D6.2: EOSC Architecture Design and Validation Procedure

Author(s)	Eric Fede, CNRS
Status	Final
Version	V1.4
Date	10/11/2017

Dissemination Level

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

Abstract:

This deliverable of the EOSCpilot project has as objective to define the architectural design of the interoperation of various types of infrastructures, which could participate to the future EOSC. This task has to be seen in context with the work in Work Package 5, where a possible service architecture will be proposed.

Beyond the e-infrastructure architecture itself, this document will propose and validate technical solutions and suggest best practices for enabling interoperability across multiple federated e-infrastructures. In order to accomplish this, we will use the gap analysis of the interoperability of the existing e-infrastructure landscape (deliverable D6.1).

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 –WP6-2	
Deliverable lead	**CNRS**
Related work package	**WP6**
Author(s)	Eric Fede, Geneviève Romier, Volker Beckmann (CNRS)
Contributor(s)	Donatella Castelli, Prodromos Tsiavos, Brian Matthews, Sarah Steele
Due date	**30/09/2017**
Actual submission date	**15/11/2017**
Reviewed by	Donatella Castelli, Prodromos Tsiavos
Approved by	**Brian Matthews**
Start date of Project	**01/01/2017**
Duration	**24 months**

Versioning and contribution history

Version	Date	Authors	Notes
0.1	**04/09/2017**	Eric Fede	First draft
0.2	**05/09/2017**	Eric Fede, Geneviève Romier, Volker Beckmann (CNRS)	Second draft with first feedback
0.3	**18/09/2017**	Eric Fede, Geneviève Romier, Volker Beckmann (CNRS)	
0.4	**03/10/2017**	Eric Fede, Geneviève Romier, Volker Beckmann (CNRS)	Version considering Donatella Castelli's comments
0.5	**04/10/2017**	Eric Fede, Geneviève Romier, Volker Beckmann (CNRS)	Upgrade by V.Beckmann and E.Fede
0.6	**10/10/2017**	Eric Fede, Geneviève Romier, Volker Beckmann (CNRS)	Version considering Prodromos Tsiavos' comments
0.7	**10/10/2017**	Eric Fede, Geneviève Romier, Volker Beckmann (CNRS)	

0.9	**16/10/2017**	Eric Fede, Geneviève Romier, Volker Beckmann (CNRS)	Final version with last corrections
1.0	**23/10/2017**	Brian Matthews, Sarah Steele (STFC)	Editorials changes for consistency and style
1.1	**24/10/2017**	Eric Fede, Geneviève Romier, Volker Beckmann (CNRS)	
1.2	**26/10/2017**	Eric Fede, Geneviève Romier, Volker Beckmann (CNRS)	
1.3	**8/11/2017**	Eric Fede, Geneviève Romier, Volker Beckmann (CNRS)	Iteration with Brian Matthews
1.4	**10/11/2017**	Eric Fede, Geneviève Romier, Volker Beckmann (CNRS)	Iteration with Brian Matthews

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

1. EXECUTIVE SUMMARY

The European Open Science Cloud will federate IT infrastructures and will be a fundamental component of Europe's future research infrastructure and Horizon 2020 actions. The EOSCpilot project brings together a broad range of stakeholders from Research Infrastructures, e-Infrastructure services and e-Infrastructure providers. The WP6 work package of the EOSCPilot project addresses the issue of interoperability, and in this context will recommend solutions to interconnect various types of infrastructures provided by the many actors within European research. Interoperability of infrastructures covers many aspects in order to accomplish the challenges of an efficient construction of deployment processes and mechanisms.

The initial EOSC service architecture is the responsibility of work package 5, mainly with the production of deliverable D5.1 in Task 5.1. This deliverable (D6.2) however, has as its objective describing the framework that needs to be set to allow the interoperability between the e-infrastructures and Research infrastructures (RI) involved in the EOSC project.

The second objective of this deliverable is to define a list of procedures which should be in place to ensure that interoperability can be guaranteed. This list of procedures is not be exhaustive and should be considered as a set of good practices, which the components/infrastructures of the EOSC project will have to satisfy.

The work done so far on the architecture design, has been an analysis of the gaps which could potentially prevent the successful interoperability of the EOSC e-infrastructure architecture. The results of this study can be found in the deliverable D6.1[1].

This document is a first draft of the architecture design concerning the interoperability of the future EOSC. This current version will evolve as the project develops and the issues, questions, proposals, and approaches mentioned in this document may be revised. This shall lead at the end of the EOSC pilot project to the deliverable of the interoperability work package D6.8 "Final EOSC Architecture".

[1] The gap analysis can be downloaded here: https://eoscpilot.eu/content/d61-e-infrastructure-gap-analysis

2. CONTEXT AND OBJECTIVES

The aim of the EOSC is to provide access to European e-infrastructures through interfaces that will allow seamless usage, and will enable connectivity between services and data across disciplines and across borders. This will be a 'system of systems' that will be built by relying on an open set of existing 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.

In order to allow the connection of the existing e-infrastructures/RIs to build a pan-European system that can be easily used, several points of view have to be considered:

- e-infrastructure/RI providers, who need to know how they can (technically) make their infrastructure available within a larger system that includes similar and differing types of computing and storage environments;
- service providers who want to provide cross-disciplinary platforms that will be used on the EOSC, and who need to know how to implement them on the overall EOSC infrastructure;
- users (including RIs), who need to know about the entry points where they can make use of the services and e-infrastructures, and how and under which conditions, they can use the underlying infrastructure;
- funding bodies, who need to have a way of accounting for the usage of the e-infrastructures under their responsibilities by the EOSC.

EOSC is also open to the integration of cloud or other infrastructures coming from commercial providers. Therefore, it is also important to consider the standardization aspects from the public and private point of view.

In this document we focus on the first two aspects, i.e. what will be necessary in order to technically integrate e-infrastructures in an overall EOSC, and the requirements that need to be fulfilled in order for service providers to be able to install their packages on this system.

The initial EOSC services architecture comes under the responsibility of work package 5, with the production of deliverable D5.1 that will describe a first outline of this future architecture. This deliverable, (D6.2) has as its objective to describe the framework that needs to be set to allow the interoperability between the e-infrastructures and RI involved in the EOSC project. This means that D6.2 does not define the service interactions, and does not define how the services and components offered by existing national and European generic e-infrastructures, need to work together. Instead, the first goal of D6.2 is to identify the elements and principles that will allow the best interoperability between the e-infrastructures and RI.

The last part of this deliverable is to define how to ensure the interoperability of the e-infrastructure/RI. The list of methods or procedures to satisfy this task is not exhaustive and should be considered as a set of good practices, which the infrastructure of the EOSC project has to satisfy.

The main principles of how e-infrastructures should be interoperable within the EOSC is described in Section 2.1, and Section 2.2 gives some important aspects about the current e-infrastructure landscape in terms of interoperability.

Section 3 summarises the results of the e-infrastructure gap analysis and points out potential solutions. Section 4 will then describe methods that have to be put in place to overcome

interoperability issues, and Section 5 addresses the issue of validation to verify whether an e-infrastructure has been integrated in a seamless fashion.

2.1. Main Principles

Some requirements can be extracted concerning the interoperability by focusing on two of the aims outlined in the abstract of the project: “Improving interoperability between existing infrastructures” and “develop a number of pilots to demonstrate interoperability in a number of scientific domains”.

To meet both requirements, the global architecture of the infrastructure involved in the EOSC should satisfy the following conditions:

1. A large majority of the existing resource providers should be considered as potential EOSC participants within the global infrastructure, and it should be possible to integrate them. Resource providers of computing infrastructures, such as High Throughput Computing (HTC), High Performance Computing (HPC), Cloud based solutions on different backend technologies (CPU, GPU XeonPhi, FPGA,…), or diverse solutions for data storage and data repositories should be considered as being able to contribute to the potential infrastructure and thus participate in the EOSC.

2. This approach is also true concerning the e-infrastructures that are elements of the global EOSC context and should be able to join the EOSC initiative, and also services may also discontinue and leave the EOSC. This means that no specific dependency on any one service provider or e-infrastructure should be considered or finally allowed in the EOSC architecture.

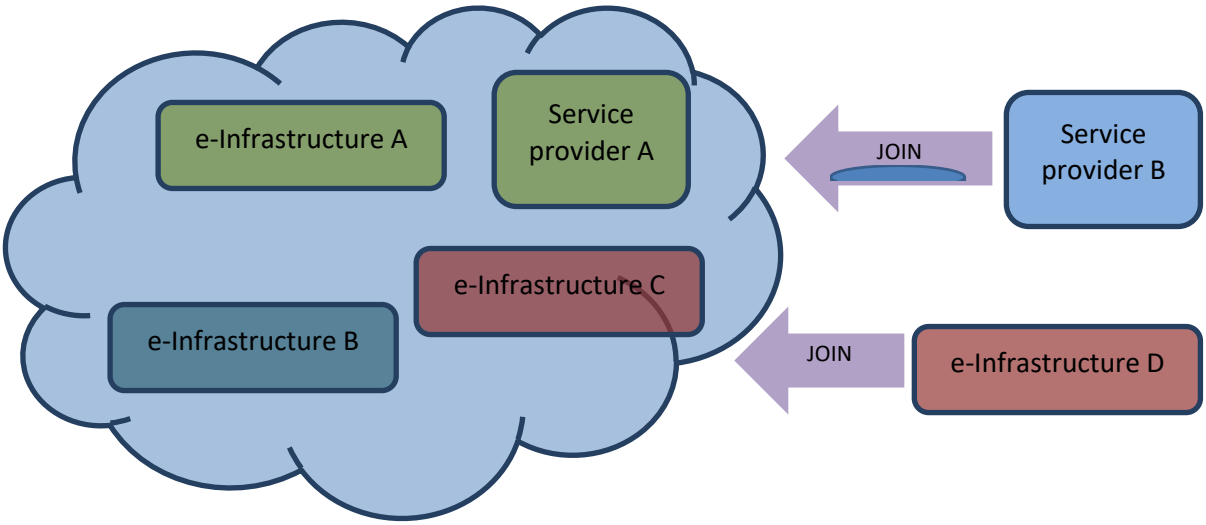

Figure 1 : e-infrastructure/RI add on EOSC context

Figure 1 displays e-infrastructure and resource providers which join a set of e-infrastructures and RIs, that are already connected in the EOSC context.

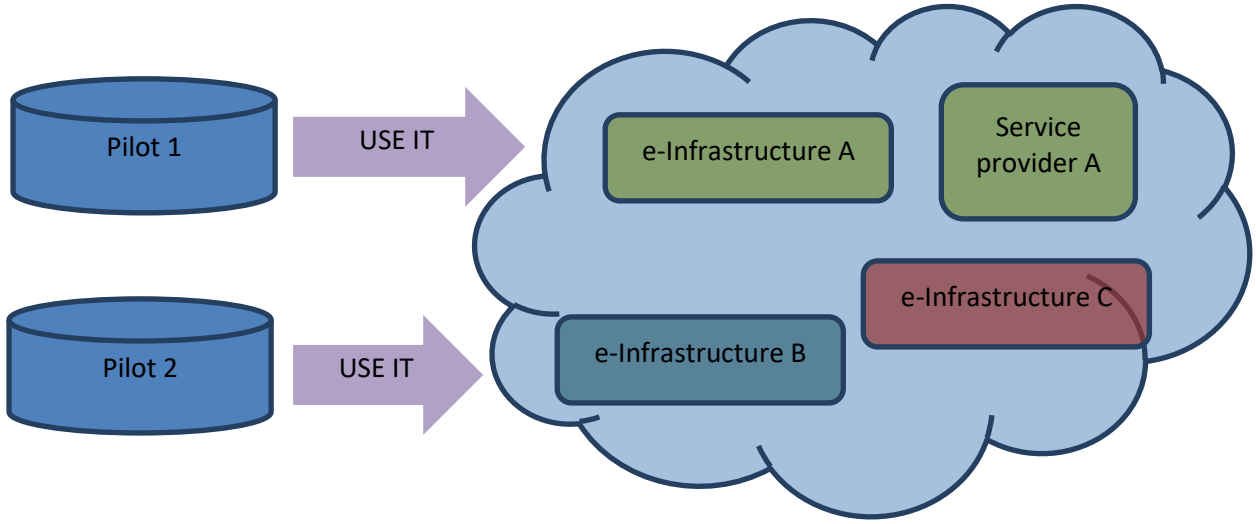

Figure 2 : Pilots on e-infrastructure/RI set

Figure 2 shows how every pilot identified in the project can be deployed within the EOSC context and benefit from it. The framework proposed by the EOSC should be independent of the usage.

These two conditions can be summarized by the requirement to have an architecture compatible with the instantiation of multiple-infrastructures for multiple-communities.

2.2. Availabilities of e-infrastructures

Because the interoperability of the e-infrastructures involved in the EOSC will be dependent on their relationships and the exchange of information between them, there is an urgent requirement on the availability and reliability of these e-infrastructures. The E-infrastructures involved in the EOSC must provide excellent reliability, a high level of availability and rapid responsiveness to problems. Beyond the e-infrastructures, this requirement concerning the availability and reliability extends also to the services provided by those Infrastructures. Consequently the e-infrastructures have to be internally organized both to provide services with a high level of performance and also to monitor this performance monitoring results must be published. Thus e-infrastructures should publish Quality of Service conditions and metrics to engage with the EOSC. These are being considered under the *Terms of Engagement* task in WP5.

Depending of the services or e-infrastructure functions, the level of availability and reliability would be different and we can imagine easily to be more necessary with the services involved on the data management than the services involved on the computing or on monitoring. The level of

availability guaranteed by an e-infrastructure has to be published in order to allow user communities or RIs to choose among e-infrastructures those that meet their needs. This level of availability must be preferably published in a machine readable form that allows portals, gateways or middleware brokers to take it into account automatically.

The next table is an example of availabilities and maximum time of degraded services/function than we can find frequently concerning e-infrastructures.

Type of functions/services	**Availabilities (annual base)**	**Functions/services interruption**	**Degradation of 50% of the services/functions**
Data management/accessibility	98 %	Less than 4 hours	Less than 6 hours
Computing	96%	Less than 8 hours	Less than 12 hours
Critical services (AAI, catalogs,…)	99%	Less than 1 hours	Less than 2 hours
Others services/functions	95%	Less than 12 hours	Less than 1 day

Table 1 : Typical availabilities and maximum time to solves an issues requested for e-infrastructures

2.3. Landscape of e-infrastructures

The landscape of the e-infrastructures today is heterogeneous. Resource providers, on which the e-infrastructures are built, have different geographical footprints (regional, national, European…) and different relationships with their user communities.

For the most part, they are providing resources in a specific technical context and/or with a focus on a small subset of computing methods to increase the efficiency of the use of resources for a specific problem. Some others are providing resources in order to be the most efficient for a small range of issues.

Concerning the computing aspect, we find that historically there is an HPC and HTC approach. Nowadays some other technologies, such as grid, cloud, and GPU computing, have matured and also have to be considered. Storage and data aspects are also relevant, including databases and “flat data” storage approaches, together with issues in relation to data management and data access. Managing a large set of data (Big Data) within some scientific domains cannot be seen just as a problem of scaling. Data management and computing models have to be revised in order that they are able to use efficiently the computing infrastructures and services which are available.

For many years, a large number of European e-infrastructures have been built with various and diverse objectives. Some examples are given below:

- **GEANT**: interconnects Europe's national research and education networking (NREN) organisations with an award-winning high bandwidth, high speed and highly resilient pan-European backbone – connecting Europe's researchers, academics and students to each other, and linking them to over half the countries in the world.

- **EGI**: delivers advanced computing services to support scientists, multinational projects and research infrastructures.
- **EUDAT**: is a service-oriented, community driven, integrated initiative which provides solutions and services for efficient and reliable storage and transfer of the data.
- **PRACE**: is a European project whose goal is to provide researchers from both scientific and commercial fields with access to high performance computing resources contributed by data centres across Europe

However, most of these heterogeneous infrastructures will have to be federated within the EOSC context. The EOSC can benefit from them as they provide some solutions to increase its interoperability, which can be reused in the EOSC context.,.

3. GAPS IDENTIFIED AS RISKS FOR INTEROPERABILITY

The European Interoperability Framework (EIF) will give specific guidance on how to set up interoperable digital public services in order to improve their quality[2]; the EOSC will take this framework into account in designing its approach to interoperability and its validation. The first step is to identify the main gaps concerning the infrastructure interoperability.

After a survey carried out among resource providers, e-infrastructures, and EOSCPilot science demonstrators, the EOSCPilot deliverable D6.1[3] identified six major gaps that should be considered in order to define the interoperability we have to achieve between the infrastructures that will build the EOSC:

- Diversity and incompatibility of the AAIs

  These risks include a large scope of subjects such as identity management and connection between diverse infrastructures, often with different technologies.

- Network services

  The network is the backbone of the interoperability between infrastructures. A reliable and flexible network is essential for the success of the EOSC project. The technical aspects of the network, i.e. speed and capability will have a major impact on the EOSC because this will interconnect the e-infrastructures. The costs of this network will be high.

- Diversity of services and providers

  EOSC "should enable trusted access services, systems and the re-use of shared scientific data across disciplinary, social and geographical borders" who require easy-access to the infrastructure for diverse communities of users and service providers.

- Diversity of access policies

  The landscape of scientific data providers today is characterized by the fact that each infrastructure or service, provides data access based on its own AAI service and with its own data access policies. EOSC aspires to find and re-use each other's data. This means that the diversity of the data management (including access policies) has to be bypassed to allow interoperability.

- Low awareness of the e-infrastructures and services

  This gap highlights the knowledge diversity of the research communities concerning the computing and data management services and infrastructures. Abilities to use the computing and data architectures differ strongly depending of the communities

- Lack of expertise and training

[2] European Interoperability Framework https://ec.europa.eu/isa2/eif_en
[3] The gap analysis can be downloaded here: https://eoscpilot.eu/content/d61-e-infrastructure-gap-analysis

Similar to the previous point, this gap highlights the lack of human networks that will ensure the provision of education and training. Sharing expertise across research communities and services/infrastructure providers is a key element to solve this issue.

For each gap, a set of actions to bridge the interoperability has been identified. These actions are summarized in the graphic and table below.

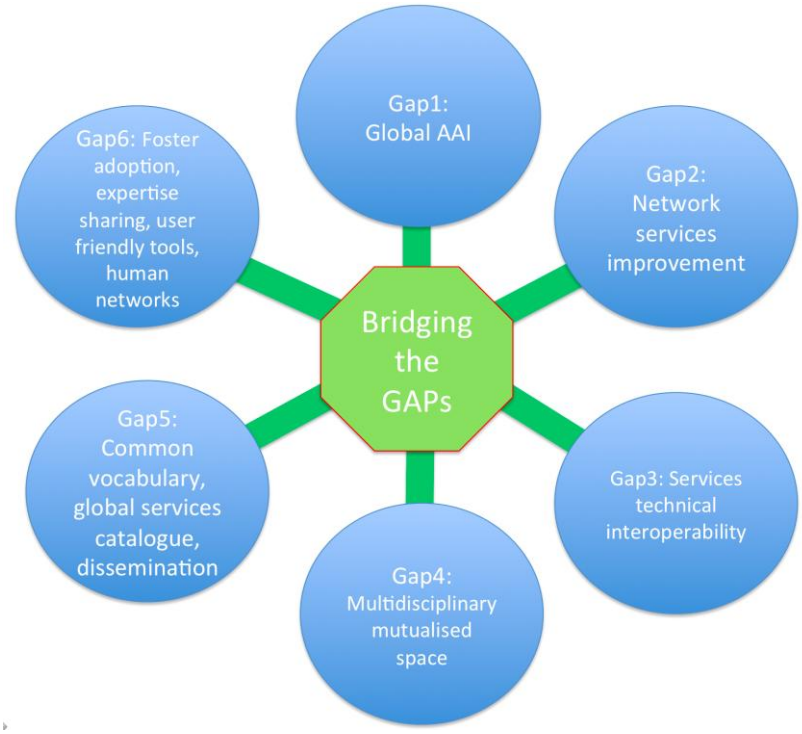

Figure 3: Actions necessary to bridge the interoperability gaps of e-infrastructures in the EOSC, as identified in the gap analysis (Romier et al. 2017, D6.1).

For each gap and bridge that will have to be built, different approaches and potential solutions might be considered. The table below summarizes these relations.

Gap	Description	Bridge to build	Interoperability approach	Potential solutions
1	Diversity and incompatibility of the AAIs.	Global AAI	Single referent for all the e-infrastructures or cross identification and authorisation between existing services	- Federation of identity : eduGAIN - Single referent: AARC

2	Network service	Network services improvement	- Improving network reliability - Improving network capability - Standardizing infrastructure connections to network	- Building the e-infrastructures and service providers on NREN and GEANT infrastructure.
3	Diversity of services and providers	Service technical interoperability	- Standardization of the interfaces to services. - Development of new open interfaces.	- Application of standards, REST APIs, WEBDAV protocol - Service catalogues, e.g. e-InfraCentral - Forums for developing standards e.g. Research Data Alliance
4	Diversity of access policies	Multidisciplinary mutualized space	- Single referent for all the e-infrastructures or cross identification and authorisation between existing services - Split between data , data description and data localisation	- Federation of identity - Data Catalogues and tools to access and share data.
5	Low awareness of the e-infrastructure and services	Common vocabulary, global services, catalogue	- Dictionary with common definitions of technical aspect of infrastructure - Usage of catalogues	- Service portfolio, E-InfraCentral - Data catalogues and repositories, e.g. OpenAire - Common work by the main e-infrastructures to build a common published vocabulary that allows everybody to understand each other.

6	Lack of expertise, training, easy tools, human networks	Foster adoption expertise sharing, user friendly tools, human networks	- Cross technical sharing between users and infrastructure manager - Computing and data management models sharing	- Tutorial - Forum - School User friendly portals Interware that allow users to use HPC, HTC, clouds through a single tool (such as Dirac)

Table 2 : Relation between gaps (as identified in D6.1) and how to bridge them

4. METHODS

This chapter focuses on the methods we have to consider based on the gap analysis, to allow interoperability between the e-infrastructures.

4.1. Standardization of the interfaces

Some of the existing e-infrastructures are building on a model (grid approach for EGI, EUDat for data) where they are used mainly by one or more specific user communities. This proximity between the infrastructure and its needs has led these infrastructures and communities to set up a specific communication mechanism. In an interoperable approach, all e-infrastructures must be able to communicate optimally with the user communities, which means that the usage of the e-infrastructure has to be easy from the application's/user community's point of view in order to optimize the use of resources, and to fulfil their users' requirements.

Ensuring a clear and direct relationship through some single channel of communication between infrastructure and user is, in terms of interoperability, clearly a response to the gap "Diversity of service and provider".

To meet this condition, which has social as well as technical and political aspects, it is essential to take into account different approaches concerning the relationships with the user community that may exist across the different e-infrastructures. From a technical point of view, it is necessary to define and standardise the different elements that are in the interfaces to the e-infrastructure, in order to normalise e-infrastructures usages.

From a practical point of view, this implies the definition and standardization of interfaces between infrastructures that provide the same function. In this context, some solutions already exist. For example, an infrastructure allowing the calculation, on HPC, HTC, GRID, and in a cloud, must be able to be addressed in a similar way. The same should be applicable to the management and storage of data, which must be able to be addressed uniformly.

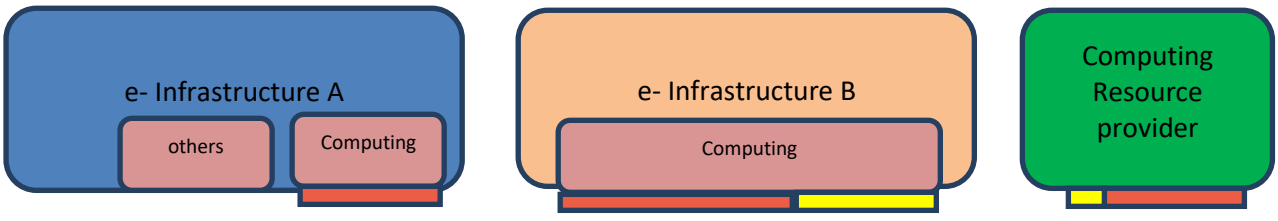

Figure 4: e-infrastructures and service provider interfaces

For example, in Figure 4. the e-infrastructures A and B and the computing resource provider all must be able to use the common open interfaces (coloured red), to receive and to respond to the

same request concerning the computing. Specific requests to local interfaces (marked in yellow) are still possible.

This approach of the e-infrastructure/service providers, partially covers another gap which is identified as "Low awareness of the e-infrastructures and services". It is essential to define the requests that e-infrastructures should have in order to be interoperable. In the figure above, this consists of completely defining what the common interface in red must be in terms of requests and in terms of response.

Many types of request require feedback from the infrastructures. These exist in relation with the function sought (execute a task, store a data set, access it …) but we also find a whole series of requests that relate to workflows that may exist between the infrastructures (trace back error, follow-up of actions related to the calculation of the data, capabilities of the infrastructure ....).

4.1.1. How to achieve standardization

This standardization cannot be easily reached, mainly because the e-infrastructure exists already with its own interfaces, and has been built to be compatible with some specific types of requests and will not yet be compatible with others. Several approaches can be undertaken to maximize easy interoperability between infrastructure/services:

1. First, using the API interfaces and protocols for every service provided by the infrastructure which already use underlying standards. Some standards like the REST API and the WEBDAV protocol are commonly used by commercial and non-commercial computing services. Allowing infrastructures/services to use these “de facto” standards will increase the potential interoperability of the infrastructures.
2. Development of new interfaces on infrastructures should also be considered. Fortunately, in many cases this development can be a “parser” to convert the current infrastructure interface to a more standardized one.
3. In some rare cases, developing a new interface may not be possible, and the solution will have to involve the deployment/development of a new service, which should be considered as a broker allowing an indirect usage of the infrastructure.

A full and precise description of these interfaces has to be provided by the e-infrastructures and services. Every API has to be described including the set of subroutine definitions and protocols, which compose the API. To increase the potential of interactions and consequently of interoperability with other e-infrastructures and services, APIs have to be provided for a larger number of coding languages including de facto standards like C++ and python today.

A list of tools allowing their users to use diverse e-infrastructures could be built and published on the EOSC project website. Many communities use such portals or interware. A first list has to include what is currently used, including information about to which e-infrastructures or type of resources it gives access to. This list has to be updated regularly.

4.2. E-Infrastructure dependencies

A clear picture of service provider dependencies is required to ensure the e-infrastructure interoperability. Because the services themselves are provided via e-infrastructures, the relationships between the e-infrastructures need to be published and openly accessible.

One way to achieve this is to build a relationships map of e-infrastructures which are included in the EOSC context. As inputs to this relationship map the e-infrastructures would have to publish at least:

- Each dependency on another infrastructure,
- For each of these dependencies a critical level has to be defined (low, strong, critical)
- A description *has to be included which functionalities/services will be degraded if the relationship should be terminated or interrupted*

4.3. *Networks*

The interoperability of infrastructures requires communication between them. This therefore gives to the network service a primary and major role. The organization and coherence of research networks in Europe is excellent, and all e-infrastructures are building on them, so the EOSC must not undermine this progress. Having GEANT as participant of the EOSCPilot project is a sign of confidence and an assurance in network provision.

4.3.1. Reliability of the network:

The EOSC system is proposed as a federated environment for scientific data sharing and computing usage. The confidence that the user/application will have of this federated environment will depend on the availability of the services provided. The first of them will be the network; the reliability of the network has to be total.

Today all around Europe, a set of National Research and Education Networks (NRENs) provide a reliable and efficient network for the research infrastructures. Beyond the national aspect, these NRENs have been working for many years, via a European partnership (GEANT), to interconnect this network in order to provide a reliable European network on which the e-infrastructures can be deployed.

A high level of reliability requires continuous monitoring of the network in order to minimize degradations or the shutdown of the network. That's means that network providers but also the e-infrastructures itself have to deploy a set of tools to monitor the network. From these data a mesh representing the connection status cross every e-infrastructure can be displayed in order to have a global picture of the status of the network connection between the e-infrastructure.

A strong interaction with all the stakeholders (NREN and e-infrastructures) of the network is also required. Regular interactions need to be held with all the stakeholders with the aim to organize the elements which can increase the network reliability.

4.3.2. Capabilities of the network

Interoperability of services/infrastructure distributed on geographical sites spread all over Europe assumes that the network provides a certain level of performance "to hide" the distributed aspect of the services/infrastructures.

NRENs and GEANT can today deploy some technologies on the European network backbone but some sites, which are connected with the appropriate network technology and which can guarantee a high level of performance (e.g. throughput, response time, backup links) between the sites. GEANT provides some VPN technologies today, which can be brought to the site (L3VPN, L3VPN) in order to provide the backbone of the EOSC.

The network services on which e-infrastructure interoperability must be based, should be

compliant with the technical solutions implemented by all the NRENs involved, in order to have a European dimension. These services, on which the overall performance of the architecture depends together with its range of technical solutions (security, identification, etc.), should not create a differential between e-infrastructures, as interoperability may be adversely affected.

Tools and probes that monitor the reliability/availability of the network can provide very useful and historical information concerning the capabilities of the network. E-infrastructures need to have access to these data in order to anticipate the evolution of network requests. Figures 5 and 6 give examples of network monitoring tools which can be used for this purpose.

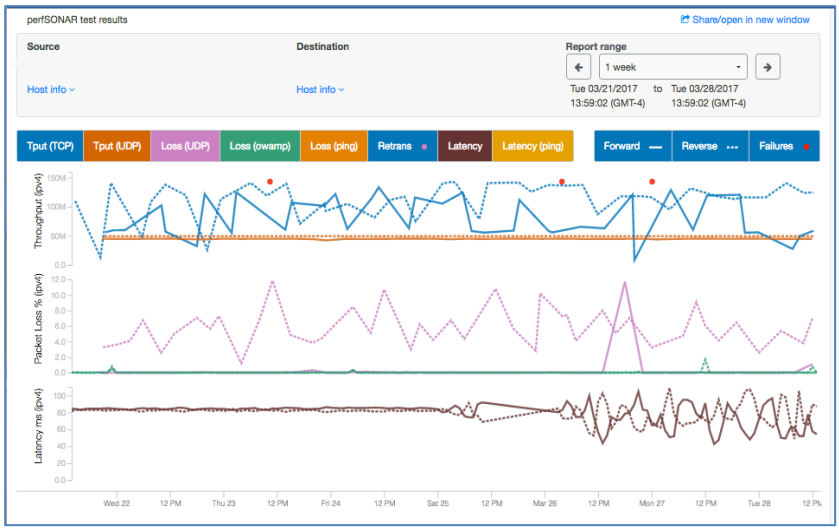

Figure 5: Example of a network monitoring with the perfsonar tool

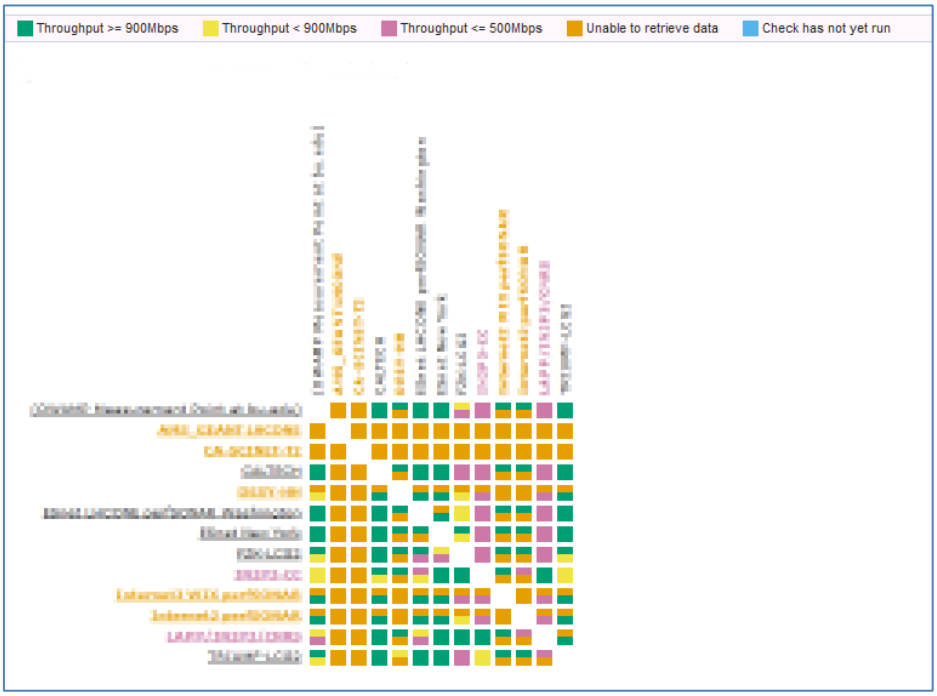

Figure 6: Example of a mesh display (Throughput)

4.3.3. Connectivity

Connectivity is also an important aspect for infrastructures. Beyond the services which are deployed by the network provider, the way that infrastructures are connected to the network is crucial to ensure they can be operated correctly. Infrastructures will be interoperable because they can rely on network services but also because they are interconnected through linked capacities, which are similar or close. The majority of the infrastructures or services providers, which are connected via a National Research and Education Networks network to GEANT, satisfy this requirement. However, we must remain vigilant regarding the infrastructures, which can be accepted as part of EOSC.

Efficient e-infrastructure interoperability requires the knowledge about how an e-infrastructure is connected to the other. To achieve that each e-infrastructure has to publish a description of this interconnection including a set of mandatory information:

- Availability of multiple paths to the e-infrastructures, dedicated links for specific usage, including rules how to use these.
- Redundancy of the path to the NREN
- Maximum throughput capacity
- Availability of dual stack (IPV4/IPV6) for the whole e-infrastructure (or for which services)
- Usage of public addresses: full infrastructure or only for external service
- A network contact

4.4. Authentication and Authorization Infrastructure

The identification of users and the authorizations attached to them are a major challenge within the EOSC. Ideally it is setting up a common user identity across all the infrastructures participating in the architecture of EOSC, and this is particularly necessary in order to allow the management of rights (authorizations).

It follows from this that there is a requirement for a single referent for all the e-infrastructures which would make it possible to identify the users and also to distribute the rights attached to them. This service for an Authentication/Authorisation Infrastructure (AAI) has to be global and should be remotely accessible by each service and e-infrastructure.

If the aim is to have only one system of identification for all infrastructures of computing that can potentially participate in the EOSC, we should consider the current status. This status is that the identification (and authorisation) services on e-infrastructure are – heterogeneous. However, the interoperability concerning authorisation and identification of these infrastructures could be attained if some bridges between the diverse AAI services can be built. This type of bridge can be based on different technical approaches, federation of identity, federation of federation, and also by some mechanisms of token exchange, or peer-to-peer id trusting.

GEANT is supporting a solution based on interconnecting identity federations around the world. This service “eduGAIN” enables the trustworthy exchange of information related to identity, authentication and authorisation. With 46 members, eduGAIN is the main mechanism to federate for research and education collaboration around the world. This service could be put forward as a

candidate to provide a solution for AAI which will increase the potential interoperability of the infrastructures.

The AARC[4] project is creating a common Authentication and Authorisation framework for Research and Collaboration communities. This approach should be also considered, because they would seem an appropriate approach to bringing about a single referent for all the e-infrastructures. This would make it possible to identify the users and also to distribute the rights attached to them.

The gap concerning the diversity and incompatibility of the AAI within the EOSC context is one of the most important and probably the most difficult to bridge. The ultimate goal within the EOSC are applications that are able to run all related workflows in the EOSC context with only one initial identification phase. This goal is still very ambitious but should be considered as the target when evaluating the interoperability of an e-infrastructure.

Management of the rights attached to the data (reading, writing, modification) is probably one of the more important items linked to the EOSC. Identified and managed data is a critical point from the applications/pilot point of view, and the diversity of the policy access of the data has been identified as a major gap.

4.4.1. Data access policies

Data management is one of the major EOSC characteristics and expectations. Allowing data transfers, movement, writing, erasing and access across different infrastructures and services is a challenge. Performance aspects exist of course, but these are more relevant concerning the network and service capabilities. Moving, and ensuring that the data access policies are compatible between the infrastructures, requires a strong level of confidence between them and a minimum set of policies that are common to every service.

The publication by every infrastructure/service of the data policies that are in place, is a minimal requirement to allow for a level of confidence. It is necessary that a thorough definition will be in place, of the core common elements on which the data policies will be applied, and also a definition of the actions of data management. These elements include for example:

- Definition of a “user”
- Definition of a “group”, “organisation”
- What means are in place for the actions of writing, erasing, accessing data
- Data licencing conditions

In addition, the user/pilot/application side should be defined. The life cycle of the data (data management) as executed by the applications should not depend on specific data policies of a given site in a sense that applications have be built to be compatible with the different site policies. In order to achieve this, sites and/or infrastructures have to publish clearly theirs policies and all useful information which can interact with data management aspect. Developing the workflow of the application, mainly concerning the data management, within guidelines will enable compatibility between application and infrastructure.

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

A clear split of the data and metadata is one of the good practices; another is the systematic usage of a data catalogue that allows placing data in many different data storage infrastructures.

4.5. Training and expertise

Some communities are already aware of the landscape of the e-infrastructures. In some cases, for example, these communities developed the infrastructure specifically according to their own requirements. But some other research communities are today are not aware of the capabilities of these infrastructures. Training and competencies within the EOSC are considered further in WP7.

An effort of dissemination to these communities is required, addressing topics such as how the e-infrastructures work, how to use them, and what are the good practices for building workflows for their applications.

Three types of expertise sharing are considered here:

- First, concerning the pilots/applications. The gap concerning the expertise on the usage of the computing infrastructures depends on the community. A useful action will be knowledge-sharing concerning the data model, computing model and computing methods. Feedback from the communities, which are familiar with the infrastructure usage, could be very useful.
- The second type of expertise sharing concerns the infrastructures. This type of meeting between the infrastructures/services is necessary for some technical aspects (definition of interfaces, solving technical misunderstandings…) and it is also essential to create a level of confidence between the infrastructures. Interoperability of the infrastructures will be efficient only if the infrastructures are mutually confident.
- The last type of knowledge transfer is linked to the training aspect, and the aim here is to optimise the infrastructures usage by the applications/pilots. Here, the users explain their aims, and the expert will explain how to do this as efficiently as possible within the EOSC infrastructure.

The last type of sharing expertise described above corresponds typically to a tutorial session or school event. This type of meeting can be repeated regularly (typically two sessions per year), but some specific training should also be set up as required (for example, because it concerns a rare usage case or high-level technical subject).

Online access to the tutorials (published under free licences) is desirable to reach, mainly to open the content to the people and communities which cannot participate in the event. It is also useful as a method to evaluate by the popularity of the tutorial, the interest and to advertise the contents of the next session.

The first two can be considered as forums where some information will be shared. Having the support from communities with certain competencies to share their expertise is one of the main challenges. Technical aspects could drive the cooperation and discussion between the infrastructures. Usually these meetings have to produce a set of documents (minutes, proceedings) which allow to build a knowledge base where the interested communities and infrastructures can find useful information to be stakeholder or to join the EOSC. This knowledge base has to be organised and accessible through a unique point of contact such as a website. The content has to be provided published under free licences to allow for a large distribution. All these events must be publicized largely to be sure every potential user is made aware of them.

A certain amount of monitoring of the training activities has to be performed. For meetings and trainings, the following information has to be gathered and made accessible:

- Number of participants
- Communities involved
- Number of documents in the knowledge base and number of consultations of each of these

This information will provide important metrics to evaluate efforts engaged to bridge gaps associated to lack of expertise.

5. ENSURING INTEROPERABILITY

Ensuring the interoperability of the e–infrastructures requires an important effort of communication between each of the resource providers or infrastructures, so as to define the service interactions, and to normalize (or including a level of abstraction) the interface between each of them.

But the interoperability on the EOSC system can be validated also from a high-level point of view. This proof should be done from the pilots/users side.

One of the elements of interoperability from the applications point of view, is the possibility to deploy a workflow (for job and data management) with any specific considerations in relation to a specific e-infrastructure. The workflow in which the pilot has some activities of computing and/or data management, has to be independent of the e-infrastructure used.

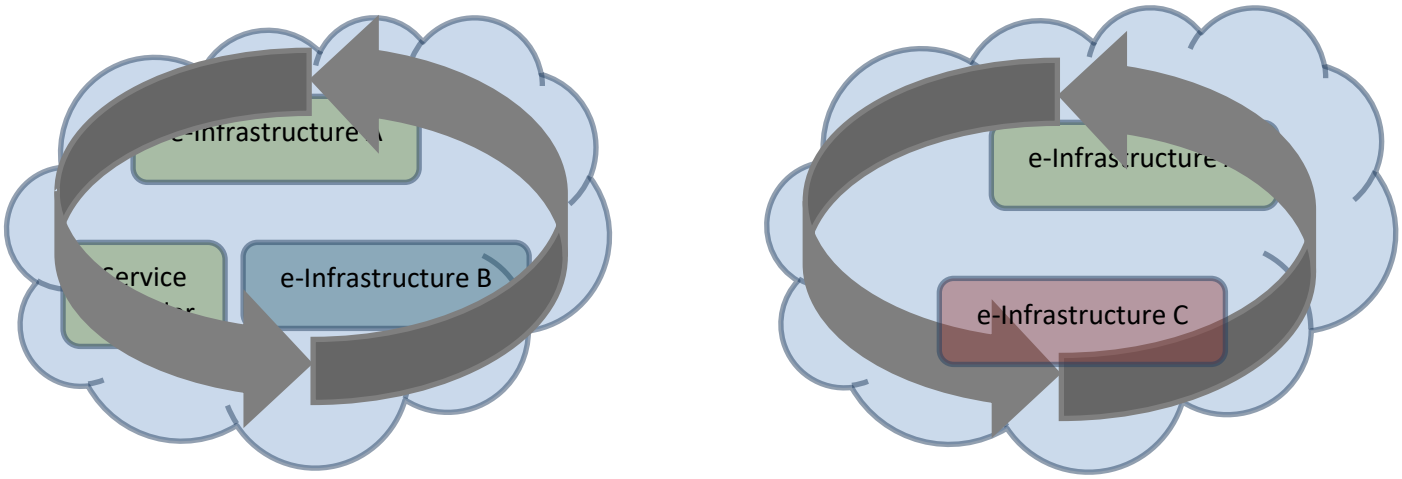

Figure 7 : Same workflow on a different set of infrastructures/services

From the infrastructure side, a good test to validate the interoperability is to check that a particular e-infrastructure could be replaced by another (allowing the same functionalities) in a transparent way from the pilot point of view. If successful, this test would ensure the interoperability in term of functionality, but wouldn't ensure that the capacity provided by the infrastructure is restored.

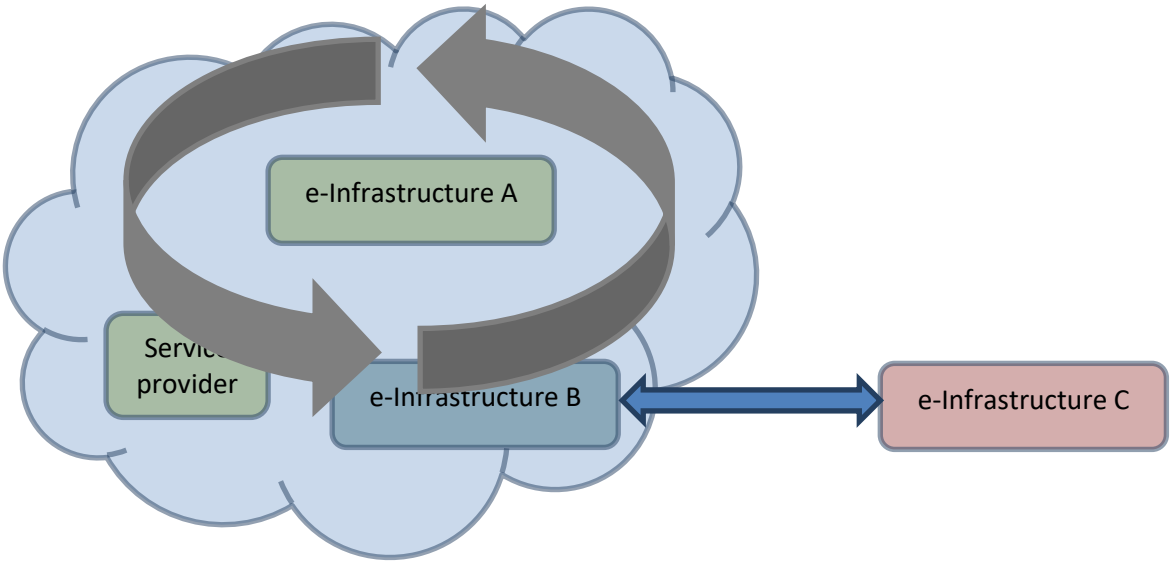

Figure 8 : e-infrastructure replacement, with same functionalities

Ensuring interoperability of the infrastructures on EOSC can be done by using one infrastructure without considering their technical specifications. These infrastructures will provide different functionalities; some of them are related to the data storage, some are concerning computing. Among the infrastructures of computing, we will have some for HTC and others for HPC, and so on. But the interoperability should ensure that one can manage a workflow (and importantly the data access) without technical dependencies across the various infrastructures involved.

5.1. How to verify interoperability

Due to the large number of e-infrastructures, services, interfaces involved in the context of the EOSC, manual testing and verification of interoperability would be extremely time consuming, cost intensive, and a highly repetitive task. This is the reason why the automated interoperability testing has to be put in place when possible. This automation could be performed from a set of tests, which are representative for the main usages of the e-infrastructures.

Different workflows, based on real applications, can be considered as candidates to be this set of tests because they have to be run on testbeds which imply several types of e-infrastructures and service providers. Pilots identified in the EOSCPilot are appropriate candidates to build this set of validation workflows.

The e-infrastructures are complex, so a hierarchical and functional test suite will be required to test all workflows . The first step would be to identify these workflows. Classification of these workflows has to be functional, data management oriented, compute oriented, and services oriented but also in functional in the level of the complexity covered. Workflows which require many types (HTC, HPC, Cloud) of computing

infrastructures, many infrastructure of data storage and have an intense and divers usage of the AAI, has to be considered more complex than a workflow which imply only one infrastructure.

A detailed set of verification criteria will be included in the final EOSC architecture recommendation (D 6.8).

6. THE ROAD AHEAD

In order to derive a functional EOSC architecture within a rather short timescale of a few years, several actions have to be taken.

6.1. Next steps within the EOSC pilot

The EOSC pilot is the framework to find solutions to the interoperability issues, and a testing ground for these solutions. Main next steps towards an efficient EOSC architecture are:

- Gather experience from the Science Demonstrators on which solutions work – and which don't.
- Discuss with e-infrastructures and communities that have found solutions for interoperability challenges. Are these solutions scalable, applicable to a wider range of communities and countries? What are the costs connected to these solutions?
- Which existing training programs appear most successful in transferring knowledge and can they be used and extended in the context of EOSC's interoperability challenge?
- How do service interoperability aspects connect to the challenges described here?

The next step here within the EOSC pilot will be the delivery of the Initial EOSC Service Architecture (D5.1) by the end of this year. WP5 and WP6 will then work together over the course of the next year in order to achieve the 'Final EOSC Architecture' (D5.4 / D 6.8) towards the end of 2018.

6.2. Next steps beyond the EOSC pilot

The EOSC architecture will require a strong commitment from all parties involved, e.g. e-infrastructures, service providers, research communities, funding agencies, political leaders. In order to ensure that the recommendations of the EOSC pilot concerning the interoperability will fall onto fruitful soil, several actions should be taken:

- Sensitize stakeholders for e-infrastructure interoperability issues, especially in the context of widening the landscape towards the EOSC
- Communicate the benefits of overall e-infrastructure interoperability to the stakeholders – it is a challenge worth taking on!
- Support and start initiatives on the national level to solve interoperability issues here.

7. CONCLUSION

Defining the methods, means and architecture that will ensure the interoperability of e-infrastructures, which will be federated within the EOSC, cannot be done easily, particularly within the early stages of the EOSC Pilot project.

Based on the infrastructure gap analysis that has been performed in the context of the EOSCPilot project, some approaches have been identified in order to take into account the interoperability aspect of infrastructures.

During the EOSCPilot project, new interoperability gaps will appear and some that are apparent today will no longer be valid. Therefore, the different methods and approaches, which could be implemented to ensure the interoperability of infrastructure, will have to evolve. Deliverable D6.8 proposed for the end of the EOSC pilot project will thus include updates with respect to this document.