework - Implementation Strategy (EIF-IS) . The views of the EIRA© correspond to the interoperability levels in the EIF: legal, organisational, semantic and technical interoperability. Within TOGAF®[15] and the Enterprise Architecture Continuum[16], EIRA© focuses on the architecture continuum[17]. It re-uses terminology and paradigms from TOGAF® such as architecture patterns, building blocks and views.
    - This aspect will be part of the validation of the final EOSC architecture, reflected in D5.4 (“Final EOSC architecture”), D6.8 (“Final EOSC architecture”), and D6.10 (“Final Interoperability WP6 Testbed report”).

All these documents show the importance is the interoperability aspects, how it is addressed in the context of other European Commission initiatives, and how the EOSCpilot activities are carried along the same line.
In the context of the activities of the interoperability validation of the solutions deployed on the EOSCpilot testbeds we will also try to use one of the outcomes of the ISA^2 programme, Action 2016.33[18], “

[13] https://joinup.ec.europa.eu/collection/european-interoperability-reference-architecture-eira/about

[14] https://ec.europa.eu/isa2/isa2_en

[15] http://www.opengroup.org/TOGAF-9.2-Overview

[16] https://www.visual-paradigm.com/guide/togaf/togaf-91-framework/

[17] https://www.gov.pl/documents/31305/182642/d1_4_raul_abril_-_the_european_interoperability_reference_architecture.pdf/637ecb24-9666-8e03-b409-50917a18ce51

[18] https://ec.europa.eu/isa2/actions/evaluating-and-rationalising-ict-public-administrations_en

Assessment of trans-European systems supporting EU policies of the Interoperability solutions and common frameworks" – the ***Interoperability Quick Assessment Toolkit (IQAT©)***[19], who's objective is to allow Solution Owners to assess the Potential Interoperability of their software solutions supporting Public Services. In particular, we'll use the **Interoperability Quick Assessment Excel Tool**[20], an Excel tool used to calculate solutions interoperability maturity, and the **Guidelines for Solution Owners**[21], a short document that includes the **methodology** that describes the high-level steps **to be followed by Solution Owners for assessing the potential interoperability** of a solution.

1.1. E-Infrastructures Interoperability

The aim of EOSC is to provide access to European e-infrastructures through interfaces that will allow seamless usage and will enable connectivity across disciplines and borders. This system of systems will rely on a set of existing open systems, including e-Infrastructures, Research Infrastructures (RI) and providers (private and public). This set of systems should be considered as the backbone of the EOSC. The "e-Infrastructure Gap Analysis" [22] (D6.1) identifies the main barriers to exploiting and using existing e-infrastructures and distributed resources. It also gives axes to build bridges regarding the identified gaps. The "EOSC architecture design and validation procedure" deliverable[23] (D6.2) defines:

- how existing e-infrastructures, Research Infrastructures (RI) and providers (private and public) can technically make their infrastructure available within EOSC, including types of computing and storage resources.
- how to ensure the interoperability of the e-Infrastructures, Research Infrastructures (RI) and providers (private and public).

Deliverable D6.2 merely sets the stage for the design of an interoperable e-infrastructure architecture, by defining some of the main principles of how e-infrastructures should interoperate within EOSC. We summarise them for completeness, as they are in the scope of the present report, regarding too the interoperability aspects of the project pilots. More details are available in the following deliverables.

- Availability and reliability of e-infrastructures involved in EOSC, and of the services provided by them.
- AAI interoperability – develop a single referent for all the e-infrastructures or cross-identification and authorisation between existing services
- Network reliability and capability, standard infrastructure connections to network
- Standard and new open services interfaces
- Common vocabulary, global services, catalogue

Work is on-going to satisfy these requirements, e.g. a dictionary with common definitions of technical aspects of infrastructure, application of standards, REST APIs, usage of the WebDAV in the development of EOSC services, usage of service catalogues, like e-InfraCentral and EOSC-hub. This also includes the validation and evaluation of the interoperability of the existing services, through the pilots set up as part of T6.3 activities and science demonstrators.

1.1.1. Pilot for Connecting Computing Centres

[19] https://joinup.ec.europa.eu/solution/interoperability-quick-assessment-toolkit/about

[20] https://joinup.ec.europa.eu/rdf_entity/http_e_f_fdata_ceuropa_ceu_fw21_f234ba476_b2523_b4c6b_b9d76_b4f04b9f6c36d

[21] https://joinup.ec.europa.eu/rdf_entity/http_e_f_fdata_ceuropa_ceu_fw21_fd733fa1b_b62a0_b40ad_b9661_b011c684290f1

[22] https://repository.eoscpilot.eu/index.php/apps/files/?dir=/Deliverables&fileid=2987#

[23] https://repository.eoscpilot.eu/index.php/apps/files/?dir=/Deliverables&fileid=2987#

PiCO2 (**Pi**lot for **CO**nnecting **CO**mputing centres) is one of the first interoperability pilots between generic, community agnostic, infrastructures, especially Tier-1 (National HPC/HTC centres), and Tier-2 (HPC/HTC regional centres). Its main objective is the Automation of frequent, community agnostic, data flows (many large files) between Tier-1 HPC/HTC and Tier-2 infrastructures.

Network requirement for PiCO2

Since cloud infrastructures and users are distributed all over Europe, network connectivity between them is therefore a pillar on which EOSC has to be built. The stability, robustness and performance of this network infrastructure are crucial for EOSC to succeed. The network requirements can be summarized as:

- Connectivity between the academic data centres themselves, academic data centres and EOSC users, and academic data centres and commercial cloud providers depends on:
    - the access capacity deployed to connect the sites, and the network paths of the data transfer between the EOSC sites
    - advanced connectivity services usage
        - VPN usage can improve data exchange performance (see HEP and HPC community)
        - VPN improves security data exchange
- Network reliability
    - A reliable and resilient network underpinning these services is an absolute requirement
    - Network service monitoring
        - Network must be able to provide a monitoring solution in order to be able to identify if the problem is due to the network and where.

These requirements do not change from the first requirement expression. One lesson learnt from the L3VPN deployment is that the site teams need to be trained about VPN local connection configuration and, even if the duration in time of the connection is not long, the local teams need to dedicate appropriate manpower in order to establish a proper VPN connection at site level.

1.1.2. WLCG/AARCv2/EOSCpilot AAI pilot for distributed authorization and authentication

As presented in the deliverables D6.4 and D6.5, the EOSCpilot and AARC projects[24] started a collaboration activity in the field of authorization and authentication, policies and recommendations regarding their design, that took shape, in the scope of the WP6 activities, under the form of an AAI interoperability demonstrator setup as part of the AARC pilots Task 1: ***Pilots with research communities based on use cases provided***[25] - **the WLCG use case,** regarding the "*Implementation of IdP/SP Proxy, mainly to provide Token Translation Services to allow end users to login without the need of manually managing X.509 certificates*"[26]. A team of people was formed, under the WLCG coordination, to deal with the various activities – the WLCG Authorization WorkingGroup[27] (WG), motivated by:

- Evolving Identity Landscape
    - User-owned x509 certificates -> Federated Identities
    - Federated Identities linkage with existing VOMS authorizations not supported
    - Maintaining assurance and identity vetting for federated users not supported
- Central User Blocking
    - Retirement of glexec removes blocking capability (& traceability)
    - VO-level blocking not a realistic sanction
- Data Protection
    - Tightening of data protection (GDPR) requires fine-grained user level access control

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

[25] https://wiki.geant.org/display/AARC/AARC+Pilots

[26] https://wiki.geant.org/display/AARC/WLCG+Worldwide+LHC+Computing+Grid

[27] https://twiki.cern.ch/twiki/bin/view/LCG/WLCGAuthorizationWG

After an initial requirements gathering[28], and analysis of how existing solutions functionalities match the requirements[29], two main activities started:

1. Design and testing of a WLCG Membership Management and Token Translation service, facilitated by pilot projects with the support of AARC (AAI Pilot Projects)
2. Definition of a token based authorization schema for downstream WLCG services and token issuers (JWT)

We list bellow a short summary of the identified ***initial requirements*** (more details available on the WG wiki):

- VO Membership Management
    - Attributes: VO ID, ID of credential, Name, Email, Authorization
    - Support multiple federated credentials & their linkage
    - Active role selection
    - Token management achievable by the standard user
- Service Requirements
    - Attributes: Authorization plus traceability or Groups/Roles
        - Splitting things more into "identity based" and "authorization based" approaches. Third-party services (e.g. storage) would primarily consume the latter.
    - Ease of implementation
    - Use standard approaches
    - Token integrity and validity verifiable
        - Without connecting to the issuer
    - For non-web, users should not have to manage identities in addition to their login session
- General
    - Support for fine grained suspension
    - Smooth transition from current X509-based to token-based AAI

The pilot's main goal is to demonstrate how WLCG services will be accessed by users authenticated and authorized via SAML, using federated credentials. In particular, registration of users in VOMS by means of SAML federated credentials is a key objective. Another relevant goal is to provide command line access to WLCG services making use of non-X.509 credentials (SAML, OIDC). The pilot will demonstrate the required functionality by implementing 2 options to address the requirements: INDIGO IAM, pilot supported by EOSCpilot and AARC projects, and EGI Checkin service + COmanage.

The WG has monthly meetings, continuously evaluating old and new requirements, the status of the implementation of missing functionalities, and the status of the two pilots. Two meetings were of particular importance, highlighting the status of the activities and the updated list of requirements:

- The "Joint WLCG and HSF workshop"[30], in Naples, Italy – containing the presentation of the AAI pilots status[31]
- The **pre-GDB - Authz Working Group**[32] **and GDB**[33]**,** at CERN – with a presentation on the status of the IAM pilot[34] and the sign-off of the updated list of ***requirements document***[35]

Together with the requirements of the authorization solutions, the goals of the WG were updated as follows:

[28] https://twiki.cern.ch/twiki/pub/LCG/WLCGAuthorizationWG/AuthZ_pre-GDB_Requirements.pdf

[29] https://docs.google.com/spreadsheets/d/1mC2U2H12RDHsOtk1OHQM3_HVbbflHfj-Y1Fv0yW_0KA/edit#gid=0

[30] https://indico.cern.ch/event/658060/

[31] https://indico.cern.ch/event/658060/contributions/2890286/attachments/1622544/2582430/AuthZ-WG-180328.pdf

[32] https://indico.cern.ch/event/651343/

[33] https://indico.cern.ch/event/651355/

[34] https://indico.cern.ch/event/651343/contributions/3045233/attachments/1687971/2715102/IAM-PreGDB-AuthZ-170718.pdf

[35] https://docs.google.com/document/d/1hnsPWf9C7ODVXZ7JehsSEiEsQwf5UmqLfTwVDhuqHzk/edit#heading=h.9jpjmmywsm

- Evolving Identity Landscape
    - Current grid middleware does not support federated identities
    - How can we shield users from the complexities of X.509 certificate management?
    - Token-based authorization widely adopted in commercial services and increasingly by R&E Infrastructures
- Data Protection
    - Tightening of data protection (GDPR) requires fine-grained user level access control, certain provisioning practices may need to be adjusted

Some of the ***new requirements*** of the AAI solutions are listed bellow, while the whole updated list is present in the ***"requirements document"***[35] referenced above:

- VO Membership Management
    - VO Membership Management Overview
        - Periodic membership renewal should be supported, as defined by policy
        - Periodic credential verification should be supported, as defined by policy
        - Periodic AUP Signing should be supported, as defined by policy, including:
            a) user suspension upon failure to sign
            b) controlled delegation and consent
        - Integration of additional trusted data sources must be supported (e.g. Institute Affiliation Expiry from CERN HR DB)
        - VO managers must be able to overwrite the information from integrated data sources
        - For LHC VOs, the option to leverage CERN Infrastructure must be supported, such as membership expiration based on home institute affiliation
        - VO management should provide a process that, given a token or identifier, can resolve user attributes. This should be restricted to VO and Infrastructure level use.
    - User Credential types (i.e. Authentication, not server-server auth tokens)
        - Flexibility to enable a trusted IdP where appropriate, e.g. CERN SSO
        - All credentials must be able to satisfy minimum requirements for Assurance; VOs should be able to select which specific Identity Providers they enable
    - Usability
        - Documentation should be provided and maintained
        - Bulk actions for VO Managers and Users should be enabled where appropriate
- Service Providers
    - Service Requirements Overview
        - All Tokens
            a) Tokens must be supported on web and non-web
            b) Must be able to determine the token issuer
        - Identity/Access Tokens
            a) Must be short lived
            b) Should be possible to transparently provision the user with the required token
            c) Users must be able to allow services constrained delegation
            d) Should include sufficient information to allow decentralised verification
        - Refresh Tokens
            a) Must be revocable
    - Required User Attributes (for the services to operate)
        - At least one of
            a) Authorisation attributes, i.e. Roles/Groups
            b) Capabilities
        - Additional identity information should be supported, as required by policy

- General
  - Operational requirements
    - Access token lifetime < operational response minimum time, as defined by policy and in line with standard recommendations
    - Suspension
      a) Blocking of individual VO Users across all services/subset must be possible by a VO and/or the Infrastructure (i.e. Security Team) within a timeframe defined by policy
      b) Sites/Services must be able to block (potentially opaque) VO Users locally, and inform relevant parties (e.g. Infrastructure security or VO management) as defined by policy
  - Change Management
    - A smooth transition path should be defined, including backwards compatibility for a necessary timeframe

In the following months the activity of the improvement of the INDIGO IAM solution to meet all the foreseen requirements will continue, together with the deployment and validation of the various components.

1.1.3. Grid-Cloud interoperability demonstrator for HEP community

Dynamic On Demand Analysis Service (DODAS) is a Platform as a Service tool built combining several solutions and products developed by the **INDIGO-DataCloud** H2020 project. It has been extensively tested on a dedicated interoperability testbed under the umbrella of the **EOSCpilot** project, during the first year of the project.

Although originally designed for the Compact Muon Solenoid (CMS)[36] Experiment at LHC, DODAS has been quickly adopted by the Alpha Magnetic Spectrometer (AMS) astroparticle physics experiment mounted on the ISS as a solution to exploit opportunistic computing, nowadays an extremely important topic for research domains where computing needs constantly increase. Given its flexibility and efficiency, DODAS was selected as one of the **Thematic Services that will provide multi-disciplinary solutions** in the **EOSC-hub** project. An integration and management system of the European Open Science Cloud starting in January 2018.

During the integration pilot the usage of any cloud (both public and private) to seamlessly integrate existing Grid computing model of CMS[37] was demonstrated.

Overall, integration has been successful and much experience has been gained resulting in improved understanding of weaknesses and aspects to improve and to optimise.

Weaknesses, and aspects to be improved include:

- **Federation**: federated access to underlying IaaS is a key. So far we've experienced several issues. Frequently we had issues with the IaaS provider already using OpenID Connect Authorization Server and thus unable to federate additional services. We adopted ESACO[38] solution to solve such a problem. **It would be crucial to have it as a EOSC provided service**.
- **Accounting**: an **APEL**[39] **based solution** for non-proprietary IaaSes would be extremely important in the EOSC landscape. A scenario where, as example, a commercial cloud is used, would benefit of such functionality for counting the overall HEPSpec[40].

[36] https://home.cern/about/experiments/cms

[37] https://twiki.cern.ch/twiki/bin/view/CMSPublic/WorkBookComputingModel

[38] https://github.com/indigo-iam/esaco

[39] https://wiki.egi.eu/wiki/APEL

[40] https://wiki.egi.eu/wiki/FAQ_HEP_SPEC06

- **Transparent Data Access**: so far the only scalable solution we can use is XrootD[41]. However, this might not fit all possible use cases. A **more generic solution** would be a big plus.
- **Resource monitoring**: we didn't find a common solution for monitoring cloud resources. Although we implemented our own we are convinced that **a common strategy** would be extremely valuable.
- **PaaS Orchestration**: Although the current INDIGO PaaS Orchestrator has been fully integrated and show enormous advantages while dealing with multiple IaaSes, there is room for **improvement** both **in the interface and in the management of IaaS ranking**.

1.2. Data Level interoperability

The interoperability WP6 main objectives are to develop and demonstrate the interoperability requirements between e-Infrastructures, domain research infrastructures and other service providers needed in EOSC. One of the two tracks into which interoperability is mapped is 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 objective of the EOSCpilot data interoperability task (6.2) is to demonstrate how to ensure availability of scientific data to users and services through an open cloud infrastructure.

After the delivery of the first draft of the strategy and recommendations[42], four data interoperability demonstrators have been proposed to test components of the strategy:

- Evaluation of the EDMI (EOSC Datasets Minimum Information) metadata guidelines to find and access datasets
- Discovery of compliant data resources and metadata catalogues
- Research schemas for exposing dataset metadata
- Description and guidelines per metadata property

The status of the activities on these four data demonstrators was presented both in D6.5 – "Interim Interoperability Testbed report", and D6.6 – 2nd Report on Data Interoperability. The later contains a detailed description of the recommendations from the demonstrators and feedback from the EOSCpilot partners involved in their activity.
For completeness, in this deliverable we list the most important ***new requirements***/recommendations. For more details, please see the D6.5 & D6.6, available in the EOSCpilot file-repository[43].
The following recommendations are meant to guide the activity of the data demonstrators in the next phase of the project. How these recommendations will be taken in consideration by the science demonstrators will be reflected in the "Final report on Data Interoperability", deliverable D6.9.

- *Exposing EDM properties:*
  - **Recommendation 1:** The minimum properties should not be considered a mandatory set but as an ideal state to facilitate findability and accessibility. EDMI should include a core set of fewer properties easy to comply with. This way we could encourage providers to move from "Core" to "Minimum" and enrich the metadata with "Recommended" properties. "Core" properties could be aligned to the DataCite[44] mandatory properties.

[41] http://xrootd.org/

[42] https://eoscpilot.eu/content/d63-1st-report-data-interoperability-findability-and-interoperability

[43] https://repository.eoscpilot.eu/index.php/apps/files/?dir=/Deliverables&fileid=2987

[44] https://www.datacite.org/

    - **Recommendation 2:** Look for more examples using other competing standards like DATS[45], DataCite, DCAT[46] and other domain specific standards such as CERIF[47] and W3C HCLS,[48] and provide mappings of equivalence to enable users to move between different solutions.
    - **Recommendation 3:** Make a proposal how to expose an EDMI property when a property is missing in an existing standard.
    - **Recommendation 4:** Keep working on the RDA metadata guidelines specially focusing on identifiers, access rights and licenses.
- *Conversion tool to help exposing EDMI with Schema.org*
    - **Recommendation 5:** Consider the Schema.org conversion tool as a way to quickly get adoption and showcase the benefits of Schema.org and EDMI.
    - **Recommendation 6:** Encourage the adoption of Schema.org and compliance to EDMI in EOSC data resources.
- *Discovery of compliant data resources and metadata catalogues*
    - **Recommendation 7:** Make a proposal of how to display compliance based on the data resource model and based on the metadata content.
    - **Recommendation 8:** Think about how to show compliance for data resources. For instance, for each data resource we could select a few datasets that we could evaluate to identify the compliance profile of data resources.
    - **Recommendation 9:** Work with EOSCpilot WP3 on how to monitor compliance with EDMI. EOSCpilot WP3 is working on defining and developing the EOSC Open Science Monitor Framework that could help to evaluate the compliance with guidelines such EDMI.
    - **Recommendation 10:** The minimum set should be seen as a goal to achieve. To be more inclusive and promote EDMI we recommend EDMI to have a subset of properties which could be core (or mandatory). Thus, we could display several levels of compliance (Core, Minimum, Recommended and Optional).
- *Description and guidelines per metadata property*
    - **Recommendation 11**: Engage metadata experts from different communities. Identify the most challenging properties and look for existing projects and communities willing to contribute to define the guidelines (e.g. for identifiers: FREYA, ELIXIR identifiers, RDA identifiers).
    - **Recommendation 12**: Make sure we have a template with the structure and we update the rest of the properties with changes.

The following list consists in recommendations about the strategy of metadata catalogues and datasets for EOSC:

- *Metadata catalogues, data repositories and datasets*
    - **Recommendations:** The final EOSCpilot data interoperability strategy needs to be consistent and reuse terminology defined in the EOSC glossary.
- *Metadata catalogues and datasets in EOSC*
    - **Recommendation 13:** In the final recommendations, it must be made clear that EDMI aims to be a crosswalk guideline, encouraging the use of existing standards to describe datasets

[45] https://biocaddie.org/group/working-group/working-group-3-descriptive-metadata-datasets

[46] https://www.w3.org/TR/vocab-dcat/

[47] https://www.eurocris.org/cerif/main-features-cerif

[48] https://www.w3.org/2001/sw/hcls/

such as: DataCite or DCAT for generic datasets, and CERIF or HCLS for domain specific datasets.

  - **Recommendation 14:** Highlight and make clear the focus and scope of EDMI and show how EDMI complement other minimum information guidelines.

- *Strategy*
  - **Recommendation 15:** Bring together metadata catalogues participating in EOSC (catalogues from e-infrastructures and Research Infrastructures) to agree and shape the strategy proposed by EOSCpilot data interoperability. Build the case to show the ecosystem of dataset metadata catalogues is one of the key blockers to making EOSC work, and persuade funders that its implementation requires active engagement with, and funding for, metadata catalogues.

2. SCIENCE DEMONSTRATORS & INTEROPERABILITY ASPECTS

In a nutshell, 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. They play an essential role as early adopters of EOSC from a range of science areas. Their input is used to drive and prioritize the integration of the EOSC services in a common and homogeneous platform.

In the following sections we present their inputs for what regards their expectations and requirements still to be fulfilled. These inputs have been collected directly by the Science Demonstrators (SDs) scientific contacts or EOSCpilot respective shepherds, or from the periodic reports submitted to the project, when the direct input was missing.

To collect the contributions from the SDs, and better describe revised and final interoperability requirements and testbeds, we prepared a spreadsheet where they could find row-wise the available information for each SD, as extracted from the WP4 reports and deliverables. The SDs representatives, including scientific contacts and respective shepherds, were asked to update the already available information, or integrate the missing ones with a particular focus on the technical implementation, protocols and technologies used. For the sake of completeness, the spreadsheet containing the status of the testbeds and the revised interoperability requirements are reported in Annex A of this deliverable. The final status of the testbeds will be reported in D6.10 – "Final Interoperability Testbed report".

2.1. First set of Science Demonstrators

The EOSCpilot project started with five Science Demonstrators pre-selected from a call in 2016:

- Environmental & Earth Sciences - ENVRI Radiative Forcing Integration to enable comparable data access across multiple research communities by working on data integration and harmonised access.
- 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).
- 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.
- 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.

2.1.1. ENVRI Radiative Forcing Integration

The ERFI Science Demonstrator originally planned 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 data integration framework aimed to make (initially parts of) multi-petabyte climate model data archives hosted at DKRZ and IPSL accessible for EGI/ICOS based on common cloud services. This data interoperability issue has been resolved considering:

- The ***Onedata*** solution, developed in the context of the INDIGO-DataCloud H2020 project, to provide a transparent access to ICOS and ENES datasets.

- A reliable and scalable cloud Infrastructure where the ENES data download and synchronization software (synda) can be installed.

The metadata integration interoperability is still work in progress.

2.1.2. Pan-Cancer Analysis in EOSC

One of the main hurdles in the PanCancer Science Demonstrator has been **the availability** of the (extremely large) **datasets** in the computational resources offered. Not only is the input data large, but for the workflows in this SD it is **required that this data is highly available and that large local caches are able to be constructed to enable the appropriate scale**. This requirement is necessary because the enormous amount of data to be processed would take an unreasonably long time to compute on 'commodity' resources. So the **interoperability of this work, across different cloud environments, is largely dependent on the data delivery model available to it**. Another complicating factor is that the workflow needs to **access this data in a POSIX way**, meaning that it needs to be part of the local filesystem and cannot be accessed through protocols that are regularly used for remote data such as (Grid)FTP, object storage or HPC solutions such as CEPH.

To try and address this issue the SD engaged with the Cyfronet service provider to test their Onedata solution. The Onedata software stack promises to deliver large scale datasets in a high performance way which, depending on the exact setup used, can be accessed as native files with a reasonably similar performance. During this SD, resources were made available by Cyfronet and a data provider deployment was setup on the EMBL EMBASSY cloud; the source of the data. In our testing scenario, a Oneprovider was set up on the receiving end as well with an appropriately sized cache available to it so that it could deliver the remote data in the high performance manner required. Ultimately, the SD partially succeeded in running workflows through this solution, but many issues were encountered which were mostly focused around the scale of the data throughput required. Ultimately, from the results of this phase of the SD, it can only be concluded that this type of **data transparency and delivery still remains a mostly unsolved issue**. It is also important to mention that Onedata is the **only** solution in the wider scientific and HPC community even attempting to deliver this type of data transparency and that **no other alternatives for this scenario exist at the moment**.

While Pan-Cancer is relatively large scale compared to current more commonplace life science workflows, it is clear that most life science researchers will require resources of this scale on a regular basis in the next few years. If an interoperable "system-of-systems" approach would need to work for the EOSC **the data delivery problem must be solved**. Not only will scale be an important factor in this, but also the availability of appropriate authentication and authorisation mechanisms and the associated ownership accounting services to handle the legal and policy requirements surrounding much of this data.

2.1.3. Research with Photons & Neutrons

The Photon Neutron Data Science Demonstrator leverages on the photon-neutron community to improve computing facilities by creating a virtual platform for all users. The Photon-Neutron science demo is based on the concept that:

- Data sets become too large to take home;
- The EOSC forms a platform for analysis and data storage;
- The EOSC allows for easy data sharing;
- Data rates require dedicated central IT infrastructure, way beyond previous requirements;
- A wide variety of scientific users means significant number of data formats and analysis software.

The activity of the SD ended and the final report[49] provides a lot of information in the scope of this deliverable, functional and non-functional requirements, interoperability challenges and possible solutions:

- The containerization of specific 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 identifying concrete **interoperability requirements** which are of importance for a successful integration of the analyses platforms and similar systems in EOSC.
    - Tests were run on local OpenStack infrastructure and on the HPC cluster, in particular in view of automatic provisioning of distributed systems and deployment of applications. The tests have shown that the automation stack applied (Foreman, Puppet, cloud-init and cloud-images) 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, AFS, CVMFS, conventional (NetApp) SMB/NFS-shares and a local/federated Owncloud instance were used. Configurations 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 EOSC**. Work will continue to 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**.
- Regarding the **network access**, inter-network trust and speed and dynamically managed overlay networks for cloud VMs and container deployment in multi-cloud environments, the **Docker Swarm** technology to deploying containers on the local HPC cluster and cloud has been investigated even though they are more 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.
- Practical implementation and scaling of (docker) 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.**
    - For **interoperable trust** between networks and cloud providers, also comparable metrics to control, SLAs and QoS are needed.
- Technical (and political) challenges and issues encountered, and some suggestions to mitigate them:
    - **Licensing -** Substantial part of the applications in the Photon/Neutron science domain is “free to use for academic purposes” but subject to restrictive licensing conditions. To provide services based on such applications requires knowing that the consumer has agreed to licensing terms and is indeed an academic user. This can of course be controlled on a per-service basis. It would however be more **convenient and scalable to provide such attributes in a central/federated way integrated into the EOSC ecosystem.**
        - **Suggestions to mitigate:**

[49] https://repository.eoscpilot.eu/remote.php/webdav/WP4%20-%20Science%20Demonstrators/Reports-of%20Science-Demonstrators/SD1-Prefinal-and-Final-Reports/Photon-Neutron-Science-final-report.docx

        a) Urge EOSC members licensing “free-for-academic-use” software for an EOSC-friendly license agreement.
        b) Aim for agreements with software providers (of not entirely open software) to allow redistribution within the EOSC cloud.
        c) Provide means to clearly separate non-academic from academic service consumption.
    - **Security -** Cloud resources outside the EOSC core like local OpenStack instances are presumably often behind firewalls in “demilitarized zones” (DMZ). **Consuming such resources in EOSC context might require access to specific EOSC entry points opening only very specific ports or routes to well-defined IPs**.
    - **Network -** Access to cloud instances across domains and in particular privileged access would greatly benefit from **inter-network trust and dynamically managed overlay networks for cloud VMs and container deployment in multi-cloud environments**.
    - **Graphical Access -** Many of the Photon/Neutron applications require visual inspection (with GL support) of intermediate analysis steps. This can be solved on a per-instance basis, and is unproblematic for Jupyter based applications. However, **a standardized solution or relay services would be beneficial for users**.
    - **Technical Readiness Level -** The technical readiness level of evolving OpenStack modular environments is very heterogeneous and, in particular embedding container orchestration, introduces strong challenges. **Solutions to such challenges should become readily available at least within the EOSC communities.**
    - **Function as a Service (Faas) -** FaaS seems an appealing way to partition problems and deploy generic/atomic cloud functions in a highly scalable and elastic way. Though such FaaS could be implemented in a domain-specific way, **common applications would benefit from EOSC wide availability of specific functions (e.g. AI based methods)**. This implies secure and reliable namespace administration.
- **Functional requirements:**
    - Performant and secure Docker Registry is a **central requirement**.
    - Networking and interoperability would be supported by EOSC trusted Proxies, IP-ranges for networking.
- **Non-functional requirements**
    - Organization of skills dissemination, admin technical training, documentation down to the issues, bug tracking, code base provision.

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

The aim of the SD was to process openly available texts availing of cloud-based tools for semantic enrichment, linking and annotation, to set up a personal virtual research environment eventually shared with others, to address specific research questions.

The activities done during the SD, previously reported in the D6.4 and D6.5 deliverables, raised the following general recommendations and future requirements:

- **Improve usability of cloud infrastructures** in terms of user interfaces and reusability among components to simplify and speed up installation and deployment processes.
- **Provide advanced controls to monitor** the status of each component and of the **infrastructure as a whole.**
- Interoperability among components would work more efficiently in platforms where **a modular approach, for deploying and running on demand the required components and services,** is supported. D4Science is an example of this **interoperability** platform. It exposed valuable features

to use TEXTCROWD together with OpenNLP[50] and OpenNER[51] web services and the GATE engine[52]. All these features **facilitated the deployment and execution in a controlled environment**. Interoperability on this kind of platform is also guaranteed by the presence of the source documents and the output results on the same environment, paramount for the efficiency and the reusability of the information.

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

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 and generic services can be combined each other in order to meet these needs in a manner that is discipline agnostic, i.e. can be used by others without modification.

Objective: - The high energy physics science demonstrator aimed 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.

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

Limited successful usage was made of the individual services but it was not possible to integrate them in order to achieve the objectives of the SD.

The objective and the goals outlined above were not fully achieved, nor were the "stretch targets" identified in the SD presentation at the kick-off meeting in Amsterdam in January 2017 addressed.

The **main interoperability requirement** raised by the SD is the need to make interoperable the three e-infrastructures services integrated by the EOSCpilot: OpenAIRE (e.g. Zenodo), EGI Engage (e.g. CVMFS – already offered by EGI InSPIRE SA3 (led by CERN)) and a Trustworthy Digital Repository (TDR).

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

[50] https://opennlp.apache.org

[51] http://www.opener-project.eu

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

- **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)

This Science Demonstrator proposes to investigate the **deployment of HPC modelling applications on the cloud**. Within fusion, as the models increase in granularity (both spatially and temporally) the demands on local computational resources is increasing at a faster rate than can be accommodated. With some of these models taking several weeks to run, shorter runs, which may take only a few hours with fewer resources, are often locked out for extended periods of time with traditional batch systems. To overcome this, we propose to investigate whether these types of high-performance computing jobs can be run on cloud environments, with automated busting from local resources onto those provided by the EGI FedCloud and making results of model runs accessible to users on completion.
At month 8 the SD had implemented a basic setup of the service. Test applications (using containerized GEANT-4 code) and 'production like' workflows using Serpent-2 have been run to demonstrate the feasibility and performance of running HPC codes on generic (i.e. non-HPC optimised) cloud computing clusters. Indeed, some of the users have been highly impressed with the scalability and ease of use and would like to make future use of the infrastructure developed. During the setup of the service, several issues required the support of the tool/service providers as already documented in D6.5. In order to gain wider adoption of the service by the fusion community there are a few issues which would **need to be addressed**, as outlined below:

- **AAI** - The need to have X509 certificate to access EGI FedCloud resources is a significant limitation. Currently the fusion community is developing its own portal which we would hope to integrate to at least one of the existing EOSC AAI services, but until then there is no way a general fusion user can obtain a certificate.
- **Resource scale-out** - We have been running with relatively small resource requirements. To meet the needs of a production service, rather than SLAs or OLAs, what we really require is just access to larger clusters. Small production jobs, which fit into the memory footprint available for the cores, would need to use between 16 and 128 CPU cores. Typically, these type of jobs are run concurrently. For this use case, we have no need for high availability or service monitoring as it represents an opportunistic use of the resources provided.
- **Simpler service integration** - It has often been difficult, and in some cases impossible, to integrate services provided with other tools such as Indigo Orchestrator. More work needs to be done to ensure the core and support services are interoperable.

While it is anticipated that these codes will not run as efficiently, the lack of queuing behind more computationally expensive jobs is anticipated to mean science from these can be produced at a significantly higher rate. It is also important to note that not all MPI/Open-MP applications are the same; some only communicate infrequently between the processes (in some cases, only once at the end of the run), while others may need to interact after each time step. The latter are expected to be worse more adversely affected by the lack of high speed interconnects, but should still be able to run efficiently. Finally, during this demonstrator we limited the testing to applications using at most less than 100 computing cores due to the current availability of resources and the need to ensure that the resources used by the SD were freed up for other users.

2.2.2. EGA Life Science Datasets Leveraging EOSC

This Science Demonstrator provided a 'remastering' of existing datasets, attempting to recreate the original pipeline using the original versions, as well as updated versions, of the tools. This results in new, potentially more immediately useful datasets that will also have more derived metadata, which can be indexed. The **main interoperability challenges** with this work has been to **synchronize the generated metadata with a (preferably EOSC supported) metadata indexing service**. This would increase the FAIRness of the data as it would be more easily findable and more reusable. Discussions have been ongoing with the **B2FIND repository** to match the metadata output with their indexing service.

Other **interoperability concerns** were encountered when trying to use the original tools of the use-case example of this SD. **Most tools were unable to be run in modern environments** and in some cases were unable to be found in the older version. In this case, nearest working versions had to be used. To mitigate these concerns for the future, all tools were **containerised** meaning that they were packaged with all their requirements to **make them portable** across systems and hopefully into the future.

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

During the project lifetime the Science Demonstrator has developed a set of additional functionalities to address the AAI and cloud integration interoperability as reported below:

- **Cloud Integration interoperability:** According to the original work-plan, one of the main objective of this Science Demonstrator was to migrate part of the scientific workflow running on HPC resources on cloud resources provided by the EGI Federation. To address this issue, a dedicated Virtual Appliance (VA) to make the VERCE portal interoperable with the EGI FedCloud resources has been updated and registered in the EGI Cloud Marketplace – The EGI AppDB[53]. During the EOSCpilot project, this VA has been further customized by the ViSIVO Science Demonstrator, adding community experiment software, to implement the scientific work-plan.
- **AAI Interoperability:** To allow members of the community to access the cloud-based resources of the EGI Federation, the VERCE portal framework has further extended its AAI framework in order to enable single sign-on access based on the AARC BluePrint architecture. To address the AAI interoperability, the OIDC module developed by the INFN for Science Gateways based on Liferay technology has been used. This best practice has been also shared with the team of the ViSIVO Science Demonstrator since also their portal is based on the same technology.

Remaining points to be addressed, the **Data Dissemination and FAIRness:**

- The VERCE platform supports the production and management of metadata and provenance information describing the generated datasets. This information can be delivered in PROV[54] compliant representation with link to community vocabularies and ontologies. In order to achieve the data interoperability within the EOSC network, requirements and technical solution fostering the linkage, acquisition and dissemination of provenance information within EOSC data-catalogue should be defined. Such specification will be extremely relevant for the support of the FAIRness of the products beyond the VERCE pilot. Approaches on metadata and data provenance management patterns should be conducted in cooperation with the related activities of the RDA Provenance Patterns group. On this topic the group has already produced a number of patterns, which represent a starting point to evaluate current approaches and future directions (e.g. https://patterns.promsns.org/pattern/12).

2.2.4. CryoEM workflows

The aim of this Science Demonstrator is to **add FAIR principles to the Electron Microscopy data**, especially at the central facilities where the data is generated. As new data is acquired, online image processing is

[53] https://appdb.egi.eu/store/software/visivoserver
[54] https://www.w3.org/TR/prov-overview/

started. We have extended the workflow framework to generate a JSON file that can be submitted to the Electron Microscopy Data Base (EMDB-EMPIAR), the international public database for this kind of data. We have contacted them and agreed with them a validation protocol and programmatic submission of the JSON file. Additionally, we have developed a JavaScript viewer that can be integrated in the EMPIAR web page if the user submits the JSON file. Finally, we have explored the use of Common Workflow Language for reporting the image-processing pipeline. However, this choice should be further explored, especially in a community effort to define common tags and meaning.

2.2.5. Astronomy Open Science Cloud access to LOFAR data

The goal of the Science Demonstrator is to empower scientists, who are not necessarily domain experts, to **process radio astronomy data** from current and future telescopes **on EOSC computational resources** and to **increase compliance to FAIR principles for data and service disclosure**. For the EOSCpilot project we focus on LOFAR and will implement a small number of representative processing pipelines to use the Common Workflow Language (CWL) standard and be deployable as Singularity containers. In this field of research, one of the main challenges is the size of datasets that are being processed. We therefore **require proximity of processing resources to the main storage facilities**.

The main objectives and activities until now were concentrated on:

- Started with implementation of CWL based pipelines;
- Investigate applicable tools for FAIR data access;
- Deploy on HPC & HTC systems co-located with archive storage.

And from there the **preliminary requirements (functional and non-functional)** raised are:

- Support for containerized software deployment;
- IO/throughput performance;
- CWL Workflow execution and monitoring (for End Users and Service Enabling organizations);
- Delegation of X509 certificates (move away from user owned certificates);
- Scalability (IO & data volume);
- Usability of (grid) infrastructure;
- User and Service Enabler support by infrastructure providers.

2.3. Third Set of Science Demonstrators

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

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

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

The Science Demonstrator pilots a demonstrator service for fast and highly scalable exchange of data across repositories storing research datasets, manuscripts and scientific software.

Its **interoperability demand** would be for a cross-disciplinary network of repositories to efficiently and reliably exchange scholarly communication; highly scalable exchange of data across repositories, working on the design and implementation of a scalable client for accessing metadata and full texts. Status and first achievement were presented in the D6.5 deliverable. No other interoperability requirements were raised.

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

For the work of this Science Demonstrator, large amounts of image data and metadata were extracted from the IDR[55] using the API functionality. Some issues were encountered when the images could only be requested in a lossy compression format whereas the workflow used was optimised to work with raw uncompressed images. After some image wrangling it was possible to use the compressed images, which were shown to be correlated in the majority of cases and thus deemed sufficient for this phase.

The workflows used in this Science Demonstrator were developed in a more traditional cluster environment and for their move to the EMBASSY Cloud this environment was emulated using an existing LSF-like cloud deployment. However, deploying this system on demand and adjusting it to the specific needs of this SD was time consuming and the SD expressed concern at the time taken to get up and running with the actual analysis. **A more turnkey solution for research coming from a cluster environment** could help to smooth this bump in the adoption of a cloud system.

2.3.3. VisIVO: Data Knowledge Visual Analytics Framework for Astrophysics

The approach proposed by the Science Demonstrator is the **integration** in the EOSCpilot e-infrastructure **of a visual analytics environment** based on VisIVO (Visualization Interface for the Virtual Observatory) and its module VLVA (ViaLactea Visual Analytics). Status and first achievement were presented in the D6.5 deliverable. During the project lifetime the Science Demonstrator re-used/adapted the solutions adopted by the EPOS/VERCE SD to address the AAI and cloud integration interoperability as reported before. No additional interoperability requirements were raised.

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

The goal of this Science Demonstrator will be to unify various modelling workflows that are the current state-of-the-art in hydrology. There are two main interoperability concerns:

- **Interoperability of datasets**, which are currently published under different licences as well as different formats and distributed by different services. One of the main challenges of this SD will be to unify these various sources so that they can be used by multiple workflows.
- **Interoperability of workflows**. At the moment many models are presented in different formats; this SD will attempt to make these workflows useable under one system through a combination of containerisation and format harmonisation.

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

The goal of this Science Demonstrator is to **further extend the functionalities of the Visual Media Service**[56] and to open it to a larger community. The focus of the Visual Media Service is to support the

[55] https://idr.openmicroscopy.org/about/

[56] http://visual.ariadne-infrastructure.eu/

easy/automatic publication on the web of visual media files (3D models, hi-res 2D images and RTI images) and to provide web-based visualizers to support visual analysis.

Concerning the **interoperability of data** and **workflows**, our application domain is characterized by a large number of different data file types and data visualization tools. Therefore, the goal of the Visual Media Service is to contribute to this issue by providing **easy conversion tools** to a set of common formats and, more important, to a common visualization tool (accepting different types of data and using a common GUI). This could greatly simplify and improve the capability of the Cultural Heritage community[57] to work in a cooperative and remote manner over visual media.

Concerning **interoperability of data**, one extension on which the SD has worked in Months 6-7 is the **support of the *collection* concept** - all the files related to a semantically-related group of components are treated in a unified manner, both at data uploading and data visualization time.
The current *workflow* in Cultural Heritage, or in Digital Humanities, is composed by four classical phases:

- Visual data production (digitization phase, producing digital 2D or 3D models);
- Exchange of data among cooperating research units (still mostly by file transfer);
- Visual analysis of data (done locally and in isolation, with different tools); and
- Exchange of results/discussion.

We plan to change this workflow by removing the need of exchanging explicitly data among partners, but opening a web-based resource where data can be uploaded for everybody and accessed with a common visualization instrument.

Finally, our evaluation of the EOSC services and infrastructure is still preliminary. In the first phase of the SD we have mainly used the **authentication** functionalities (we found them fully functional to our scope and, thanks to the possibility to consult with the D4Science staff, we had any problem in using this support). In the last phase of the SD, after users upload of data and more intense usage of the service, we will evaluate the performances requirements and possible bottlenecks. Based on this feedback, we will decide if it will be needed to scale up to a more sophisticated remote processing model.

[57] https://en.wikipedia.org/wiki/Cultural_heritage

3. Conclusions and Future Work

The aim of this deliverable is to provide an updated picture of the different actors involved in the EOSCpilot project, that, through their activities, aim in shaping the EOSC environment, improving the services and e-infrastructures it consists of, and also provide requirements and recommendations, based on the experiences they gained during the project, to help in the prioritization of the new features of the existing services and of the development of new services that are aligned with the needs and expectations of researchers.

After the initial interoperability requirements gathering (D6.4), the present updates (D6.7), and the description about the status of the different pilots' setup (D6.5), D6.10 – "Final Interoperability Testbed report", the last deliverable of WP6, will conclude by providing the validation of the e-infrastructures and services deployed. For this final assessment we will take into considerations:

- Tools/services developed as part of other EC projects to implement the interoperability aspects, e.g. the Interoperability (IOP) Quick Assessment Toolkit, developed in the context of Action 2.1 of the Interoperability Solutions for European Public Administrations (ISA) Programme;
- The Technology Readiness Levels (TRLs) of the different services deployed during the pilots. In this case the criteria made available by the EOSC-hub project for on-boarding new services in the EOSC Service Catalogue will be considered (see Annex A.2).

Annex 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.
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 deliver 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

Annex A.2 - EOSC-hub – Technology readiness levels

The level of completeness and maturity (the "Technology Readiness Level" – TRL) of the EOSC-hub services is validated according to a set of mandatory criteria (A to D as described below), which were equally applied to all catalogue entries.

TRL8 "system complete and qualified": evaluation criteria

- **A (system complete)**: The value proposition of the service is defined with the related functional capabilities. A web page with the description of the service is available including its features.
- **B (system qualified)**: The service was deployed in an operational environment and successfully used in real- world scenarios by end-users, all its components achieved the expected performances level within the scope. Links to either an available running instance of the service or to the release notes are available.
- **C (system complete)**: User manual and admin manual (where relevant) are available and enable effective use and operation of the service within the defined scope.
- **D (system qualified)**: Helpdesk channels are available for support, bug reporting and requirements gathering.

Application of TRL 7, 8 and 9

All services included in the service portfolio have been evaluated for their TRL. The services in the portfolio have been assessed as TRL8 at the moment of writing:

- Services not fulfilling one of the four criteria are considered TRL7 and have been included in the portfolio only when documented, clear and feasible plans to reach TRL8 by September 2017 were provided.
- Services already deployed in production with an associated SLA or OLA are considered TRL9.

In summary, every service in the service portfolio at the moment of the project submission has at least a pre- production service available for the users. New services added to the services portfolio during the duration of the project will be evaluated using the same criteria.

Annex A.3 - EOSC SDs testbed commonalities – v1

SD	Interoperability demand	Interoperability solutions	Compute Nodes	Storage	Network	Data	Software	AAI	Services	Protocols	Licences	Formats	tbd	Problems encountered
ENVRI dynamics of greenhouse gases, aerosols and clouds and their role in radiative forcing	cooperation between environmental infrastructures.; Interoperability between observations and climate modeling		OneData EGI node (Virtual Machine); A VM provided on EGI FedCloud was configured to fetch data-sets from the original data repository into the OneData volume.	CYFRONET configured OneData disk space in the EGI DataHub service to allocate 2TB in the Ceph installation		transfer IS-ENES climate model data hosted on the DKRZ node of the Earth system Grid Federation (ESGF) on a OneData EGI node (Virtual Machine); Model-data integration by use of HPC; Model-data integration by use of HPC; 1.1TB of IS-ENES climate data models have been uploaded in the CYFORNET data infrastructure	Synda (a command line tool to search and download files from the ESGF archive)		compile and compare model output from different sources	http or gridftp (exploiting Synda)			ICOS RIs check/validate transferred datasets	Initial issues with the Oneprovider - solved by developers of Onedata, and released in version "RC8" of the service;initial lack of resources to support the SD - solved with the contribution of CYFRONET
PanCancer PCAWG (Pan-Cancer Analysis of Whole Genomes) analyzing >2800 genomes; establish a portable cloud-based federated solution for collaborative cancer genomics and associated health data management,	compute and data environment accessible to European scientists for analysis		EMBL-EBI Embassy Cloud in the UK: 700vCPU cores, 2.6TB RAM; Cyfronet: 1000 vCPU cores, 4TB of RAM; ComputeCanada : 1000 vCPU cores, 3.9TB of RAM	EMBL: 4.9 TB volume storage; Cyfronet: 1 PB NFS storage; ComputeCanada: 62 TB volume and 200 TB NFS storage	EMBL: 32 floating IPS	data from the ICGC pediatric brain cancer cohort were downloaded to the ComputeCanada cloud; public data from the 1000 Genomes project was loaded onto the Cyfronet environment from the EMBL/EBI data servers utilising Cyfronet's Oneprovider software; ComputeCanada: 400 data sets, 60TB; Cyfronet: 50 TB; 100 TB data processed with Butler on <400 samples each at ComputeCanada and EMBL/EBI	The Butler scientific workflow framework has been set up and tested at three globally distributed cloud computing environments that are based on the OpenStack platform. Including monitoring solution and self-healing modules by Butler. Used OpenStack, OneProvier						Currently testing the new set-up of Cyfronet	CYFRONET addressed scalability issues in the context of the HNSciCloud project;stability issues encountered at the Cyfronet environment with the allocated storage and data staging mechanism (Oneprovider); Solved a storage throughput issue faced at the ComputeCanada environment; Uniform identity and access management across environments remains an unsolved issue

Photon Neutron interoperability storage systems and integration with function as a service to manage workload on the cloud. Tool: Kafka messaging, server sent events.	Interoperabiltiy between different storage systems		At DESY: Local OpenStack cloud and compared with present workflow on HPC cluster "Maxwell"; Currently in the process of significantly scaling up the on-site OpenStack infrastructure due to successful proof of concept, Used in SD: 192 physical CPUs (284 logical); 1.15 TB RAM: Currently upgrading to: 512 physical, 924 logical and 7.3 TB RAM	AFS, CVMFS, conventional (NetApp) SMB/NFS-shares, dCache and the local/federated Nextcloud instance (also backed by dCache) in order to provide data exchange between VMs and external resources outside the OpenStack environment	~10 Floating IPS, which are during the project phase internal VLAN only. During interoperabilty pilot in WP6: currently moving resources into DMZ, opening new cloud cell and preparing access to/from external resources. No Infiniband available in the cloud by now, low latency gpfs only on the HPC Cluster, export as NFS to cloud only..	Sets of similar texts concerning related topics (e.g.: archaeological excavation); Small size (datasets of KBs or Mbs); metadata creation	CrystFEL: used at various synchrotrons and FELs to analyze datae from serial (femto-second) x-ray crystallography; OnDA (online data analysis) utility; OpenStack; OpenWhisk, Docker, Puppet, Foreman; Jupyter Notebooks	LDAP, Kerberos, OpenStack Keystone, DESY idp	web-services for easy consumption and visualization of the data. Run containerized workflows (tested singularity, using docker, examining Kubernetes); private Docker registry	tcp, http, xrootd, webdav; RESTapi and AMQP used to communicate with services	complex, non-redistributable software stack, which is free to use by academia and non-profit organizations	NeXus/HDF5; Docker (yaml), HEAT templates (HOT), qcow2 for cloud images		data policies in laboratories; licences of the software mapped to new usecases (e.g. distribution via docker)
Textcrowd enable a semantic enrichment of text sources and make it available on the EOSC; text documents that are the main part of datasets used in Digital Humanities and Cultural Heritage research; improve semantic item searchability through the ARIADNE catalogue	Standard NLP encoding to make data interoperable with other NLP tools		Deployed on EGI VM	Interested to use B2DROP to store/share Metadata; also stored in ARIADNE Registry, PARTHENOS Registry		Sets of similar texts concerning related topics (e.g.: archaeological excavation); Small size (datasets of KBs or Mbs); metadata creation	modular and extensible framework; advanced machine learning scenario. POS and NER frameworks		Sets of similar texts concerning related topics (e.g.: archaeological excavation); Small size (datasets of KBs or Mbs); metadata creation			semantic format: CIDOC CRM (ISO 21127:2006) encoding in RDF format; Machine readable and consumable data		

WLCG long-term preservation and re-use of HEP data, documentation and associated software	Use PIDs in a non-communitiy specific manner; A virtualised environment that runs in Cloud, Grid and many other environments.			B2SHARE and Zenodo		100 TB and more of OpenData			Scalable “digital library” services where documentation is referenced by a Digital Object Identifer (DOI);OpenAIRE (e.g. Zenodo), EGI Engage (e.g. CVMFS – already offered by EGI InSPIRE SA3 (led by CERN)) and a Trustworthy Digital Repository (TDR).					significant delays in finding a host for the Trustworthy Digital Repository (TDR); final report states: “The objective and the simple goal outlined above was not achieved”; services are not interoperable

Prominence Demonstrate ablity to run small scale containerised applications on cloud resource	Ability to burst onto EGI FedCloud; Access to CCFE cloud for external communities	Initial tests used EC3 to deploy elastic SLURM clusters on EGI FedCloud sites. Have also used VM Operations Dashboard and OCCI command line tools. Switched to HTCondor using hooks to interact with a wrapper around Infrastructure Manager to gain greater flexibility and automated error handling. Tested on Azure and Google Cloud Platform in addition FedCloud sites.	133 vCPU cores at CESNET-MetaCloud, 8 vCPU cores at IN2P3-IRES and CESGA ; CCFE cloud and local OpenStack; also commercial clouds.	Up to ~1TB NFS server node created for each job if needed for temporary files and have also tested BeeGFS On Demand and GlusterFS as alternatives to NFS; Ceph based storage at STFC with S3 and Swift API	infiniband; But all VMs on EGI Fedcloud need public Ips atm	Generated data accessed through automatically-generated temporary URLs (Swift). Data is test data not intended for publiction or release.	MPI/OpenMP jobs, Fusion MPI modeling codes, Geant 4 simulations; containerized (udocker, also have tested Singularity and Shifter);SLURM and HTCondor; Grafana for visualization, InfluxDB for storing, metrics collection through Telegraf	Different credentials are being used to deploy infrastructure on EGI FedCloud (X.509 certificate) and storage (access key and secret key)	Intended as demonstrator - may become a production service in the future	ssh	N/A	Code dependent	Provide more widespread access to internal and later also external users; ploicies for different cloud providers; Apply INDIGO PaaS Orchestrator; enhance IAM and use SLA manager	Solved initial problems deploying dockerized tools on EGI FedCloud; Bug fixes on SLURM cluster RADL in the INDIGO IM Github repository; corrected the IM role in Ansible Galaxy to enable EC3 to deploy clusters on EGI FedCloud sites; waiting for existing sites in EGI FedCloud providing low latency networking and determine availability automatically; need to harmonize AAI amongst platforms; MPI versions in container depend on the version on the host; all VMs on CESNET-MetaCloud lost due to a problem with FedCloud; recently udocker cannot sucessfully pull images from DockerHub unless the images are cached within DockerHub's infrastructure; poor documentation for many INDIGO-DataCloud software products; INDIGO PaaS Orchestrator doesn't support FedCloud yet; resource limitations of fedcloud.egi.eu VO meant that existing solutions for deploying elastic batch systems on clouds not suitable for us (need to use small numbers of cores at many different cloud sites simultaneously, automatically handle failures caused by quota limits, too many keypairs, no free floating IPs left etc); getting access to example fusion codes has been time consuming
EGA Life Science Datasets develop pipelines and security mechanisms to upload (FAIRify) and process genome datasets and metadata on cloud resources and make available for portable clients			Running at the Barcelona Supercomputing Center (BSC)	Interest to integrate the B2FIND Metadata Catalog		Data sets are large, too large to be processed seamlessly on currently available cloud resources; Auxiliary files needed for datasets	GoNL pipeline, containerized (Singularity). NextFLOW workflow manager; Multiple versions of the domain specific data analysis software used					BAM files	Addressing security challenges and policies; Currently working on interoperability between metadata schemas and B2FIND	Multiple Versions of the analysis software posed a challenge on findability and reproducability, the same applies to auxiliary files involved; Not all modules and tools ready to run in containers and integrate in cloud environments;

EPOS/VERCE Produce data sets on cloud resources and get ready for FedCloud integration; Post-processing workflows on Cloud resources; test with large data sets and concurrent users			IN2P3-IRES, SCAI and HG-09-Okeanos-Cloud, of the EGI Federation configured to support the verce.eu VO	Self managed iRODS instance	Public Ips and port exposure needed on FedCloud resources for DCI-Bridge; need for support of Per-User Sub-Proxies	simulated seismic waveform images, wave propagation videos, 3D volumetric meshes, sharable KMZ packages and parametric results	Provenance system (S-ProvFlow), dockerized; Using DCI_BRIDGE VM image available on EGI AppDB	extended AAI framework with EGI AAI Check-In using the OIDC module; work in progress, based on Unity		RESTapi (SprovFlow)				VM image required by gUSE Cloud integration updated (dci-bridge, Obspy & GridFTP included, LVM issue); iRODS GridFTP DSI do not support Per-User-Ssubproxies
CryoEM workflows Share image processing workflows and enhance reproducibility by linking to acquired data and software and fully describe image processing steps	Shared service accessible for different facilities in Europe			Public database EMPIAR		Workflow files in JSON	Scipion to manage image processing pipelines used at electron microscopy; Webviewer to visualize files					JSON; Tested Common Worfklow Language (CWL)	Finalize work on jsonbased detailed workflow description and web viewer	
LOFAR Astronomy OpenScience Cloud; share existing LOFAR data in accordance to FAIR principles and enable container-based online data processing	Build a distributed system to locate, access and extract scientific results from the LOFAR archive; Run on variety of platforms: MacBook, Linux, SURFsara resources HPC cloud and Cartesius supercomputer	Build on existing tools like Xenon, CWL, Docker, Virtuoso		B2SHARE resources; PSNC/FZJ resources			Domain specific tools: Prefactor (calibration), Presto (framework to search specific signals in LOFAR data), Spiel (radio telescope simulator); Used Docker, Singularity and uDocker; Toil workflow engine	B2ACCESS				CWL; made debian packages from used software	Data access and storage, UI enhancements; Notebook based processing	Fixed bugs in CWL (allow nested directories, sorting and alignment); Could not finish work on Mesos scheduler on HPC cloud); Security concerns using docker, prefer Singularity; GPU accelerations introduces rigid dependencies on version on host and in container (Cuda Drivers, NVIDIA kernel modules) workaround only for docker.

Generic Technology Frictionless Data Exchange Across Research Data, Software and Scientific Paper Repositories	cross-disciplinary network of repositories to efficiently and reliably exchange scholarly communication; highly scalable exchange of data across repositories	Design and implementation of a scalable client for accessing metadata and full texts					Sets of similar texts concerning related topics (e.g.: archaeological excavation); Small size (datasets of KBs or Mbs); metadata creation			ResourceSync; OAI-MPH			Testing done using TextCrowd data set; develop benchmark suite and finalise evaluation	Need to sync a large amount of small files (using ResourceSync Batch); Benchmarking different systems on identical test data wrt network latencies
Life Sciences – Genome Research - Bioimaging Mining a large image repository to extract new biological knowledge about human gene function	perform comprehensive machine learning analysis on large cloud-based collection of image-based genome-scale data sets; demonstrate the use of the infrastructure for users to run their own analysis via the cloud on publicly available image data sets	demonstrate the use of the infrastructure for users to run their own analysis via the cloud on publicly available image data sets; OpenLava workflow scheduler	resources on EMBASSY cloud ~200 low-memory CPUs at the beginning; ~20 high-memory CPUs at later stages	IDR repository	Openstack SDN	IDR repository	OpenLava to create LSF-like environment			NFS		Images in TIFF format	Scaling the SD deployment; provide better image access through API enhancements	resolved network performance issues; setting up environment incl. NFS, network, software installation; Dependencies on software versions; Firewall related difficutlies
VisIVO Data Knowledge Visual Analytics Framework for Astrophysics visual analytics environment based on Virtual Observatory and ViaLactea	publishing a data analysis and visualization environment fully integrated with a relevant set of astrophysical datasets		Local cloud resources not federated, using EGI FedCloud resources	EGI data resources; sustainable storage for 0.5-1.5 TB		~400GB of astrophysics survey data	DCI_BRIDGE used by the ScienceGateway to run scientific workflows		Published Image on EGI AppDB				testing with multiple concurrent users	bug fixes, documentation, and workflow refinements especially when exploiting cloud infrastructures

Computational Hydrology FAIRifying top-down approach of the eWaterCycle project and bottom-up approach of the SWITCH-ON		Sets of similar texts concerning related topics (e.g.: archaeological excavation); Small size (datasets of KBs or Mbs); metadata creation		OneData storage			Domain specific tools SWITCH-On and eWaterCycle; OneData; Containerized deployment; support CWL, BMI, OpenDA					CWL; NetCDF		
Visual Media service for sharing and visualizing visual media files on the web	provide researchers with a system to publish on the web, visualize and analyze images and 3D models in a common workspace	run workflow on D4Science infrastructure		personal per user storage needed (D4Science User Workspace)		Per user data browing needed	Visual Media Service OESC v.1	Authentication services of D4Science and Google						