D6.1: e-Infrastructure Gap Analysis

Author(s)	Geneviève Romier, CNRS Xavier Jeannin, RENATER Mauro Campanella, GARR Cristina Duma, INFN
Status	Final
Version	V1.0
Date	30/06/2017

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:

The objective of this deliverable is to perform a gap analysis on technical and political barriers that prevent the interconnection of e-infrastructures and the application of the FAIR principles with the aim to guide the design of an architecture that will allow overcoming these gaps. The work is mainly based on a questionnaire sent to a set of e-infrastructures, to the Science Demonstrators and communities related to the long tail of Science. The analysis includes the technical aspects (authentication, authorisation, data, compute, network, core infrastructures and workflows management systems), the political, social and cultural aspects (access policies, knowledge of the e-infrastructures landscape, adoption) and, based on the analysis done, a first list of initiatives and projects to collaborate with.

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-D6.1	
Deliverable lead	**CNRS**
Related work package	**WP6**
Author(s)	Geneviève Romier, CNRS Xavier Jeannin, RENATER Mauro Campanella, GARR Cristina Duma, INFN
Contributor(s)	Vincent Breton, Eric Fede, Violaine Louvet, Volker Beckmann (CNRS), Sibel Yasar, Frank Schlünzen, Patrick Fuhrmann, (DESY), Licia Florio, Tryfon Chiotis (GÉANT), Afrodite Sevasti (GRNET)
Due date	**30/06/2017**
Actual submission date	**30/06/2017**
Reviewed by	**Cees Hof (DANS)** **Jorge Sanchez (Corallia Clusters Initiative)** **Ludek Matyska (CESNET)** **Pierre-Etienne Macchi (CNRS)**
Approved by	**Brian Matthews**
Start date of Project	**01/01/2017**
Duration	**24 months**

Versioning and contribution history

Version	Date	Authors	Notes
0.1	18/05/2017	Geneviève Romier (CNRS)	Table of Content internally approved
	23/05/2017	Geneviève Romier	Start of content in the wiki

	31/05/2017	Violaine Louvet, CNRS	Participation to the national or regional resources providers part
	12/06/2017	Vincent Breton, CNRS	Revision of the political, social and cultural aspects of the analysis
	14/06/2017	Cristina Duma, INFN	Indigo Datacloud description
	14/06/2017	Xavier Jeannin, RENATER Mauro Campanella, GARR	The interconnection (GÉANT and NRENs) Network analysis
	14/06/2017	Sibel Yasar, Frank Schlünzen, Patrick Fuhrmann, DESY	Corrections
	14/06/207	Geneviève Romier	Draft version in the deliverable template
	14/06/207	Eric Fede, CNRS	Global revision
	15/06/2017	Cristina Duma, INFN	Technical annexes
	16/06/2017	Geneviève Romier	Last draft version
	16/06/2017	Licia Florio, GÉANT	Additions
	16/06/2017	Geneviève Romier, CNRS	GÉANT description
	19/06/2017	Volker Beckmann, CNRS	Global revision
	21/06/2017	Volker Beckmann, CNRS	Global revision
	23/06/2017	Afrodite Sevasti, GRNET	Global revision and Annex part A.6
	27/06/2017	Geneviève Romier, CNRS	Summarized answers to the questionnaire previously in sections 2 & 3 transferred in annex A
	29/06/2017	Geneviève Romier, CNRS	Corrections following the reviewers comments and
	30/06/2017	Eric Fede, CNRS	Comments
	30/06/2017	Volker Beckmann, CNRS	Final review

Copyright notice: This work is licensed under the Creative Commons CC-BY 4.0 license. To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0.

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

LIST OF FIGURES

LIST OF TABLES

EXECUTIVE SUMMARY

The objectives of this "e-Infrastructure Gap Analysis" are to perform a gap analysis of the current issues preventing exploitation and usage of existing e-infrastructures and distributed resources, on technical and political barriers that prevent the interconnection of e-infrastructures and the application of the FAIR principles with the aim to provide architectures that will allow overcoming these gaps. In this document, task 6.1 of the interoperability work package provide an inventory of best practices, and provide feedback to the Services work package (WP5) for the definition of services in the EOSC portfolio that are related to interoperability across e-infrastructures.

These are the key results at technical, political, social or cultural levels:

1. The needs of authentication and authorisation at a global level as well as between any two infrastructures show that all the e-infrastructures landscape could benefit from a global Authentication and Authorization Infrastructure (AAI). This AAI should be widely and consistently available at international, national and even local levels and fully federated between them. The approach should encompass all existing AAIs and take into account the various legal aspects and the policies required by research communities. This will help to work across infrastructures despite the lack of common access policies. The work of the AARC (Authentication and Authorisation for Research and Collaboration[1]) projects will provide the framework and architecture for AAI providers to benefit to all the scientific communities. Implementation of this architecture by AAI providers will be a keystone for the EOSC success.
2. Efficient data transfers and network service delivery are also key components for the success of the EOSC. The generic and specialized requirements of EOSC with regards to network connectivity and data transfers need to be thoroughly analysed with respect to the current offerings of GÉANT and the NRENs. Last mile issues, namely the network infrastructure between end user sites or data/computing/storage facilities and the national/pan-European network infrastructures are also of paramount importance. The overarching orchestration of core networks, last mile network infrastructures and end sites towards reliable, end-to-end network service delivery and lifecycle management requires a close collaboration of the involved operators.
3. The variety of the data and computing services and the diversity of providers have to be taken into account. The needs of new services such as data mining and container management have appeared. Accounting and traceability are necessary. Certain communities require web portals and user-friendly workload managers.
4. A multidisciplinary mutualised (or discipline agnostic) space in the EOSC could allow the EOSC to be a benefit to all researchers in Europe including those working on the long tail of science.
5. The EOSC success will rely on technical realisations and efficient collaborations. It will also rely on making scientists in Europe aware of its existence and benefits, especially in terms of services and infrastructures provided by the EOSC. In this context, the adoption of the EOSC by the scientific communities is vital. This adoption requires the consideration of the users' needs such as appropriate approach to provision of resources and services, security and privacy for example. Here key factors are communication, dissemination, and training of the infrastructure providers and scientists. This has to be combined with the suitable conditions to foster experience and expertise sharing among scientists, include user support and infrastructure technical experts.

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

1. SECTION 1 - INTRODUCTION

According to the proposal the objective of this deliverable is to perform a gap analysis on technical, political social or cultural barriers that prevent the interconnection of e-infrastructures and the application of the FAIR principles with the aim to provide architecture that will allow overcoming these gaps. In this document, task 6.1 will provide an inventory of best practices, and provide feedback to the Services workpackage (WP5) for the definition of services in the EOSC portfolio that are related to interoperability across e-infrastructures.

The outline and the methodology were defined during the WP6 kick-off meeting[2] in Amsterdam on February 20/21, 2017. It was decided to send a survey to the contact persons ("shepherds") of the 5 science demonstrators and to a number of e-infrastructures and scientific communities. The survey document has been iterated among the colleagues involved in the analysis and discussed with a WP5 representative. The survey was performed in two versions: one for the research communities and one for the e-infrastructures or centres that provide computing or storage resources. The research communities' survey version[3] was sent to the shepherds of the 5 science demonstrators and the IndexMed[4], Eurofidai[5], and Phenome[6] consortia that may be considered as part of the long tail of science. The e-infrastructure survey version[7] was sent to representatives of PRACE, EUDAT, EGI. It was also sent to different Grid and HPC computing centres. The list is available in Annex

The questionnaire and the received answers of the infrastructures[8] and the Science demonstrators[9] are available on the project repository. More than 20 answers have been received that represent a broad range of cases. 8 answers were received from the Italian Grid sites, France Grilles sites and French regional computing centres contacted via their mailing lists. The questionnaires sent directly to known persons have a better level of answer (10/13).

In addition to the questionnaire, the T6.1 teamwork relies on personal exchanges or discussions during meetings and available information and documents of the e-infrastructures and sites and more specifically on documents elaborated by different projects or communities such as INDICO Data Cloud or ELIXIR-EXCELERATE deliverables.

It also takes into account other important advices such as the "First report and recommendations of the Commission High Level Expert Group on the European Open Science Cloud" drafted by the Commission High Level Expert Group on the European Open Science Cloud and the e-IRG Roadmap. The interconnection part has been drafted by the GÉANT partners according to their knowledge of the needs of the different stakeholders.

The deliverable is organised in three main sections and annexes.

This introduction is the SECTION 1 of this document. SECTION 2 presents sites, e-infrastructures (FR, IT, DE), European e-infrastructures (EGI, EUDAT) and the interconnection (GÉANT and NRENs). SECTION 3 is dedicated to the Science demonstrators and user communities' inputs. The analysis including the current

[2] https://repository.eoscpilot.eu/index.php/apps/files/?dir=/WP6%20-%20EOSC%20Interoperability/Meetings/WP6%20kick-off%20February%202017&fileid=2442

[3] https://repository.eoscpilot.eu/remote.php/webdav/WP6 - EOSC Interoperability/Task 6.1 e-infrastructure gap analysis %26 interoperability architecture/Input-for-EOSCpilot-gap-analysis-science-demonstrators.docx

[4] http://www.indexmed.eu

[5] https://www.eurofidai.org/en

[6] https://www.phenome-fppn.fr/phenome_eng/

[7] https://repository.eoscpilot.eu/remote.php/webdav/WP6 - EOSC Interoperability/Task 6.1 e-infrastructure gap analysis %26 interoperability architecture/Input-sites-and-e-infras-for-EOSCpilot-gap-analysis-v3.2.docx

[8] https://repository.eoscpilot.eu/index.php/apps/files?dir=/WP6%20-%20EOSC%20Interoperability/Task%206.1%20e-infrastructure%20gap%20analysis%20%26%20interoperability%20architecture/Input-infrastructures

[9] https://repository.eoscpilot.eu/index.php/apps/files?dir=/WP6%20-%20EOSC%20Interoperability/Task%206.1%20e-infrastructure%20gap%20analysis%20%26%20interoperability%20architecture/Input-science-demonstrators

situation and context is presented in the SECTION 4 and concludes with the main findings. The SECTION 5 summarises the conclusions of this study.

Annex A contains 4 subsections. A.1 is the list of organisations and infrastructures contacted for the survey; A2 is the description of the main European e-infrastructures; A.3 is the description of AAI technical solutions deployed in the main e-infrastructures; A.4 is the description of storage and data management solutions deployed in the main e-infrastructures; A.5 is the description of computing solutions deployed in the main e-infrastructures; A.6 is the description of GÉANT Connectivity / Network management & monitoring services; Annex 7 is the description of the inputs of the regional and national resources providers; Annex 8 is the description of the input of the European e-infrastructures ; Annex B contains a glossary

2. SITES, E-INFRASTRUCTURES (FR, IT, DE, CH), EUROPEAN E-INFRASTRUCTURES (EGI, EUDAT, PRACE) AND THE INTERCONNECTION (GÉANT AND NRENS)

2.1. National or regional resources providers

This section presents the national or regional resources providers' inputs. The input comes from the questionnaires and may be completed with interviews or documents. National or regional resource providers include many types of infrastructures:

- High-Performance Computing (HPC) and High-Throughput Computing (HTC) centres as well as data centres.
- Resources centres with a large spectra of resources (HPC, HTC, cloud computing, storage of different types) that offer many services to their users generally in their region, university campus or town and for certain communities such as WLCG for example. Several of them act as Tier 0, Tier 1, 2 or 3 data centres or computing centres for international scientific experiences or in large pan-European federated e-infrastructures such as PRACE or WLCG, for example.
- Federations of centres that offer services built on top of mutualised resources. The mutualisation may concern a part or all resources and may concern one or several services. The federation may be at any level of geographic area.

In this analysis as explained in the introduction we sent our questionnaire to a set of centres and e-infrastructures that are listed in Annex 1 and we received inputs from main centres such as IDRIS (HPC), CC-IN2P3 (HTC) in France, Jülich Supercomputing, DESY and KIT in Germany; a Swiss Data Centre: Centre Common Data Centre Infrastructure (CDCI) for Astronomy Astroparticle Physics and Cosmology at Geneva University; regional centres such as PSMN ENS Lyon, IN2P3-IRES (IPHC) and the *Mésocentre Clermont-Auvergne* (MCA) in France, the two INFN centres of Rome and Padova in Italy; regional federated centres such as GRICAD in Grenoble or joint answer from Strasbourg (University and IPHC); France Grilles as the French National Grid Initiative (NGI).

The centres have different geographical scopes and different implications regarding the user communities depending on their level in the Tier hierarchy in the Research Infrastructures organisations for example. This explains a certain inconsistency that can appear at the first reading of all their answers. This large variety of answers gives in fact a view of the large panel that may contribute to the EOSC. The national and regional centres that ware solicited in this analysis come from only four countries and represent a tiny fraction of the actual heterogeneity of the whole European landscape.

In order to avoid a too long deliverable, the inputs are summarized in the Annex A 7 of this document. All original answers to the questionnaire are stored in the projects file repository[10]. Citations of the questionnaire answers are included in the text of this document.

2.2. European e-infrastructures

The inputs presented in this chapter originate from the answers to the questionnaire filled in by the EGI and EUDAT e-infrastructures representatives. The footprint of these e-infrastructures is pan-European. They federate centres and resources providers with national or regional scopes. The main European e-infrastructures are described in Annex A.1 and their inputs are summarized in the Annex A7.

[10] https://repository.eoscpilot.eu/index.php/apps/files/?dir=/WP6%20-%20EOSC%20Interoperability/Task%206.1%20e-infrastructure%20gap%20analysis%20%26%20interoperability%20architecture/Input-infrastructures&fileid=3895

2.3. The interconnection (GÉANT and NRENs)

In Europe, the National Research and Education Networks and their main interconnecting network GÉANT offer an excellent connectivity footprint between user communities, data/computing e-Infrastructures and the Internet.

The GÉANT/NREN networks are engineered to offer high-speed connectivity and a set of above-the-net services, in particular for trust and identity such as eduGAIN or roaming access such as eduroam.

Connectivity of R&E users to the GÉANT/NREN footprint differs from country to country, from direct access of user facilities (e.g. campus) to the local NREN to access via one or more regional networks. Peerings between interconnected networks are established to ensure that there is a highly efficient network infrastructure across Europe.

One of the EOSCPilot project objectives is to demonstrate the interoperability of data cloud services and infrastructures and their benefits in a number of scientific domains. As cloud services and infrastructures are by nature distributed over Europe, network connectivity between the sites that provide and consume the cloud services is therefore a foundation layer on which EOSC is to be built. The stability, robustness and performance of the network are key elements for the EOSC success.

2.3.1. **Site interconnection**

From the network point of view, the main requirement is the transfer of data between the EOSC sites and widely distributed users. The sites can be user/researcher sites, instrument sites or data/cloud/storage infrastructures, academic cloud or commercial cloud.. The combination of the NREN and GÉANT network infrastructures provides the core, high quality, efficient interconnect. From the NREN edge all the way to the different EOSC sites the so-called last mile for network connectivity exists, through regional, campus or other facility network infrastructures. In certain cases, NREN and GÉANT also interconnect directly to commercial cloud providers via L2 or L3 peerings for more efficient connectivity compared to the one through public Internet providers.

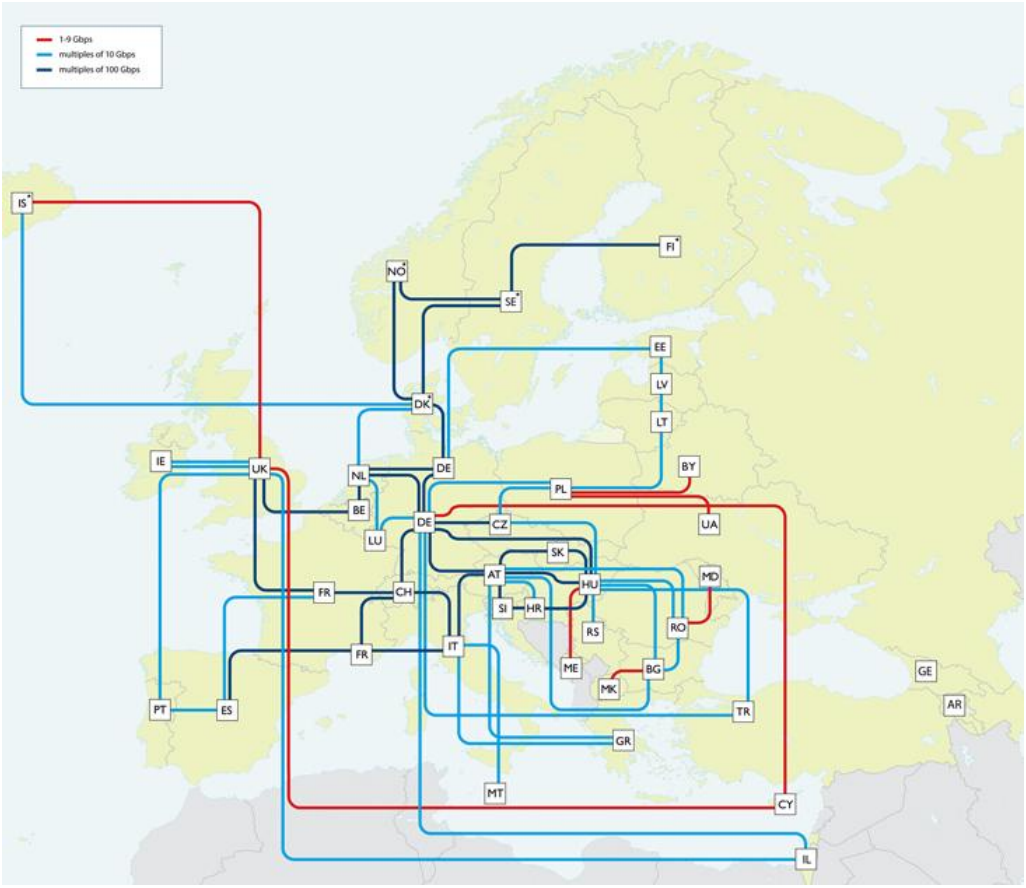

Figure 1: GÉANT topology map

By default, EOSC site interconnections are expected to be implementing using public IP/Layer 3 connectivity services such as GÉANT IP[11]. In these cases, EOSC sites need to support public IP address endpoints advertised by the local (campus/facility) network to the upstream regional/NREN networks. Internal networking at the EOSC site and configuration of the site gateway to make sure data transfers are possible between local EOSC site endpoints and remote EOSC site endpoints are the minimum requirements. However, general-purpose IP transit between EOSC sites will not always meet the data transport needs of EOSC-related use cases.

Such use cases might include bulk data transfers with guaranteed capacity, or within guaranteed time limits, configurations so that computing/storage resources between remote sites are combined in a single data processing/storage pool or cases, multi-point data transfers or cases when data should not be transported over the public networks together with commodity traffic (e.g. for reasons of security). For such cases, where performance, stability and security are required, more advanced service offerings are available by GÉANT/NRENs, such as GÉANT point-to-point services[12] and Virtual Private Network (VPN) Services[13]. EOSC is expected to introduce additional requirements to the existing advanced services offered by GÉANT/NRENS as briefly described in the following sections.

2.3.2. **Virtual network infrastructure delivery lead-time**

The GÉANT-NREN network infrastructures offer various options to create virtual private networks. VPN instances can be realized at the optical layer, Ethernet or IP according to users' requirements. While advanced services coverage throughout Europe by GÉANT/NRENs is quite extended (e.g. 16 countries support the GÉANT Multi-Domain Virtual Private Network offering at the moment of writing) reachability of

[11] https://www.geant.org/Services/Connectivity_and_network/Pages/GEANT_IP.aspx
[12] https://www.geant.org/Services/Connectivity_and_network/Pages/GEANT_Point-to-Point.aspx
[13] https://www.geant.org/Services/Connectivity_and_network/Pages/VPN_Services.aspx

VPN service access points from known and future EOSC sites depends upon the existence of physical footprint as well as support for VPN technologies at the last mile (from the NREN edge up to the end site).

When a user books digital resources in a cloud (storage, computing, …), these resources are available in a matter of a few minutes. The usage of a virtual network infrastructure (a VPN for instance) in order to interconnect the sites is an improvement regarding performance, stability and security for the users. The delivery lead-time of this virtual network infrastructure must be also short.

2.3.3. **Resiliency and availability**

For access to data/cloud/storage resources and sites, the user expectations for high availability of network connectivity services are very important. Throughout the GÉANT/NREN footprint, availability of both the general-purpose IP transit services but also specialized point-to-point or VPN services is exceptional. This is ensured by infrastructure and service engineering principles (such as redundancy and fast reroute)

2.3.4. **Data transport performance**

EOSC use case requirements for data transports related to performance can vary a lot. In some cases, very high throughput requirements are expected while in other cases the requirement might be for very short latency. Connection to cloud facilities through the public Internet may lead to insufficient performance even for EOSC use cases without special performance requirements.

By utilizing the capabilities and specialized service offerings of GÉANT/NRENs, as already mentioned, it is possible to meet advanced performance requirements. For example, GÉANT already provides access to commercial clouds in the *Helix Nebula Science Cloud* project and the benefit of this type of connection rather than using the public Internet is recognized by end-users who are partners of this project. However, the complexity, time scales and cost for specialized service delivery will differ. Such specific implications and requirements from the network in the EOSC context need to be further elaborated.

2.3.5. **Reduce cost of data transfers**

Utilizing commercial cloud provider offerings in the context of the EOSC can incur significant costs. Such costs include, apart from the computing/storage resource reservations, the data transport costs between user locations and cloud provider facilities. GÉANT/NRENs interconnect with commercial cloud providers with specific agreements, which waive to a large extent the cost of data transfers for end users.

2.3.6. **Security**

Finally, storage and transferring of data across EOSC sites may impose security requirements, such as the deployment of dedicated and costly hardware at EOSC sites (e.g. firewalls). Specialized network service offerings by the GÉANT/NRENs are able to isolate EOSC traffic when needed from the rest of the Internet traffic. Thus, an EOSC site can minimize its security CAPEX (firewall expenses) and benefit from better performance, as the traffic is not inspected by an additional hardware element.

2.4. Projects, standards and Open Source technologies developed to facilitate the interoperability of services

Here are the descriptions of outputs of projects, standards and Open Source technologies that may be of interest to solve interoperability issues. Several of these outputs or technologies may be not disseminated enough or are too recent to be largely adopted. The goal of this section is to avoid re-inventing the wheel. The projects are presented by alphabetical order.

2.4.1. AARC - AARC2 - Authentication and Authorisation for Research and Collaboration

AARC[14] was an EC funded project that brings together 20 different partners from among National Research and Education Networks (NRENs) organisations, e-Infrastructures service providers and libraries.

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

AARC champions federated access and works together with national identify federations, research infrastructures, e-infrastructures and libraries to identify building blocks and policy best practices needed to implement inter-operable authentication and authorisation infrastructures (AAIs). Research communities' use-cases drive the AARC work to test technical and policy components and their integration into production research and e-infrastructures.

AARC work will help avoid a future in which different e-Infrastructures and (new) research collaborations develop and operated independent (and not inter-operable) AAIs. For that reason it is really important for the EOSC to work together with AARC-AARC2. This will be facilitated by the fact that several partners of EOSCpilot are also involved in AARC-AARC2.

During the last two years, AARC has worked with the research and education community to design and promote an integrated authentication and authorization framework that enables international research collaborations to adopt federated access solutions in an interoperable and sustainable manner. The DJRA1.1 deliverable (Analysis of user- community requirements) contains an e-infrastructures requirements analysis that can be a basis for the EOSC. The MJRA1.1 milestone (Existing AAIs and available technologies for federated access) presents the existing standards and technologies in this domain and a comparison of all those elements.

Working closely with research and e-infrastructure providers, and scientific communities, AARC has designed an AAI architecture that enables the eScience projects, research and e-Infrastructures utilize federated access and reap the full benefits of eduGAIN, while providing all the necessary, higher-level architecture that is required by international research collaborations. The AARC Blueprint Architecture (BPA)[15] provides set of interoperable architectural building blocks for software architects and technical decision makers, who are designing and implementing access management solutions for international research collaborations. A series of guidelines linked in the blueprint page[16] go along with the blueprint that will be the starting point of the AARC2 project.

Along with the Blueprint Architecture, AARC has provided technical recommendations and guidelines for implementers and a set of policy frameworks and sustainability models that can support the production operation and interoperability of the initial AAI solutions that are already being deployed by research and e-Infrastructure providers. Furthermore, through the continuous engagement with the scientific communities via the training & outreach activities, but also through the focused technical piloting activities, AARC has attracted critical mass for the adoption of federated access.

AARC2 is an EC funded successor project of AARC and it will continue the work of AARC. AARC2 will work closely with infrastructure and technology providers to ensure that they have the necessary architectural and policy building blocks available that are required to implement secure, sustainable, scalable and interoperable AAI solutions for international research collaborations. AARC2 will bring even more pilot activities with many communities and infrastructures and will facilitate the production deployment and the sustainable operation of such solutions by the infrastructure providers for the benefit of the research collaborations.

2.4.2. INDIGO-DataCloud (INtegrating Distributed data Infrastructures for Global ExplOitation)

In Cloud computing, both the public and private sectors are already offering Cloud resources as IaaS (Infrastructure as a Service). However, there are numerous areas of interest to scientific communities where Cloud Computing uptake is currently lacking, especially at the PaaS (Platform as a Service) and SaaS (Software as a Service) levels. In this context, INDIGO-DataCloud (INtegrating Distributed data Infrastructures for Global ExplOitation)[17], a project funded under the Horizon 2020 framework program of the European Union, aims at developing a data & computing platform targeted at scientific communities, deployable on multiple hardware, and provisioned over hybrid e-Infrastructures.

[15] https://aarc-project.eu/wp-content/uploads/2017/04/AARC-BPA-2017.pdf
[16] https://aarc-project.eu/blueprint-architecture/
[17] https://www.indigo-datacloud.eu/

The main challenges that the projects is trying to address are:

- **Orchestrating and federating Cloud, Grid and HPC** [public or private] resources
- Overcoming the **current barriers** limiting the **adoption of PaaS solutions** and of **virtualized cloud resources** in **large data centres**
- Facilitating **flexible data sharing & access** between group members & infrastructures
- **Managing dynamic and complex workflows for scientific data analysis** & combining data from multiple sources and storage locations
- Avoiding software and vendor lock-in
- **Supporting federated identities** and providing privacy and distributed authorisation in open cloud platforms
- Exploiting distributed computing and storage resources through **transparent network interconnections**.

In this context the INDIGO - DataCloud offering is to develop a **data/computing platform targeting scientific communities, deployable on multiple hardware and provisioned over hybrid (private or public) e-infrastructures** by using an user-driven approach - a structured interaction with the users**..**

- Developing and delivering of **software to simplify the execution of applications on Cloud and Grid based infrastructures,** as well as on HTC and HPC clusters, and,
- **Extending existing PaaS solutions integrating services provided by existing e-infrastructures** (e.g. EGI, EUDAT, PRACE and Helix Nebula)

Over ten scientific communities are involved in the technical requirements collection for the new services belonging to four main different domains: Biological & medical science, Social science & humanities, Environmental and earth science, Astrophysics

The first software release of the project, code named MidnightBlue, comes after an initial phase of requirement gatherings which involved several European scientific collaborations in areas as diverse as structural biology, earth sciences, physics, bioinformatics, cultural heritage, astrophysics, life sciences, climatology, etc. This resulted in the development of many software components addressing existing technical gaps linked to easy and optimal usage of distributed data and compute resources. These components are now released into a consistent and modular suite, offered as a contribution toward the definition and implementation of an efficient European Open Science Cloud. The first INDIGO-DataCloud release provides **open source components** for:

- **Data centre solutions**, allowing data and compute resource centres to increase efficiency and services for customers.
- **Data solutions**, offering advanced access to distributed data.
- **Automated solutions**, allowing users to easily specify and deploy complex data and compute resource requirements.
- **User-level solutions**, integrating scientific applications in programmable front-ends and in mobile applications.

Key technical highlights:

The Data Centre. INDIGO is providing many new features/services for resource centres:

- *Improved scheduling* for allocation of resources by the popular open source Cloud platforms. OpenStack and OpenNebula. This provides both better scheduling algorithms and support for spot-instances.
- Support for *improved IaaS resource orchestration capabilities* using standards orchestration engines through the use of the TOSCA standard, for both OpenStack and OpenNebula.
- *Improved QoS capabilities of storage resources* for better support of high-level storage

requirements, such as flexible allocation of disk or tape storage space and support for data life cycle.
- Improved and *transparent support for Docker containers*. This includes for example the introduction of native container support in OpenNebula.

The Data Services. INDIGO provides a complete set of data-related features that includes:

- *Distributed Data Federation* through several protocols, in order to support both legacy application and advanced standard interfaces such as CDMI or just simple web interfaces.
- The possibility to *federate diverse storage technologies* (such as POSIX, Object Storage, CEPH, etc) in a seamless way, letting users exploit data and storage resources wherever they are available.

Automated Solutions. INDIGO provides a rich set of high-level automated functionalities. Some of the most innovative are:

- Improved capabilities in the *geographical exploitation of Cloud resources*. End users need not know where resources are located, because the INDIGO PaaS layer hides the complexity of both scheduling and brokering.
- *Standard interface to access PaaS services*. INDIGO uses the TOSCA standard to hide the difference on the different way of implementing services at the PaaS level.
- Support for *data requirements in Cloud resource allocations*: computational resources can be requested and allocated where data is stored.
- Integrated use of resources coming from *both public and private Cloud* infrastructures.
- Deployment, monitoring and *automatic scalability of existing applications*.
- Integrated support for *high-performance Big Data analytics*.
- Support for *dynamic and elastic clusters of resources*. HTCondor, Torque and Mesos cluster are supported.

High-level user oriented services. Researchers and data managers are able to access resources through:

- *Toolkits* (libraries) allowing usage of the INDIGO platform from Scientific Gateways and desktop applications.
- An open source *Mobile Application Toolkit* for the iOS and Android platforms, serving as the base for the development of Mobile Apps.
- *User-friendly* front ends for building *programmable, general-purpose multi-domain Science Gateways*.

All the INDIGO components are integrated into a comprehensive **Authentication and Authorization Architecture**, with support for *user authentication through multiple methods* (SAML, OpenID Connect and X.509), support for *distributed authorization policies* and a *Token Translation Service*, creating credentials for services that do not natively support OpenID Connect.

The **INDIGO-DataCloud software** is released[18] under the Apache 2.0 software license and can be deployed on both public and private Cloud infrastructures.

The second, and last major software release, **ElectricIndigo** builds and expands on the first version of the software. In this respect, it enhances stability**,** adding more **programmability, scalability, automation and flexible network management**, to help resource providers and scientific communities address challenging problems and deliver new services:

- **Application-level Interfaces for Cloud Providers and Automated Service Composition:** Easily port applications to public and private Clouds using open programmable interfaces, user-level containers, and standards-based languages to automate definition, composition and instantiation of complex set-ups.

[18] https://caifti.gitbooks.io/indigo-datacloud-releases/content/

- **Flexible Identity and Access Management**: Manage access and policies to distributed resources using multiple methods such as OpenID-Connect, SAML, X.509 digital certificates, through programmable interfaces and web front-ends.
- **Data Management and Data Analytics Solutions**: Distribute and access data through multiple providers via virtual file systems and automated replication and caching, exploiting scalable, high-performance data mining and analytics.
- **Programmable Web Portals, Mobile Applications:** Create and interface web portals or mobile apps, exploiting distributed data as well as compute resources located in public and private Cloud infrastructures.
- **Enhanced and Scalable Services for Data Centres and Resource Providers**: Increase the efficiency of existing Cloud infrastructures based on OpenStack or OpenNebula through advanced scheduling, flexible cloud / batch management, network orchestration and interfacing of high-level Cloud services to existing storage systems.

New to ElectricIndigo: **ElectricIndigo** includes more than 40 modular components, distributed via 170 software packages and 50 ready-to-use Docker containers, adding the following **new features** to the previous INDIGO release:

- FairShare Scheduler for OpenNebula
- Network Orchestrator Wrapper (NOW) for Intra-site Networking Management in OpenNebula
- Command Line Interface for submitting TOSCA Templates to the INDIGO PaaS

3. SCIENCE DEMONSTRATORS AND USER COMMUNITIES INPUTS

This section is dedicated to the input of the Science Demonstrators of the EOSCpilot project and a few scientific communities more representing the long tail of science. In order to avoid a too long deliverable, the inputs are summarized in this document. All original answers to the questionnaire are stored in the projects file repository[19]. Relevant citations of the questionnaire answers are included in the text of this document.

3.1. EOSCPilot Science demonstrators

The input of the science demonstrators come from the discussions during the WP6 kick-off meeting in Amsterdam in February 2017 and the answers to the questionnaire. They can be completed with the help of the demonstrators' descriptions. We received answers from 4 of them: High Energy Physics - WLCG, Social Sciences – TEXTCROWD, Life Sciences - Pan-Cancer and Physics - The photon-neutron

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.

Five Science Demonstrators started at the project outset (listed below), with more to be included through EOSCpilot Open Calls for Science Demonstrators. In this deliverable the first set of demonstrators only is taken into account (the others are currently not selected):

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

3.1.1. High Energy Physics - WLCG

During the WP6 kick-off meeting the representatives of this demonstrator explained they will need to see which resources are used first, which e-infrastructure provider will be chosen and they will provide feedback to WP6 and also ask for support if/when needed. For this demonstrator authentication and authorization may not be very relevant issue as data are openly available (public, anonymous access).It is possible new gaps will surface later.

In the questionnaire this demonstrator expressed only technical needs:

They foresee using EUDAT services such as B2SAFE, B2HANDLE and B2SHARE as well as EGI services such as CVMFS (and CernVM itself to run in "the Cloud"). They expect services that are currently operated by EGI and EUDAT to "interoperate".

They made a comment concerning the questionnaire, which they don't really understand.

[19] https://repository.eoscpilot.eu/index.php/apps/files?dir=/WP6%20-%20EOSC%20Interoperability/Task%206.1%20e-infrastructure%20gap%20analysis%20%26%20interoperability%20architecture/Input-science-demonstrators

3.1.2. The Social Sciences – TEXTCROWD science demonstrator

TEXTCROWD announced during the joined WP4-WP6 meeting in February 2017 in Amsterdam that it has no interoperability needs.

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

They confirmed this position answering the questionnaire.

3.1.3. Life Sciences - Pan-Cancer

Their main interoperability needs are:

- Uniform APIs across different providers for all the exposed services
- Compliance of services with the GA4GH APIs reference implementation (http://genomicsandhealth.org/)

They provided the ELIXIR compute platform deliverable as a reference.

At the technical level the main gap (mandatory solution for the use case) is that data set sizes exceed capabilities of current infrastructures, thus it is imperative to enable interoperable operation to maintain the ability to analyse data sets required by today's and future scientific studies in the field of cancer research.

In addition at the political level the geographical dispersion of the data implies locale specific requirements that need to be bridged to enable global-scope studies like PanCancer (mandatory solution for the use case).

The interoperability solutions this demonstrator is currently testing/building for their user community is: support for federated analysis scheduling and management is being built to ease the burden of cross-site deployment.

The solutions they recommend: Currently implemented support for cross-cloud operation by adopting frameworks like Terraform and Saltstack that allow deployment to a multitude of technical environments. They classify the impact as mandatory for the use case.

3.1.4. Physics - The photon-neutron

Their main interoperability challenges are:

• at a technical level, identified as important benefit to the use case, data interoperability at the binary level is to extend implemented through the use of HDF5 (https://www.hdfgroup.org/) and NeXus (http://www.nexusformat.org/). However, the data interoperability is still rather limited. This could be greatly improved through interoperability service, e.g. a kind of data type registry with machine digestible schemata. In this particular case, the HDF Product Designer appears as the most suitable implementation.

• The Photon and Neutron science applications are very diverse requiring a wide spectrum of different resources. Dataset sizes range from a few MB to hundreds of terabytes per dataset. The processes can range from very low CPU requirements, over embarrassingly parallel to massively parallel applications, such that the most appropriate platform varies from experiment to experiment. As such, infrastructure interoperability services would be highly beneficial for the Photon and Neutron use case.

• Accounting data would be helpful. In most cases compute resources would be provided by the user facilities in form of "Data analysis as a Service". Users might or might not be charged for such services. In any case, this would require to be able to roughly predict associated costs for a wide spectrum of application and use case scenarios.

A social gap is that currently the use of cloud resources for data analysis in the Photon Science domain is very close to zero.

In addition they identified a lack of technical skills in the community:

• For some scientific applications direct incorporation of cloud APIs in the code might be an interesting topic, but scientific application developers are certainly lacking the knowledge how to do this (at all or efficiently). Might be an interesting topic for hands-on materials or tutorials (if feasible at all).

They classified this as useful.

The interoperability solutions they are currently testing/building for their users are:

- HDF Product Designer (https://wiki.earthdata.nasa.gov/display/HPD/HDF+Product +Designer)
- The ICAT instances at the different facilities are loosely federated, with UmbrellaID (https://umbrellaid.org/) as the envisaged AAI supplier. However, the Indigo AAI solution should be evaluated.
- H5serv (https://h5serv.readthedocs.io/en/latest/) for convenient access to HDF5 containerized data without major code changes.

3.2. Local and regional users of the infrastructures

Here is described other inputs related to other users. This reflects an input more related to the long tail of science, either not related to a specific discipline (discipline agnostic) or less organised communities than the project Science Demonstrators (IndexMeed, Phenome and CDCI).

Researchers working tail of science researchers may do so not only on their own or in a single laboratory. They may be skilled in advanced computing. But often they are not members of a Research Infrastructure and it is not easy for them either to work with a workflow that needs National HPC resources, large datasets storage and regional HPC resources either to prepare their production or to visualise their results for example. Another situation is the need to share data or software between different teams that are distributed in different regions, countries of organisms. Emerging consortia are also part of the long tail of science. The inputs below present several such examples.

3.2.1. IndexMeed

Indexing for Mining Ecological and Environmental Data (IndexMeed)[20] is a scientific consortium in the Ecological and Environmental domain. IndexMeed's aim is to index biodiversity data (and to provide an index of qualified existing open datasets) and make it possible to build graphs to assist in the analysis and development of new ways to mine data. Standards (including TDWG) and specific protocols can be applied to interconnect databases. Such semantic approaches greatly increase data interoperability. They organised in June 2014 a scientific workshop on databases interoperability in ecology. The input was sent by Romain David on behalf of IndexMeed.

Their more important technical challenge is to get more proof of concepts to convince that interoperability is a key for new approaches based on graphs. Realising this proof of concept with different kind of collaborators (STICs and environment scientists) could be facilitated by a step-by-step assistance from EOSCpilot. It is often the first runs that are the most difficult to organise.

They express another very important need in term of skills: "As data literacy is low in our community, transformation can only be possible with a real and hard support to develop human capacity and man power" and they are strong advocates of the FAIR principles: "FAIR Principles must be applied, which will have an impact on collaborative work and research characteristics" and of the use of semantic and standard among communities that they classify as mandatory.

In their answer they express a strong need of support to the EOSCpilot project to help them to identify the consensus solutions according to the heterogeneity of biodiversity data and the multitude of producers and to use the e-infrastructures such as EUDAT and EGI.

[20] http://www.indexmeed.eu

Their answer is well documented and their principles they express in their comment are well aligned on the EOSC principles. They seem to be a typical example of a long tail of science community willing to use the future EOSC.

3.2.2. Phenome

PHENOME[21] is a French "national biology and health infrastructure" of the "Investments for the future" program certified by the GIS Biotechnologies Vertes and the competitiveness poles Céréales Vallée, Qualiméditerranée, Végépolys and Vitagora. It is linked to the AgriPhen SCience Demonstrator included in the EOSCpilot proposal but doesn't reflect all its aspects. The input was sent by Christophe Pradal and deals with the Plant Phenotyping domain.

Their most important challenge at technical level is "Easy access to different cloud/grid infrastructure (e.g. France Grilles, IFB) with the possibility to scale depending on user requests. The impact class of this challenge is important benefit to the use case.

At political level, they classify at the same level the need to build a central web site where users can run public OpenAlea workflow that demonstrates the usefulness of such infrastructure. OpenAlea is a Software Environment for Plant Modelling[22]. OpenAlea is an open source project primarily aimed at the plant research community, with a particular focus on Plant Architecture Modelling at different scales. It includes modules to analyse, visualize and model the functioning and growth of plant architecture.

The social challenge is more important (mandatory for the use case): Scientists would adhere to such infrastructure if they can store their data, have an easy access to it with simple certificate (just connect to your institute account like with IFB), and use it for their daily work, rather than using local clusters. The *Institut Français de Bioinformatique* (IFB[23]) is the French representative in the ELIXIR infastructure. The iPlant (now CyVerse - http://www.cyverse.org/) initiative[24] is a success from this point of view. Perhaps also having domain specific tutorials (like Software carpentry - https://software-carpentry.org/) will enhance a lot the dissemination of such infrastructure. Software Carpentry is a volunteer non-profit organization dedicated to teaching basic computing skills to researchers. Its goal is to make scientists more productive, and their work more reliable. Founded in 1998, it runs short, intensive workshops that cover program design, version control, testing, and task automation. The Software Carpentry Foundation was created in October 2014 in the United States to act as a governing body for the project. They are developing a 'middleware'/library to transparently distribute scientific workflow on cluster/Grid/cloud using OpenAlea (they have obtained a PhD topic funded by #DigitAg).

Another solution is the use of iRODS as a universal solution to store, share data and provenance information during the computation and outside, especially with API request.

They recommend Simple certificate management.

3.2.3. Common Data Centre Infrastructure (CDCI) for Astronomy Astroparticle Physics and Cosmology, Université de Genève, Switzerland.

The CDCI is attached to the astronomical observatory of the University of Geneva[25]. The CDCI emerged from the ISDC, which started its activities as the INTEGRAL Science Data Centre and is now actively contributing to several other space missions and ground-based projects with a prime scientific focus on high-energy astrophysics.

N.B.: The answer is also presented in the infrastructures input. As it is dedicated to a scientific community this centre is at the interface between the scientific communities and the infrastructures.

[21] https://www.phenome-fppn.fr/phenome_eng/
[22] http://openalea.gforge.inria.fr/dokuwiki/doku.php
[23] https://www.france-bioinformatique.fr/en
[24] http://www.cyverse.org/news/iplant-collaborative-has-become-cyverse
[25] http://isdc.unige.ch/

In their answer they explain that other services are not considered due to:

- Intrinsic rules of collaborations around the projects and the associated privacy / data property issues,
- Absence of clearly defined framework for funding the cloud-based computing by funding agencies,
- High price of CPU time and storage at the national academic science cloud services (e.g. SWITCHengines) and,
- Absence of the guarantee of long-term commitments by the cloud service providers (the astronomical projects and space missions are developing on the time scale of decade(s)).

A standardized / transparent / stable in time approach to provision of cloud computing resources, with pricing models fitting well into standard funding scheme of space science and astronomical projects (through national science foundation(s), space agencies, European Southern Observatory etc.) would increase proliferation of the use of cloud computing for the CDCI services. Our services (access to astronomical data and data analysis pipelines) currently run locally, but we envisage to use cloud computing to cope with peak loads (e.g. at the moments of re-processing of many-year data sets of space missions, before public data releases etc.).

We also plan to use cloud services to deploy data analysis pipelines for astronomical “big data” projects currently in the development stages.

The main interest will be to use EOSC to provision “on-demand” increases of CPU power and storage for peaks of data processing activities for all CDCI astronomical data projects and for deployment of data analysis pipelines in response to queries for the astronomical data products received through the web data analysis interface.

Another particularly important aspect addressed by our CDCI infrastructure is long-term preservation of data together with the full data analysis software system. Our activity is related to the data of space missions and astronomical observatories, but the problem is generically important and should be addressed across science disciplines. Dedicated study of reliable long-term (decades time scale) preservation of data analysis software systems together with the full raw data sets would be interesting to address within a broader context of the EOSC. Best solutions: We are currently exploring the deployment of data analysis pipelines using a Docker-type container approach, which appears promising for interoperability and which also provides a solution for long-term preservation of data analysis software.

3.2.4. PiCo2

PiCo2[26] is a technical pilot about interoperability between generic (community agnostic) infrastructures at Tier-1 and Tier-2 level, in the long tail of science. It is presented here as a T6.3 use case. It reflects a frequent need of users of at least two computing centres that are willing to share data between a National HPC centres and regional HPC centres.

The challenges regarding interoperability are mainly about data transfer: how can we improve the data transfer in terms of security, reliability and performance? How can this be done at national and European level in a transparent way for our users? This is the real goal of the project.

Another technical challenge is about authentication and authorization issues: how can we authenticate and authorize users in a global manner from one infrastructure to another? Pico2 will benefit of the work of other WP in EOSCpilot project.

At political level, the challenge will be about access policy: who will be authorized to use the infrastructures?

[26] https://repository.eoscpilot.eu/remote.php/webdav/WP6%20-%20EOSC%20Interoperability/Task%206.3%3A%20Interoperability%20pilots/PiCo2/EOSC_Technical_Note_Interoperability_Pilot_PiCo2.pdf

In order to answer this question, we will deploy a federation of IRODS area at national level and between Tier-1 and Tier-2. This will allow our users to transparently transfer large files. This is a generic solution and every scientific community can be concerned by their use.

About authentication, the use of identity federation at European level will certainly be of great interest.

3.3. Other inputs (Indigo DataCloud, ELIXIR compute platform...)

Here are other valuable inputs collected by other projects.

3.3.1. INDIGO-DataCloud

The INDIGO-DataCloud project was designed with a clear objective in mind:

“The proposal is oriented to support the use of different e-infrastructures by a wide-range of scientific communities, and aims to address a wide range of challenging requirements posed by leading-edge research activities conducted by those communities. Indeed, the force driving this proposal is the interest of these communities and the organizations supporting them, from many different fields in science, from biomedicine to astrophysics, from cultural heritage to climate, participating in very relevant initiatives at European level, such as INSTRUCT, ELIXIR, EMSO, DARIAH, LIFEWATCH, etc."

The following table show the partners and related Research Communities that participate in the INDIGO-DataCloud project, and in particular that contributed to the activity of requirements gathering:

#	Partner	Research Community	Area/Type
P0	CSIC	LifeWatch	Environmental ESFRI
P1	UPV	EuroBioImaging	Biological and Medical Sciences ESFRI
P2	CIRMMP	INSTRUCT	Biological and Medical Sciences ESFRI
P3	INAF	LBT CTA	Physical Sciences ESFRI
P4	Univ. Utrecht	WeNMR	Biological and Medical Sciences #
P5	CMCC	ENES	Climate and Earth System Modelling
P6	ICCU	Galleries, Libraries, Archives and Museums	Social Sciences and Humanities Cultural Heritage
P7	EGI.eu	EGI	All Areas
		Virtual Teams	
		Competence Centres	
P8	CNR	ELIXIR	Biological and Medical Sciences ESFRI
P9	INGV	EMSO	Environmental Sciences ESFRI
P10	RBI	DARIAH	Social Sciences and Humanities ESFRI

Table 1: INDIGO-DataCloud Partners and related Research Communities

After collecting the description of the use cases of each research community involved and performing an initial analysis, a (long) list of around 100 requirements per Case Study was compiled. The resulting table includes several entries per Case Study, and detail the Research Community, Requirement enumeration (#), description, priority rank (Mandatory/Convenient/Optional), current solution, gap, etc. The requirements

are classified as “Mandatory”, if they are felt as such in the input provided by the corresponding Research Community. This may mean that either the requirement is satisfied by a current solution and needs to be preserved, or that it must be implemented in the new solution.

The complete list with all details can be found here.

In the activity of requirements gathering from the research communities and their implementation into new or improved services an Agile Development approach has proven to be useful in promoting adaptive planning, evolutionary development, early delivery, continuous improvement, and encourages rapid and flexible response to change. It refers to a set of software development methods in which requirements and solutions evolve through collaboration between self-organising, cross-functional teams. “Agile Champions” are the key of efficient agile transformations. Typically, they are the ones who try to continuously find better ways of working with others around him/her and adjust the practices to suit their environment. A “Champion” is identified from each research community to lead the “Agile-like” effort in coordination with user communities’ team management and with the development teams. They assure that community requirements are taken into account and that INDIGO solutions are of interest to the Research Communities that they represent.

In order to unify the understanding in communication, a common terminology is introduced:

- A Case Study refers to an implementation of a research method involving an up-close, in-depth, and detailed examination of a subject of study (the case), as well as its related contextual conditions.
- The Case Study is based on a set of User Stories, which are tokens representing requirements and describing the value of functionalities from a user’s perspective, i.e. how the researcher describes the steps to solve each part of the problem addressed.
- Requirements are a list of technical points, being derived from the different User Stories.

After the first iteration that provided the first list of requirements, a second one followed. By this process, new or refined User Stories were introduced, and in particular as Champions become familiar with the potential solutions, additional notes/explanations are introduced refining the requirements and their scope.

3.3.2. ELIXIR

"The ELIXIR Compute Platform: A Technical Services Roadmap for supporting Life Science Research in Europe" is the D4.1 deliverable of the ELIXIR-EXCELERATE project.

It includes use cases and information related to interoperability in the Life Science domain. The document (a work in progress) is available as a Google doc[27]. It will be updated during the next summer (2017).

The deliverable is built on the initial requirements from ELIXIR-EXCELERATE Technical Use Cases (TUC) to define an ELIXIR technical services roadmap. These TUC represent "generic technical activities" extracted from scientific use cases. In this document they identify issues preventing the exploitation and usage of existing e-infrastructures and distributed resources that may be taken into account in the EOSCpilot gap analysis. The TUC are detailed in the Appendix B section of the D4.1 deliverable.

The ELIXIR Compute Platform relies on the one hand on services that are provided by existing e-infrastructures such as EGI, EUDAT and GÉANT and on the other hand on services run by the ELIXIR nodes. Consequently interoperability between all those services is a key to build the Platform.

Several TUC are related to Authentication and Authorization interoperability through a global AAI: TUC 1, TUC 2, TUC 3 and TUC 10. All together allow a user to use the Compute Platform Services without to be prevented from their use. It is important to note that ELIXIR’s work on its AAI is based on the AARC recommendations ELXIIR AAI and ELXIRI ID is a first step for a Life Science AAI. This is a good pilot for the future EOSC.

[27] https://docs.google.com/document/d/1gMKFrcbzuN9BSREU1VDnlml-bl6KSOnfyQbJGh20L5s/edit?usp=sharing

TTUC 8 and 15 are related to file transfer that relies on sites files transfer interoperability. TUC 20 and 23 imply that Cloud IaaS on the one hand interoperate and HTC/HPC clusters on the other hand interoperate too through federations.

Three TUCs are especially related to the use of European e-infrastructure services for ELIXIR: TUC 9 (Infrastructure Service Directory), TUC 21 (Operational Integration), TUC 22 (Resource Accounting).

All these needs are relevant in the framework of the EOSCpilot. Technical details are given in the document.

4. ANALYSIS

4.1. Technical aspects

Following section provides the analysis of the technical aspects important in the context of infrastructure interoperability. For each part the analysis explains the issues, lacks or needs, the gap to bridge and points out technologies and key Open Source possible solutions, best practices and feedback. One has to keep in mind that the infrastructures are not at the same level regarding interoperability and that operational solutions running already in several sites may be ignored by others or irrelevant for them. For that reason already existing solutions are given to serve as examples.

4.1.1. Authentication and Authorisation Infrastructures (AAI)

AAI appears to be *the* main challenge regarding interoperability and the main issue that prevents exploitation and usage of existing e-infrastructures and distributed resources. Almost all answers to the questionnaire point at this problem at different levels. There are individual solutions such as CheckIn, Unity, INDIGO DataCloud IAM, B2ACCESS, UmbrellaID or eduGAIN, as well as others. Because it is a well-known issue already studied for a long time, many initiatives and projects were and are conducted on this topic and partial solutions are currently in production and applied in e-infrastructures or in Research Infrastructures.

The subject includes Identity Management (IdM), as well as Authentication and Authorisations. It deals with the support of multiple technologies, methods and protocols. As explained by EUDAT, each (e-) infrastructure, service provider, and community has adapted different kinds of methods and technologies for IdM and authentication and authorisation, creating more or less different user domains. As pointed out by EGI, the progresses in this domain are quickly advancing the state of art.

One of DESY's challenges is to find an AAI solution that provides access to compute, storage/data and web-services likewise. This would preferably be a Single-Sign-On solution, allowing federated access at the European level. Certificates are not really an option for some of the scientific communities. For some large and highly distributed communities, the currently used mechanism of providing external accounts does not scale any more e.g. for the XFEL at DESY.

The eduGAIN interfederation service provided by GÉANT, enables researchers to collaborate using the very same accounts they are using at their home organizations. Like the GÉANT network provides connectivity across borders, eduGAIN interconnects the national identity federations, creating a global network of identities and institutionally provided attributes that can be leveraged by scientific communities and e-Infrastructures when designing and implementing their collaborative services. All the federated IAM solutions mentioned in the questionnaire rely on eduGAIN as the solid foundation for federated access. GÉANT explains that eduGAIN enables the trustworthy exchange of information related to identity, authentication and authorisation (AAI), simplifying access to content, services and resources for the global research and education community. GÉANT is continuously enhancing the eduGAIN portfolio with more services and capabilities, such as the InAcademia and eduTeams service, and SIRTFI (security incident response frameworks) based on the SCI (Security for Collaboration in e-Infrastructure) principles.

EGI notes that the project could experiment with a system in which researchers may use a single digital identity through the EGI AAI CheckIn service to access resources federated by different data providers and computing infrastructures. EGI points out that CheckIn is a solution already implemented within the EGI framework.

EUDAT explains that it has developed the concept of credential conversion (e.g. token translation) to support multiple methods for user authentication and service integration. For this EUDAT developed the B2ACCESS service to enable access to the EUDAT CDI. The B2ACCESS service has been in production since October 2015. To enable access between EUDAT, EGI and the PRACE infrastructure and services, EUDAT started, with support from the AARC project, integration work between EGI and PRACE AAI systems to provide access across these e-infrastructures. EUDAT has also started to pilot the use of the B2ACCESS

service with the EPOS community to provide an EPOS unified IdM domain and to bridge access to services between EPOS and CDI. To foster this collaborative work, it would be of great interest to extend this to other communities, infrastructures, and services.

The AARC projects that are presented in Section 2.4.1, are specifically working on this matter and ELIXIR has already implemented its recommendations in its production AAI.

The EOSCpilot project may take advantage of these activities and, if possible, contribute with use cases and inputs. It is important to avoid duplication of work and to include experts in the field. The EOSCpilot is probably a good place to extend the use cases and foster implementation of the AARC results and to disseminate this information.

4.1.2. Data

Data Interoperability is the topic of task T6.2 in work package WP6. For this reason only technical points not related to data itself are taken into account in this section.

Many technical tools and solutions (and embedded protocols) are already in production in the different e-infrastructures in Europe. Descriptions of the technical solutions deployed in the e-infrastructures can be found in Annex 4. Consequently, the diversity and the heterogeneity of the protocols, tools and solutions already used in the different infrastructures, the lack of interoperability in services that rely on these tools and solutions, and the absence of interoperability in the AAI systems used by each infrastructure is a common challenge and the main visible motivation that prevents usage of distributed resources.

Several other challenges are listed in the questionnaires:

- The large size of certain datasets,
- The need of high performance data transfer with different related issues (large amount, transfer of access rights…),
- Data mining possibilities,
- The long-term preservation of data together with the full data analysis software system.

All these issues may be grouped in different topics:

Data storage and data management

Many solutions and tools are cited in the questionnaire as relevant and already used in production by different e-infrastructures and Research Infrastructures. For example:

- **iRODS**: IN2P3-IRES explains that iRODS provides a powerful rule system for managing data and data workflows. The PiCo2 service pilot project proposes iRODS as a solution and EGI and EUDAT apply it as well,
- **Onedata:** cited by EGI.eu and INFN-Padova,
- **dCache**: DESY explains they are running dCache as a highly scalable storage backend, serving close to 20 PBytes, to support preconfigured or flexible storage qualities in conjunction with ownCloud/nextCloud as an interface for syncing and sharing as well as to store and to retrieve data.

One of DESY's challenges is about HDFgroup[28]: HDFgroup has some services like HDF5server and HDF product designer which allow native and distributed HDF5 access and registries of schemas for HDF5. Since HDF5 is the most widely used format (and the only de facto standard for binary data) it would be important to support and to provide services based on these. There are only a few attempts so far, but HDF has demonstrated the capabilities in a cloud environment. The Photon Neutron Science Demonstrator needs to extend data interoperability at the binary level through HDF5 and NeXus. However, the data interoperability is still rather limited. This could be greatly improved through interoperability services, e.g. a kind of data type registry with machine digestible schemata. In this particular case, the HDF Product Designer appears to be the most suitable implementation.

[28] https://www.hdfgroup.org/

Unistra expresses the need to share different file systems between two sites that are willing to collaborate.

IN2P3-IRES is currently testing the OpenIO storage solution[29]. It may provide an efficient way to distribute storage over geographical locations and provide high performance access to storage. It provides a standard S3 API.

It is clear that the EOSC will have to manage this diversity of tools and solutions that are used with success (even if they are not always completely suitable for all purposes) for a long time and the EOSC has also to take into account the fact that these tools will probably evolve in time.

Even if those tools are successfully used, all needs can currently not be solved:

- For example, EGI highlights the lack of solutions for hosting privacy sensitive data in a secure manner in a distributed, multi-supply environment.
- Data accounting is needed (it would be helpful for e.g. the Photon-Neutron Science Demonstrator) and this has to be possible also in a distributed environment.
- Data size in certain use cases such as PanCancer: data set sizes exceed capabilities of current infrastructures, thus it is imperative to enable interoperable operation to maintain ability to analyse data sets required by today's and future scientific studies in the field of cancer research.

Data transfer and data placement

Data transfer may be difficult between two sites for different reasons:

- The lack of common identity management or authorisation management. This is discussed in the AAI chapter.
- Transfer of metadata between services must include access rights on the data as explained in detail in the the Jülich Supercomputing Centre answer. It depends on the AAI but **this is also a requirement for the data management and storage tools and for the APIs.**
- The lack of common APIs. This is discussed in the paragraph about standardisation and interoperability between APIs for access and data transfers.
- The capacity of the network, especially for large amounts of data. For example, PanCancer datasets' sizes exceed current capabilities. This question of networking capacity is discussed in the Networking section.

But data transfer alone is not sufficient: the CC-IN2P3 notes that using computing resources from everywhere means that data should be accessible also from everywhere, which requires a well-designed network infrastructure and/or smart processes to manage data placement. About the same topic EGI proposes that the project could experiment with solutions such as Onedata and iRODS for federation of data across scientific domains and within one domain, to enable ease of discoverability, access and bringing data near to computing. A piloting activity involving different data providers and computing infrastructures would be beneficial from this point of view.

The geographical data dispersion of the PanCancer Science Demonstrator could be an opportunity to test solutions for this problem.

Standardisation and interoperability between APIs for access and data transfers

As EUDAT ranks among the most important challenges it is currently facing: within the (e-) infrastructure and services different types of APIs are being supported for access and data transfers. To enable easy access and data flows across services and (e-) infrastructures a number of APIs should be standardised. To make efficient data transfers possible, APIs and protocols for data transfers should support third party transfers. This does not mean a single API or single protocol. Multiple APIs can be adopted, but a limited number should be set as a standard across the (e-) infrastructures and services. For sustainability, these APIs should be preferably based on HTTP and on open and/or well-adopted standards.

[29] http://openio.io

The CC-IN2P3 needs common and standardised APIs for accessing computing and storage resources. That means that the APIs available for data access must be shared with computing services.

According to the data transfer and data placement paragraph the APIs available for the transfer of data must also transfer the metadata between services and must preserve access rights to the data.

The EOSCpilot project may propose a set of APIs that should be then supported by the services providers according to EUDAT recommendations.

Data Mining

The *Mésocentre Clermont-Auvergne* (MCA) explains that data interoperability is an emerging requirement from several MCA user communities. Of particular relevance is the development of distributed research infrastructures (LifeWatch, ENVRI) in Environmental sciences that generates very challenging interoperability requirements. Despite the existing efforts within the Research Data Alliance (RDA[30]) and other international projects and consortia (e.g. LifeWatch[31]), there is a missing common strategy or tool to provide the interoperability of data collected from multiple observatories. It is important in the area of data interoperability to learn from GAFA[32] what has been successful. Data mining approaches of unstructured data using tools like ElasticSearch should be further explored.

Long-term preservation

Long-term preservation is cited by the CDCI as an important factor for astronomical data, which has to go hand in hand with preservation of the full data analysis software system. CDCI's activity is related to the data of space missions and astronomical observatories, but the problem is generically important and should be addressed across science disciplines. A dedicated study of reliable long-term (decades time scale) preservation of data analysis software systems together with the full raw data sets would be interesting to address within a broader context of the EOSC. Long-term preservation is certainly a topic that interests a wide range of communities.

Quality of service

DESY suggests continuing its efforts in evaluating the interoperability work in terms of Quality of Service in storage solutions. If the EGI propagated Onedata storage federation solution turns out to be used by other e-Infrastructures, DESY suggests continuing its efforts to make the dCache *Quality of service* capability available through the federated Onedata system.

Despite the fact that only one answer mentioned it, the Quality of service in storage is obviously a necessary point that the EOSC must take into account to be able to attract user communities.

Summary

In the research communities and e-infrastructures a large variety of tools and solutions is successfully has been used for quite some time. However, certain cases are not currently covered in a distributed environment (hosting of privacy sensitive data, data accounting) and the data size in certain use cases exceeds the capabilities of current infrastructures. Data transfer may be difficult due to the lack of transfer of the access rights and of common APIs and due to the low capacity of certain parts of the network. There is a need of smart data placement, as well as of standardised APIs. Data mining, long-term preservation and quality of service are also of interest.

4.1.3. Computing

Descriptions of the technical solutions deployed in the e-infrastructures are given in Annex 5.

[30] https://www.rd-alliance.org
[31] http://www.lifewatch.eu
[32] Google, Apple, Facebook and Amazon.

Communities may have very diverse needs concerning computing. For example, the Photon and Neutron Science Demonstrators applications are very diverse requiring a wide spectrum of different resources. The processes can range from very low CPU requirements, over embarrassingly parallel to massively parallel applications, such that the most appropriate platform varies from experiment to experiment. As such, infrastructure interoperability services would be highly beneficial for the Photon and Neutron use case.

At European level and often at national or regional levels the infrastructures that deliver embarrassingly parallel or High Throughput Computing (HTC) and now Cloud Computing are different from the infrastructures that deliver massively parallel or High Performance. HTC, cloud and HPC technologies are different too. Even though PRACE interconnects the main HPC centres in Europe and EGI has considerable experience in HTC and cloud service solutions that allow achieving interoperation at international level, issues remain open for resources providers as well as for users.

Resources providers such as INFN-Padova for example would like to have a solution that allows users to run jobs both on cloud and HTC resources with the same interface. Unistra would be interested in testing web portals allowing users to submit jobs, and mechanisms to integrate HTC jobs into HPC workflows. They have been working in cooperation and synergy with the University of Freiburg and the KIT, organizing workshops on how to run HTC jobs on HPC Machines. They wish to extend the audience of such workshops and to keep the trans-national character to enforce collaborations in the European framework. To this aim, they underscore that they have all the expertise and know-how to contribute to this kind of initiatives.

IDRIS is interested to develop the original framework originally experimented between CNRS-IDRIS, as a Tier-1 site, and a Tier-2 site located in Bordeaux (South-West of France) to post-process on a regional site the data generated by numerical simulations executed at IDRIS, around a data framework based on iRODS. This has been proposed in the context of task T6.3 (interoperability pilots) during the WP6 Kick-Off meeting in February. In the PiCo2 project, we plan to replicate this framework in another place. In a next step there will be experiments how to extend it across two Tier-1 sites, two Tier-2 sites, and possibly a grid-based infrastructure. Such experiments would be transferable to other scientific communities.

Resources providers are willing to serve their user communities whatever their computing needs are. This is a positive input for the EOSC. It is important that bridges between the different infrastructures are developed to allow users to use all types of resources they need in their scientific workflows.

Cloud

EGI highlights the fact that an area that still requires investigation and the definition of an interoperations blueprint is hybrid cloud federation involving commercial and publicly funded clouds.

INFN-Roma would like to have the possibility to automatically scale out infrastructure based on specific workloads and to be "multi tenant" ready (support different type of analysis with different distros/libs). They are now experimenting an automation / scale out solution based on Ansible and VMWare.

The CC-IN2P3 proposes that the EOSCpilot should experiment with cloud (virtualisation on demand) and storage resources.

Since 2013, IN2P3-IRES has developed expertise on OpenStack, the IaaS Cloud framework. Their infrastructure is already connected to several federations (France Grilles, EGI, IFB). They are interested to share their expertise and resources with EOSC partners and state that OpenStack is a promising Cloud framework that can be used to federate distributed infrastructure resources. It provides a convenient to provide HTC and HPC resources to users and administrators. Additional services (like SlipStream or components from the INDIGO-DataCloud project) may be used to simplify the deployment of infrastructure across the sites.

DESY is operating OpenStack and tries to get as many INDIGO extensions integrated as possible (like the TOSCA HEAT translator).

EGI suggests to test brokers for instantiation and management of virtual machines (VMs) taking into account data locality and recommend as a good solution IaaS Provisioning tools (e.g. IM, Terraform or

OCCOPUS) that unify and simplify the access to heterogeneous cloud frameworks using a single infrastructure description and a single workflow for managing resources at all providers. These make use of a variety of standards for consistent access and management of different clouds (e.g. OCCI, TOSCA)

EOSC may play an important role in requiring standard based solutions.

Container management

The CDCI is currently exploring the deployment of data analysis pipelines using a Docker-type container approach, which appears promising for interoperability and which also provides a solution for long-term preservation of data analysis software.

EGI proposes that the projects could experiment with Container Management in a multi-provider environment (like a Kubernetes[33] Federation) to enable a single platform for running container-based applications on the infrastructure.

DESY suggests providing the results of INDIGO-DataCloud in making container management in OpenStack as well as investigating in the INDIGO improved scheduler and preemptionable instances for OpenStack.

4.1.4. Network

4.1.4.1. Interconnection

The interconnection between EOSC sites research and academic users is a key requirement. For most R&E sites general-purpose IP connectivity to EOSC sites utilizing the GÉANT/NREN infrastructures is already available. In several cases, more advanced service offerings is already available. However, a number of gaps are observed, an analysis for which follows.

Last mile issues

As already mentioned, gaps in the infrastructure or service offerings between end/user sites or EOSC sites and the core GÉANT/NREN infrastructures exist. Intermediate networks (campus, regional etc.) might not offer adequate capacity/throughput or quality to address the EOSC use case needs. In cases of inadequate capacity, infrastructure investments will be required. When high throughput is required, dedicated equipment to serve scientific traffic might need to be deployed in parallel to commodity infrastructure at the last mile. In cases of qualitative or other requirements (e.g. the delivery of VPN services end-to-end) the last mile should be equipped with technical capabilities (e.g. software defined networking) to implement QoS and features such as virtual private networking. Operational capabilities to deliver and monitor specialized services across the last mile are also required.

Connectivity across commercial cloud sites/providers

EOSC is expected to span across public and private clouds. Access to commercial clouds is at present through Internet providers. Data transfer throughput and security between commercial clouds and the R&E user community often presents a number of challenges. However NRENs and GÉANT have devised a cloud connectivity strategy and implement a number of technical solutions to eliminate relevant gaps. These include:

- Flexible peering connections to commercial cloud providers via GÉANT, specific NRENs or Internet Exchanges, in which NRENs can opt-in (or out) to IP (L3) traffic exchange with each different cloud provider. In this setup all national or regional network traffic to/from a cloud provider is treated in the same way.
- Delivery of dedicated L2 or L3 VPN connections between a user end site and the border of the cloud provider network, over the GÉANT/NREN networks. Such connections can be specialized on a user-site to cloud provider basis and can also accommodate qualitative requirements.

[33] https://kubernetes.io/

Regarding traffic transport charges, agreements are already in place between GÉANT and a significant collection of commercial cloud providers, in the context of the GÉANT IaaS Framework Agreement[34] to reduce connectivity costs, waiving data egress or ingress fares.

Automation and orchestration

The EOSC is expected to grow to a complex federation of e-Infrastructures and service providers, with the goal to offer identical services to all researches, independently of location. To efficiently control and manage the EOSC resources and services, orchestration and federation across providers will be required, as well as robustness and resiliency of the different offerings.

At an abstract level, services delivered to end-users will be a combination of digital (computing, storage) and network (data transfer) resources. User experience on cloud delivery models requires that computing/storage offerings are bundled with network access both in terms of ordering and in terms of operations. On demand storage/computing resources should be accompanied by on demand network access to them. In cases where data transfers and remote network access require specialized virtual private network services (with or without quality of service), the delivery of virtual network infrastructure across multiple networks needs to be automated to the greatest extent possible and orchestrated with the delivery of digital resources, so that the user has an one-stop-shop experience (see **Errore. L'origine riferimento non è stata trovata.**) .

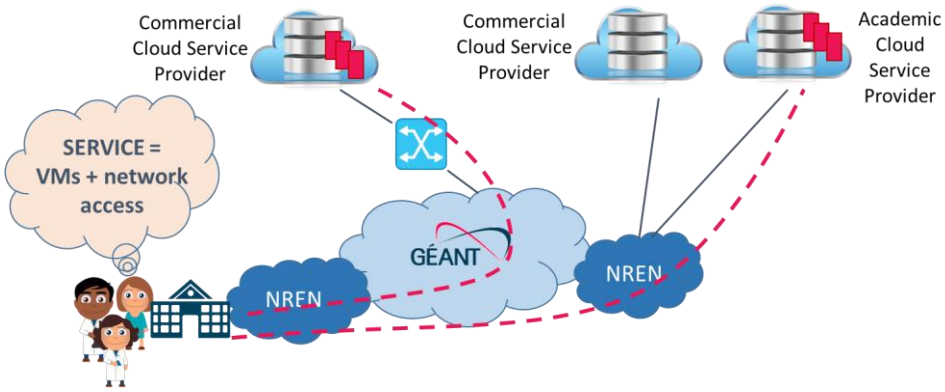

Figure 2: Users require an orchestrated service offering

At present, dedicated network access to cloud resources from remote end sites, requires coordination among the involved intermediate operational entities and often results in extended lead times for service delivery. Furthermore, during service operation, service requests and monitoring/troubleshooting requirements are also propagated across the involved service providers hindering seamless service delivery to users. It is therefore expected that integrated service delivery and operations for EOSC use cases may be greatly improved in terms of efficiency and agility introducing standardization, orchestration and automation (where possible) thus minimizing manual intervention and lead times. Such an approach requires an interoperability framework between the network and the cloud operators.

Service provider/operator orchestration is currently a gap for e-Infrastructures. It is also a gap outside of the e-infrastructure ecosystem, in the industry/private market of cloud services. Hence, relevant industry initiatives are currently under development:

[34] https://clouds.geant.org/services/geant-cloud-catalogue/infrastructure-services/

- The Metro Ethernet Forum (MEF) Lifecycle Service Orchestration (LSO) [35] track of specification of open APIs to automate the entire lifecycle for services orchestrated across multiple provider networks (and multiple technology domains within a provider network).
- The TeleManagement Forum (TMF) Open APIs initiative as a practical approach to seamless end-to-end management of complex digital services[36]

At the same time multiple open source and commercial solutions for next generation OSS/BSS stacks and orchestration are currently available, entering a stage of maturity, or under development.

As the demand is strong, the industry is actively working on standardization of the inter-provider Application Programming Interfaces needed for interoperation and relevant implementations, while at the same time different intra-provider delivery automation solutions are being deployed.

The uptake of relevant specifications (MEF, TMF) by GÉANT/NRENs as a means to achieve interoperability with other service providers is currently on-going. A lightweight but sufficient adoption of APIs at the business (catalogue management, order management, customer/user management) and the operational layer (fulfilment, inventory management, trouble ticket management) are under adoption. At the same time, public and private cloud APIs need to converge so that cloud to network and network to cloud service signalling can become possible at large scale. Convergence of inter-provider APIs within the EOSC ecosystem is a crucial gap. It is expected that not all providers (e.g. some private ones) will move towards adopting standards and this needs to be addressed through proper adaptation functions.

GÉANT work on realistic use cases for orchestration is in progress at the moment of writing and can serve as a boilerplate for interoperation in the EOSC context.

In a short-term, two-stop-shop scenario (see **Errore. L'origine riferimento non è stata trovata.**), the users can use any EOSC provider portal to request both computing/storage and connectivity resources. Orders can be broken down into individual order items and submitted to endpoints of the associated providers via standards-based API calls. In the case of network connectivity over GÉANT/NRENs, the relevant endpoint receives the connectivity order item and initiates an orchestration process to deliver the end-to-end connectivity service (e.g. a L2 VPN) from the end-user site to the cloud facility over the campus, NREN, GÉANT, possible Internet Exchange or other network infrastructures. Orchestration will also invoke the relevant cloud provider API endpoints to stitch the GÉANT/NREN delivered circuit with the cloud facility network edge. Overall orchestration achieves the seamless coordination of provisioning actions among all involved parties, however, some of these provisioning actions can still be human/manual-driven. Still, via orchestration, the user has visibility on the progress of his request at any point in time, the relevant provider actions and time frames for order completion. After service delivery, orchestration can be very useful for monitoring, troubleshooting, SLA management and other aspects of the service instance lifecycle.

[35] https://www.mef.net/third-network/lifecycle-service-orchestration
[36] https://www.tmforum.org/open-apis/

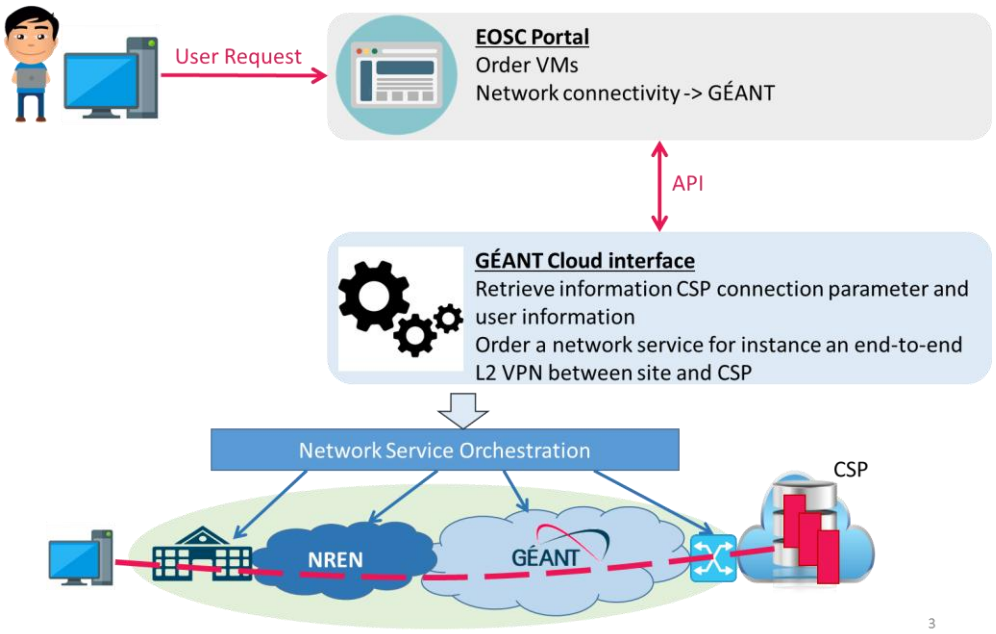

Figure 3: One possible scenario for orchestration

4.1.4.2. Network services

EOSC may profit from a set of multi-domain network services provided by the NRENs and few in-field trials.

VPN services

The VPN services of GÉANT are GÉANT L3VPN, GÉANT Plus (P2P L2VPN), GÉANT *Bandwidth on Demand*, and MD-VPN. These services are adequate to create infrastructures that may be used to interconnect specific projects and the cloud sites that form "the backbone of EOSC".

The request for more extensive and dynamic provisioning has to be assessed and may lead to increased automation over a wider footprint in Europe. Currently GÉANT *Bandwidth on Demand* and MD-VPN fulfil the initial requirements.

Bandwidth guarantees

The connectivity services can be offered with protection and with capacity guarantees, as in these cases a resource is used in exclusive mode, but the cost is significantly higher in this case. The GÉANT *Lambda service* (circuits at the optical layer) is suited to deliver point-to-point circuits with assured capacity guaranteed, but it is available only at sites connected by fibre optics. Currently it is a manual service that requires planning and requires a substantial time for delivery. However, it is very effective for connecting permanent cloud sites with very large capacity requirements, as in the particle physics case.

Low latency service

The network may be engineered to reduce latency between specific end-sites, having a lower limit of the geographical distance between the machines. The tuning is manual, causing a long provisioning time.

4.1.4.3. Network Monitoring

Network monitoring is necessary at operational level and at business level (Service Level Agreement). The current situation does not permit a unified view of monitoring information. In some cases, monitoring requires to have a monitor probe inside each network service instance. GÉANT is working on several solutions (PerfSONAR, SQM project, e.g.) that can be validated in the EOSCpilot

4.1.4.4. Security

Security measurements against threats typical of networking, like Distributed Denial of Service, and data encryption are available as a network services. The security architecture of the EOSCpilot will need to

develop and integrate and holistic architecture for the combined e-Infrastructure.

4.1.4.5. Trust and Identity

The NRENs have developed and adopted technologies based on open standards for an extensive confederation of identity federations. Various network and application services (e.g. eduroam, Service Providers in identity federations.) rely on these technologies. The identity and authorization management us expected to be harmonized in EOSC, based on AARC/AARC2 results..

4.1.4.6. Summary

Network services are a key component for the success of EOSC. NREN and GÉANT are able to contribute to EOSC with a stable, robust and rather ubiquitous set of services.

The gap analysis provides indications for required interventions on last mile infrastructure (where needed), operational processes and service delivery support systems.

- True end-to-end service delivery (e.g. end-user to EOSC site) is still a challenge, as data transit not only R&E networks but also often a last mile domain (campus, mobile provider, regional network). Policies and technologies need to be refined to ensure end-to-end network service delivery in the context of EOSC is available to all researchers.
- EOSC stakeholders need to collaborate
    - To improve the quality of the last mile connectivity;
    - To define functional requirements and policies towards commercial clouds;
    - To define a strategy on the use and interconnection of public/academic clouds;
    - To develop and define specialized network connectivity requirements;
    - To enforce security at the network level and at the whole EOSC.
- GÉANT/NRENs must develop the capabilities to deliver specialized network services (e.g. virtual private networks) in a 'cloud delivery model' fashion and in orchestration with cloud resources.
- Orchestration and automation developments will speed-up network service delivery, however the process to achieve true on demand delivery may be complex due to software dependencies and integration policies of individual providers.
- Network and services' monitoring has to be further developed and integrated as an unavoidable starting point for service management, SLA enforcement and accounting.

4.1.5. Core infrastructures

Accounting

Accounting including data and network accounting is a global issue for the EOSC. EGI proposes APEL to do accounting of the usage of heterogeneous resources of different types (e.g. computing, data etc.). The AARC project provides information on how to share accounting data between two infrastructures. Accounting is certainly a challenge at technical and organizational level as well as legal level.

Traceability

The CC-IN2P3 notes that for security purposes, the traceability of all actions (for example login, data modifications), must be guaranteed, and must respect all local state laws. Traceability is another challenge for EOSC for the same reasons than accounting.

4.1.6. Workflows management systems, portals across different systems and data analysis pipelines

Portals and comparable user-friendly services are cited in several answers:

CNRS-IDRIS, started to install and experiment with workflow managers and portal technologies (Unicore, Nice, SynfiniWay, SysFera DS) since the beginning of the year 2000, that is to say 17 years ago, as well as with distributed file systems across different platforms (Avaki, Andrew File System, GPFS) a few years later.

From their experience, the various workflow managers and portal technologies tested have most of the time not convinced the users. Nevertheless, there are still communities (in particular in life sciences) for

which we know that this is not only a potential benefit for users, but a real requirement to attract them and to offer to non-technical colleagues a way to easily access significant computational and data management resources, which they will not use otherwise.

IndexMeed raised the question how to streamline production and data storage, and DESY states that providing access to private and experiments' data through modern Web 2.0 and cloud mechanisms will be more and more required. The 'afs' approach is no longer sufficient. Scientists must have the ability to share their data, including massive amounts of data, with individuals and groups. Synchronisation with local and mobile devices is a must.

For France Grilles the real challenge was to provide the services on top of the resources to hide the grid complexity from end-users. This was achieved through the deployment of user-friendly services like Dirac and environments like VIP on top of multinational (Biomed) or multidisciplinary (France Grilles) virtual organizations.

DIRAC is cited in different questionnaires as recommended solution. For IRES the DIRAC interware is a very impressive tool that permits to address different types of computing resources (Cloud, HTC). Having container support (like Docker) in DIRAC would also permit to use easily and effectively HPC resources, but this step clearly needs further development.

INDIGO-Datacloud provides user-friendly front ends for building programmable, general-purpose multi-domain Science Gateways.

Web portals and user-friendly workload managers such as DIRAC are required by user communities or the INDIGO's FutureGateway (Programmable Scientific Portal)[37].

4.2. Political, social, and cultural aspects

In this section political or social aspects that prevent exploitation and usage of existing e-infrastructures and distributed resources are outlined and possible solutions or ways to improve the situation are presented.

4.2.1. Access policies

The conditions for a user, a group of users, or a community to access computing resources generally depend on:

- his/her scientific field, his / her scientific community membership or the project he/she is involved in,
- institute or university he / she is affiliated to,
- the laboratory, the region and the country where he / she works,
- his / her awareness and knowledge of the computing landscape and the different e-infrastructures and their services offering including his /her skills to apply to and to be selected in an HPC call.

Despite the fact that almost all sites and infrastructures of the academic community are funded with public money, the possibility to access their resources is very heterogeneous from a public user point of view. As a consequence, the possibilities for a user to manage a scientific workflow that needs access to diverse types of resources in parallel or sequentially are restricted by access policies that apply to each of the resources needed. The rules that restrict access to resources may also prevent collaborations with other researchers from other institutes, sites, countries or disciplines. In reality, funding agencies define the rules to be applied to the resources for which they provide funding. This is usually the case for states, ministries, regional authorities, European projects and large organizations but also for the laboratory or the unit close to the researcher.

For example, access policies to PRACE and HPC Tier 1 centres (http://www.prace-ri.eu/application-procedure/) are based on excellence: access is granted to a few users or projects after a technical and peer

[37] https://www.indigo-datacloud.eu/future-gateways-programmable-scientific-portal

review scientific selection process. Users or projects are given a predefined allocation to be consumed in a limited period of time.

Another example, in France the calls (twice a year) are under the responsibility of GENCI[38], the entity in charge of the three national HPC centres (CINES, IDRIS and TGCC). Only users belonging to a French organisation can apply. Of course all nationalities can apply to the PRACE calls for Tier 0 resources.

Because the source of funding is often regional, Tier 2 policies are mainly based on the user geographic location: a user may be granted access if he/she works in the geographic footprint of the Tier2 centre. In addition, funders may add requirements on the scientific discipline or the academic affiliation of the potential users. Tier 2 centres may have to deal with different funding schemes that apply different rules to the different resources bought with their budgets. As a consequence, the resource centres must know for each new potential user his / her institute or university, the field of his / her research and the geographic location of his / her unit or laboratory to be able to grant him/her access to a suitable resource while respecting the rules. This fragmentation is present at all levels. Resources in a laboratory may be reserved exclusively to the team that purchased it in the framework of its projects.

Germany is a federated country where education and research are competences of regional authorities (Länder) that manage different types of activities such as the eScience Initiative in Baden Württemberg or HeFIS, the "*Hessische Forschungsdaten Infra-Struktur*" in Hessen. In addition, research organizations are also running their own projects such as the development of the FORDATIS private cloud by the Fraunhofer Society or the Helmholtz-Data-Federation.

Large federations of resources such as EGI and the NGIs (grid infrastructures or cloud federation) are differently organised: users are grouped in Virtual Organisations (VO) that can be considered as common interest groups either at scientific level (scientific community) or at geographic level or at technical level (technology sharing). Users are granted access to the resources negotiated by the VO with the resources providers. The membership of the VO is the entry point. The VO system allows in a quite simple way users from different institutes and countries to share resources, data, software, workflows, and results and to collaborate in the end. But even in this type of organisation, issues remain as explained by EGI or IPHC in their answers to the questionnaire. For example, EGI mentions that in some cases, infrastructures integrated into EGI and open to all disciplines may offer different services depending on the discipline.

Research Infrastructures gather researchers around a thematic topic and act as representative of all their members to negotiate the resources and build their own dedicated infrastructure. From the user point of view, this simplifies the day-to-day work. But it doesn't foster either the cross-disciplinary collaborations or the widening of the communities outside the borders allowed by the membership rules.

Long tail of science users or emerging communities encounter issues if their needs exceed or are different from the services offered in their institutes. These groups of users may be considered as "e-infrastructure-less" groups if they are not members of existing large communities.

Partial solutions exist to solve this issue:

France Grilles, the French NGI, started a multidisciplinary VO in 2012 able to welcome users that cannot be members of an existing VO for disciplinary or geographical reasons. This VO runs successfully on mutualized resources committed voluntarily by French sites (grid and cloud). This VO and the related set of services meet the needs of the long tail of science, individual researchers or groups, including support and training. Expertise sharing is the motto of the participants.

EGI set up a portal dedicated to the long tail of science on a different way. Individual researchers may ask for services that are allocated on demand for a limited duration. Services and resources are provided by resource centres on a voluntary basis. The portal is open to all researchers that have no access to such a service in their own country.

[38] http://www.genci.fr/en

These two examples show that solutions exist to lower the barriers for the long tail of science.

e-IRG in its 2016 roadmap identifies the “lack of common policies” as one of the issues and highlights the advantages of a single organisational umbrella in a country to take care of all e-infrastructures (network, data, computing, security) at the national level or at least organisational integration and good coordination. This is a top-down approach.

The EOSCpilot may certainly propose and test a complementary bottom-up approach. The elements could be:

- A global efficient Authentication and Authorization Infrastructure that federates existing AAI. According to all stakeholders this is a necessary infrastructure and it can certainly be built step by step. As pointed by EUDAT it is not just a technical challenge (already discussed in the 4.1.1 paragraph) but to maintain a certain level of assurance (LoA) across the different IdM domains, also an organisational challenge. The EOSCpilot should take its part in this challenge and foster such an organisation in strong links with AARC.
- A multidisciplinary mutualized space (or discipline agnostic) in the EOSC that potentially welcomes all researchers of all disciplines to provide services that cannot be fulfilled by other means. The PiCo2 pilot is a good starting point for such a bottom-up approach.

4.2.2. Knowledge of the computing landscape and of the different e-infrastructures

As expressed especially in the answers of GRICAD and France Grilles there is a large and major challenge in raising awareness about the existing infrastructures and services, knowledge and skills, support, usability and usage. That's a major and invisible barrier in the sense that if the users do not know the existence of an infrastructure or services, they cannot express the need to use it in addition to their usual computing resources or services. For example, the research communities around Photon and Neutron sources are not familiar with cloud services. The applications developers in their communities are lacking the knowledge to integrate cloud APIs in their coding work.

The issue is the same for the resources providers. They cannot help users to diversify the services they use (for instance, find services better fitted to their specific needs) and to use new resources in addition to the resources they propose if they are not aware of their existence or if they do not know how to manage the data transfer with another site or if the collaboration with another site is too difficult. The EOSCpilot could provide practical examples and technical tools to solve their difficulties.

Another convergent point is that despite the fact that the questionnaire was widely disseminated the number of replies was limited: although three mailing lists of e-infrastructures and sites (Italian and French Grids and French regional computing centres) were contacted repeatedly, only a few recipients provided input. Everybody knows that it is difficult to get a high response rate to a questionnaire, but we may also wonder if the targeted people are aware of the European infrastructures landscape and of the EOSC.

Several persons contacted individually explained they were not involved in the project and did not feel concerned by the survey. An effort should be done to better disseminate the benefits of the European infrastructures and deploy a particular effort to the sites and e-infrastructures that could be part of the future EOSC. As closest point of use of the computing and storage resources and services they are multipliers and their knowledge of the EOSC is mandatory in the future. A promising response is that one answer (Strasbourg) proposes to contribute to the EOSCpilot in the transnational aspects. This shows that once informed sites may be interested.

Another gap in this domain is the existence of different vocabularies depending on the e-infrastructure or the services. A common vocabulary or at least a glossary referenced on all web sites and documentation should help all stakeholders to understand each other. INDIGO-Datacloud have set up such a common terminology in its area. It is a good example of the feasibility of a common vocabulary.

4.2.3. Adoption

The EOSC success depends on technical, organisational and political solutions but also on users’ adoption.

This is finally the main gap or challenge that the EOSC has to address. It has to be adopted by users! In fact the best infrastructure cannot be a success if it is not useable, useful and used.

- Useable means that the services offering (and the way to access it) meet at least partially the needs of all stakeholders.
- Useful means that the service offering is user friendly enough and not too difficult to deploy and maintain from the infrastructure providers' point of view.
- Used means that it is adopted by a large portfolio of users and at least by the main targeted users. In other words, it means that nothing prevents the usage and that the benefits and the added-value are high. It also means that the user communities trust enough in it and in its long term sustainability to adapt their own infrastructure and services to interoperate with it.

The feedback provided by Phenome expressed clearly that scientists would adhere to such infrastructure if they can store their data, have an easy access to it and are provided with an example of what an easy access could be. This highlights the technical need of a global AAI that federates or acts as an umbrella on the top of all AAI researchers are used to. The simplest, the use of the different components of the EOSC, the easiest the deployment of a workflow on the top of the EOSC, the larger the number of happy users!

In order to foster the EOSC adoption the support of the users has to be well organised and efficient: ticketing and requirement collection systems have to be globalized.

In addition communities generally need training to use the infrastructures services they are not used to. The IndexMed answer points out the fact that users need training. The Photon and Neutron scientific applications developers are lacking the knowledge to include cloud APIs in their developments. The Phenome answer gives the example of Software Carpentry that trains researchers in a very popular way. There is a clear need of training in all possible forms to support users and developers to better use the panel of resources and services the EOSC will offer them.

In parallel it is important that human support resources such as local teams of engineers are available at the closest level to build the bridges between the scientific communities (scientific interoperability) and the infrastructures (technical interoperability) as expressed in the answer of the Mésocentre Clermont-Auvergne. These academic level e-science personnel should be recognised as an indispensable part of the EOSC. Regional centres, which are the closest to the end users, especially working on the long tail of science, can / must be the first level for information, training and support about adapted and accessible e-infrastructures.

Concerning the adoption, we have to consider the universality of the EOSC versus the specificity of user community needs: In their position papers the main European e-infrastructures describe the EOSC as universal: "Comprehensive: The Open Science Cloud will be universal, specific to no single scientific discipline or research field. It will promote inter- and multi-disciplinary science and encourage innovation and integrated knowledge creation among all research communities, also capturing the long tail of science and citizen science." (Position Paper: European Open Science Cloud for Research). But a number of disciplines has specific needs concerning their fields of research. For example Life Science, especially human data analysis and storage require specific configurations and tools and a level of assurance that is generally not required by other disciplines. This gap has to be taken into account as much as possible.

In addition as explained by IDRIS, the current Science demonstrators either have no real interoperability requirement (TextCrowd) or have real ones but have already agreed and designed their own choices and solutions.

The IDRIS point of view is that more valuable inputs can come from ad hoc projects selected after an investigation of the requirements of some representative users and some relevant communities as this was done for the PiCo2 project.

The infrastructure and the services provided by the EOSC will have to be able to interoperate with similar services used by those communities and by other communities to realise the interoperability that will be

achieved through the deployment of scientific applications across the disciplines.

This will require an intermediate layer of engineers who are knowledgeable about the EOSC catalogue of services and who work with all user communities. This is pointed out in the France Grilles answer for example. It is really important to create a human network across disciplines and across infrastructures that gathers engineers and scientists who are willing to share their expertise and know-how. This is the best solution to foster interoperability according to France Grilles.

At a more political level the adoption issue may be linked to the limitations due to the footprint of the EOSCpilot: the EOSCpilot project accomodates a large number of partners and takes into account a large scope of needs brought by the Science Demonstrators and all partners. But the needs of the other European stakeholders also have to be considered: there could be a gap between the project results and what future stakeholders of the EOSC that are not part of the EOSCpilot may expect. The discussions during the last Plan-e meeting in April in Poland about the EOSC show that this expectation is high and that there are question marks hanging especially on the financial model (business model/funding models), the inclusion of all stakeholders in the discussions (European member countries, scientific communities and end-users in the different countries, e-science centres...) and the objectives of the EOSC to be able to serve all scientific communities. This issue is probably larger than the project itself but it is probably possible to address it at least partially by discussions with those stakeholders or decision makers outside the project partnership.

Another point is related to issues in the service provisioning: The Common Data Centre Infrastructure in Switzerland points out the fact that the use of open / public cloud services is currently not considered due to different reasons. Some are linked to the funding and the cost (the absence of clearly defined framework for funding the cloud-based computing by funding agencies and the high price of CPU time and storage at the national academic science cloud services, e.g. SWITCHengines) and one is the absence of the guarantee of long-term commitments by the cloud service providers (the astronomical projects and space missions are developing on the time scale of decade(s)). The Centre adds that a standardized / transparent / stable in time approach to the provision of cloud computing resources, with pricing models fitting well into standard funding scheme of space science and astronomical projects (through national science foundation(s), space agencies, European Southern Observatory etc.) would increase proliferation of the use of cloud computing for the CDCI services.

The issue of long term availability (and funding) of services, infrastructures and resources is also pointed in the KIT answer that explains: "EC and national funding agencies have to change their thinking – project-based funding is not sufficient to make a “game changer”. And nothing less than a “game changer” in science is envisaged with EOSC." This is a very important point related to the EOSC adoption by large communities.

Security may seem to be technical at a first glance. But as underlined by EUDAT and the e-IRG in its roadmap there is a lack of common security policy (Security policies and related infrastructure in the member states are heterogeneous at the moment.) and the EOSC needs interoperability on the security level. As explains EUDAT when providing access across (e-) infrastructures a minimum agreed level of security should be applied and adhered by across the (e-)infrastructures. To achieve this goal, good initiatives are taken, for example via the Wise Information Security for e-Infrastructures (WISE) community what provides the Security for Collaborating among infrastructures (SCI) framework. WISE is governed by a steering committee and project-managed by staff from GÉANT. It is a unique experts group in this field. The *Security for Collaborating Infrastructures* (SCI) working group is a collaborative activity within the *Wise Information Security for e-Infrastructures* (WISE) trust community. The aim of the trust framework is to enable interoperation of collaborating Infrastructures in managing cross-infrastructure operational security risks. It also builds trust between Infrastructures by adopting policy standards for collaboration especially in cases where identical security policy documents cannot be shared.

By endorsing this framework, the Infrastructures can express and improve their security stance and foster trust among the global peers in facilitating interoperation and availability of services and data for research

and their collaborations. During the deliverable redaction (1st of June) SCI's updated version 2 was officially endorsed by EGI, EUDAT, GÉANT, GridPP, PRACE, SURF, WLCG and the XSEDE e-infrastructure in the United States. With this endorsement, the e-infrastructures subscribe to the governing principles and approach of SCI as a medium of building trust and exchanging information in the event of security incidents. This paves the way for other resources providers or infrastructures that are not currently members of such e-infrastructures. Endorsing this framework may be a condition for resources and services providers to be part of the EOSC infrastructure.

From these providers, it is a guarantee concerning their own resources and services.

From a customers or users points of view even though all communities are not well aware of these issues, the more advanced in ICT or those dealing with security issues will consider this type of framework as a proof of seriousness.

Managing these issues of security is certainly a real mean for the EOSC to foster its adoption.

Privacy is another point that may seem to be technical at a first glance but it requires agreements and collaboration between the different stakeholders and compliance with the different laws that apply to the storage, access to or analysis of the data. This was pointed out by the EGI, CDCI inputs and the ELIXIR WP8 and WP9 presented in the ELIXIR Compute Platform document that are related to Rare Disease and Human Genetic Data. The ELIXIR Compute Platform document is more related to technical solutions but the European Genome Phemone Archive manages agreements, protocols and policies. This is one example but other may exist that deal with other types of data whose access, analysis and storage have to be strictly controlled for whatever reason (business for private sector, embargo for scientific publications, personal data...). Managing the privacy is necessary and is already done by several (e-) infrastructures. Managing the privacy at the EOSC level should be a way to foster its adoption by users or customers with privacy needs.

4.2.4. Summary

In conclusion the main identified gaps and the main axes to build bridges may be summarized as:

Lack of common access policies: set up

- A global efficient Authentication and Authorization Infrastructure that federates existing AAI. Take care of the organisational challenge of such a global AAI according to AARC experts' propositions.
- A multidisciplinary mutualized space (or discipline agnostic) in the EOSC that potentially welcomes all researchers of all disciplines for their requirements that cannot be filled by other means.

Lack of knowledge of the computing landscape and of the different e-infrastructures: set up

- A common vocabulary (at least a common published glossary).
- Improve awareness among users.
- Global services catalogue that presents all participants' services portfolios.

Adoption:

- Improve skills, know-how and users support among administrators and sites representatives, foster experience and expertise sharing to create the conditions for technical experts to find and mutualize the right ways to support their users.
- Improve skills of users (hands-on or tutorials).
- Foster networking activities to share expertise among sites.
- Foster networking activities among users and user communities to share experience and tools.
- Consider input of ad hoc projects selected after investigation of requirements in order to complete the panel of science demonstrators.
- Take into account needs of all stakeholders even outside the partnership of the project itself.
- Adopt a standardized / transparent / stable in time approach to provision of resources and services, with pricing models fitting well into standard funding scheme of the scientific communities.
- Take into account security issues and study the opportunity to endorse WISE framework.

- Take into account privacy issues at EOSC level.

4.3. First list of initiatives and projects to collaborate with

This part gives a first list of initiatives and projects to collaborate with. These initiatives and projects are included in the following analysis because they work on important aspects of the interoperability at different levels and they can leverage experts or resources. The EOSCpilot project could not only give them complementary inputs but also use some of their solutions.

It is also important to note that several sites proposed to contribute to EOSCpilot initiative in their answers to the questionnaire.

4.3.1. AARC2

AARC2 is the second AARC[39] project. It gathers AAI experts from the main European e-infrastructures and engages with scientific communities. This second project will build on the first project results. Its kick off has just been organised in June 2017. As explained on its website AARC2 will take two main routes:

- AARC will expand efforts to engage with target communities by working closely with them to disseminate information, ensure messages are clear, deliver training, gain feedback, and implement the AARC framework. The number of partners has increased from 20 to 26, helping to make this approach easier.
- AARC will shift the technical focus from the question of how to authenticate the identity of users to the question of how to authorize permitted users to access the resources they require, across the boundaries of different infrastructures and research collaborations.

These two routes are of great interest for the EOSC because they deal with the main requirement of this gap analysis: AAI as discussed in the 4.1.1 part of this document.

AARC already provided information on how to share accounting data between two infrastructures that can be of interest in EOSC.

AARC has strong links with WISE that guaranties consistency in their work. Several partners of EOSCpilot are also partners of AARC2. This is a positive situation.

4.3.2. eInfraCentral

eInfraCentral[40] is a Horizon2020 project that started in January 2017. Its mission is to ensure that by 2020 a broader and more varied group of users (including industry) discovers and accesses the existing and developing e-infrastructure capacities. A common approach to defining and monitoring e-infrastructure services will increase the uptake of and enhance understanding of where improvements can be made in delivering e-infrastructure services.

eInfraCentral's work is related to the EOSC and deals with e-infrastructures services' interoperability. It is obviously a project to cooperate with. Common partners in eInfraCentral and EOSCpilot should facilitate the interactions.

4.3.3. INDIGO-DataCloud

INDIGO Datacloud is presented in details in the section 2.4.

We mention here just some of the project strengths

- Based on **Open Source** solutions, it will also develop Open Source software.
- Rooted in use cases and supported by **multi-disciplinary scientific communities**, big and small.
- It exploits **available, general solutions** rather than on custom, homemade specific tools or services.

[39] https://aarc-project.eu/
[40] http://einfracentral.eu/

- The framework or services offered to final users as well as to developers will have a **low learning curve**, with popular existing software suites will be supported and exploitable by INDIGO software in a transparent way.

INDIGO-DataCloud aims to allow running the software in a **hybrid, distributed Cloud** environment.

This project is often cited in the different answers and its results are expected may help many infrastructure providers. It is really necessary for EOSCpilot to share use cases and needs with INDIGO-DataCloud, better understand and propose to the Science Demonstrators project's solutions known to address any of the gaps we identify. Common partners should facilitate the interactions.

4.3.4. WISE

The Wise Information Security for e-Infrastructures (WISE)[41] community what provides the Security for Collaborating among infrastructures (SCI) framework. WISE is governed by a steering committee and project-managed by staff from GÉANT. The Security for Collaborating Infrastructures (SCI) working group is a collaborative activity within the Wise Information Security for e-Infrastructures (WISE) trust community. The aim of the trust framework is to enable interoperation of collaborating Infrastructures in managing cross-infrastructure operational security risks. It also builds trust between Infrastructures by adopting policy standards for collaboration especially in cases where identical security policy documents cannot be shared.

By endorsing this framework the Infrastructures can express and improve their security stance and foster trust among the global peers in facilitating interoperation and availability of services and data for research and their collaborations. During the writing of this document (June 1st, 2017) SCI's updated version 2 was officially endorsed by EGI, EUDAT, GÉANT, GridPP, PRACE, SURF, WLCG, and the USA's XSEDE e-infrastructure. With this endorsement, the e-infrastructures subscribe to the governing principles and approach of SCI as a medium of building trust and exchanging information in the event of security incidents.

4.4. Main findings

4.4.1. Technical aspects

Authentication and Authorisation Infrastructures (AAI)

AAI is *the* cornerstone of interoperability and the main issue that prevents exploitation and usage of existing e-infrastructures and distributed resources. However, this topic has been studied for a long time and the experts of the AARC projects are specifically working on this matter. The EOSCpilot may take advantage of this work and if possible contribute its part of use cases and inputs. It is important to avoid double work and to integrate existing experts in order to benefit from their expertise. The EOSCpilot could benefit from the work done by the AARC projects during the previous years.

Data

A large variety of tools and solutions has been successfully applied already for a long time in the research communities and e-infrastructures. However, certain cases are not currently covered in a distributed environment (hosting of privacy sensitive data, data accounting) and data size in certain use cases exceeds the capabilities of current infrastructures. Data transfer may be difficult due to the lack of transfer of the access rights with the data and of common APIs and due to the low capacity of certain parts of the network, and AAI services could solve partially these issues. There is a need for smart data placement, and for standardised APIs. Data mining, long-term preservation and quality of service are also of interest.

Computing

Resources providers are willing to serve their user community whatever their computing needs (HPC, HTC, cloud). This is a positive input for the EOSC. It is important that bridges between the different

[41] http://wise-community.org/

infrastructures are developed to allow users to use all types of resources they need in their scientific workflows in a seamless manner.

Cloud computing and Container Management are both technologies that require tests and improvements within the context of EOSC.

Network

Network service is a key component for the success of the EOSC. NRENs and GÉANT are able to contribute to the EOSC with a stable, robust, and rather ubiquitous services.

The gap analysis overall signals a possible need for increase of automatism and dynamicity. The integration with services and e-Infrastructure above the net is a clear gap. The following items provide more details on this matter:

- An extensive end-to-end service (end user to EOSC e-Infrastructure) is still a challenge, as it transits not only NRENs and EOSC itself, but also often a local domain (university, mobile provider, regional network). Policies and technologies need to be refined to ensure EOSC is available to all researchers.
- A collaboration between NREN and EOSC will help
    - To improve the quality of the last mile connectivity;
    - To define requirements and policies towards commercial clouds;
    - To define a strategy on the use and interconnection of large academic cloud;
    - To develop and define bandwidth guarantee requirement;
    - To enforce security at the network level and at the whole EOSC.
- The NRENs must be aware of user requests in order to be able to quickly provide to them a virtual network infrastructure.
- Automation and orchestration developments will speed-up network service delivery, however the process may be complex due to software development and integration policies.
- Network and services' monitoring has to be further developed and integrated as a unavoidable starting point for management, SLA enforcement and accounting.

Core infrastructures

Accounting and traceability are necessary.

Portals and user-friendly services

The community demands Web portals and user-friendly workload managers such as DIRAC.

4.4.2. Political, social and cultural aspects

The main identified gaps and the main axes to build bridges regarding political, social and cultural aspects may be summarized as:

Lack of common access policies: set up

- A global efficient Authentication and Authorization Infrastructure that federates existing AAI.
- A multidisciplinary mutualized space (or discipline agnostic) in the EOSC that potentially welcomes all researchers of all disciplines for their requirements that cannot be filled by other means.

Lack of knowledge of the computing landscape and of the different e-infrastructures: set up

- A common vocabulary (at least a common published glossary).
- Improve awareness among users.
- Global services catalogue that presents all participants' services portfolios.

Adoption:

- Improve skills, know-how and user support among administrators and centres representatives, foster experience and expertise sharing to create the conditions for technical experts to find and

mutualize the right ways to support their users.
- Improve skills of users (hands-on or tutorials).
- Foster the development and the deployment of more easy-access tools and a higher level of user friendliness especially to ease access and use of individuals and scientific communities, which have not yet been exposed a lot to the e-infrastructures.
- Foster networking activities to share expertise among centres.
- Foster networking activities among users and user communities to share experience and tools.
- Consider input of ad hoc projects selected after the investigation of requirements in order to complete the panel of Science Demonstrators.
- Take into account needs of all stakeholders even outside the partnership of the EOSCpilot project itself.
- Adopt a standardized / transparent / stable in time approach to provision of resources and services, with pricing models fitting well into standard funding scheme of the scientific communities.
- Take into account security issues and study the opportunity to endorse WISE framework.
- Take into account privacy issues at EOSC level.

4.4.3. First list of initiatives and projects to collaborate with

AARC2, INDIGO-DataCloud, eInfraCentral, and Wise are identified in this analysis as important projects to collaborate with.

5. CONCLUSIONS

We performed an "e-Infrastructure Gap Analysis" of the current issues preventing exploitation and usage of existing e-infrastructures and distributed resources with the aim to provide architectures that will allow overcoming these gaps. The analysis was based on the input from European e-infrastructures and service providers.

The main findings of this gap analysis are diverse and are summarised in Section 4.4 above. The main gaps are presented in the Figure 4.

Figure 4: Main gaps

The main challenges are arising from the wish to

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

Facing these challenges, communities have already developed solutions. Thus, we can envision overcoming the gaps by building bridges. These bridges include

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

The main bridges to be built are summarised in Figure 5.

Figure 5: Main bridges to be built

The gaps seem large but the European e-infrastructure community has already built many bridges. The EOSC is an excellent opportunity to create an environment, where these bridges are solid and maintainable for the diverse communities, e-infrastructures, and technologies, which are going to support the challenges of the future.

ANNEX A. ANNEXES

A.1. List of organisations and infrastructures contacted

The research communities' survey version[42] was sent to:

- the shepherds of the 5 science demonstrators (directly and via the WP4 management),
- the IndexMed[43], Eurofidai[44], and Phenome[45] consortia that may be considered as part of the long tail of science. We received answers from IndeMed and Phenome.

The e-infrastructure survey version[46] was sent to:

- the WP5 contact that participated to the elaboration of the survey
- the representatives of PRACE, EUDAT, EGI and we received answers form EUDAT and EGI.
- the representatives of different Grid and HPC computing centres:
- CDCI in Switzerland. We received its answer.
- Karlsruhe SCC, FZ-Jülich, DESY in Germany. We received their answers.
- CNRS-CC-IN2P3, CNRS-IDRIS, GENCI in France. We received the answers from CC-IN23 and IDRIS.
- the Grid centres in Italy via their mailing list. We received answers from INFN-Padova and INFN Roma.
- the Grid centres of the French NGI via their mailing list. We received answers from the NGI, MCI, CNRS-IN2P3-IRES and MCA.
- the regional centres in France via their mailing list. We received the answers from GRICAD, PMSN and Unistra.

Several reminders have been sent to get the answers.

A.2. Description of the main European e-infrastructures

GÉANT

The GÉANT network continues to set the standard for speed, service availability, security and reach, delivering the high performance that more than 50 million users rely on.

GÉANT 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.
The network is essential to Europe's e-infrastructure strategy, supporting open science with a terabit-ready e-infrastructure and advanced networking services for trusted access. It offers the highest levels of capacity and security users need, where and when they need it.

GÉANT's role in Europe is unique: by interconnecting Europe's NRENs, we bring Europe's brightest minds together to collaborate virtually and accelerate research, drive innovation and enrich education.

[42] https://repository.eoscpilot.eu/remote.php/webdav/WP6 - EOSC Interoperability/Task 6.1 e-infrastructure gap analysis %26 interoperability architecture/Input-for-EOSCpilot-gap-analysis-science-demonstrators.docx
[43] http://www.indexmed.eu
[44] https://www.eurofidai.org/en
[45] https://www.phenome-fppn.fr/phenome_eng/
[46] https://repository.eoscpilot.eu/remote.php/webdav/WP6 - EOSC Interoperability/Task 6.1 e-infrastructure gap analysis %26 interoperability architecture/Input-sites-and-e-infras-for-EOSCpilot-gap-analysis-v3.2.docx

GÉANT develops the services its members need to support researchers, educators and innovators - at national, European and international levels.

GÉANT's portfolio of advanced services covers connectivity and network management, trust identity and security, real-time communications, storage and clouds and professional services.

EGI

EGI is a publicly funded e infrastructure put together to give scientists access to more than 530,000 logical CPUs, 200 PByte of disk capacity and 300 PByte of tape storage to drive research and innovation in Europe. The infrastructure provides both high throughput computing and cloud compute/storage capabilities. Resources are provided by about 350 resource centres which are distributed across 56 countries in Europe, the Asia Pacific region, Canada and Latin America. EGI is a federation of cooperating resource infrastructure providers, working together to provide the leading edge computing services needed by European researchers.

The EGI resource infrastructure providers are:

- **Core resource providers**: National Grid Initiatives (NGIs) and European Intergovernmental Research Organisations (EIROs). The National Grid Initiatives (NGIs) are organisations set up by individual countries to manage the computing resources they provide to the EGI. They represent the country's single point of contact for government, research communities and resource centres as regards ICT services for e science. NGIs are the EGI main stakeholders, together with CERN and EMBL.
- **Integrated resource providers**: International organisations which contribute resources and collaborate through Memoranda of Understanding,
- **Collaborating resource providers**: Other international organisations, which have strong collaborations with EGI.

The EGI Platform Architecture was designed to support a broad consumer base with a very diverse set of requirements, and with a much smaller central coordination effort. The platforms EGI is based on are the following:

- **EGI Core Infrastructure Platform**, to operate and manage a distributed infrastructure (federated operations e.g. accounting),
- **EGI Cloud Infrastructure Platform**, to operate a federated cloud based infrastructure (more about the EGI Federated Cloud),
- **EGI Collaboration Platform**, for information exchange and community coordination (e.g. AppDB),
- **Community Platforms**, tailored service portfolios customised for specific scientific communities.

The platform architecture allows any type and any number of community platforms to co exist on the

physical infrastructure.

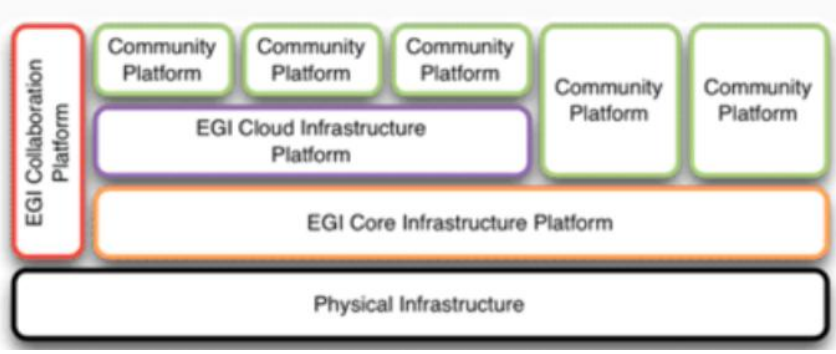

Figure 6: EGI platform architecture

The **EGI Core Infrastructure Platform** is composed of a set of fundamental technical services and support processes necessary for any type of federated e infrastructure exposing resources for shared use. This includes technical subsystems, such as Accounting, AAI, Monitoring, as well as non-technical services, such as first point of contact, Helpdesk and support services, federated security policy management and advice, as well as operational security activities, such as Computer Security Incident Response, Risk Assessment, and Vulnerability Handling. This platform, unlike other EGI ones build on top of it, does not include any technology availing ICT resources to EGI's user communities.

The **EGI Cloud Infrastructure Platform** incorporates at its heart IaaS Cloud deployments, which primarily deliver IaaS Compute and IaaS Storage services. Integrating with the Core Infrastructure Platform, it inherits all its federation properties and supplementary services without having to re implement them. Since institutional resource providers wish to retain control over which Cloud Management Framework (CMF) they wish to deploy, the federation technical architecture models institutional Clouds as abstract entities that must implement/integrate with mandatory interfaces. Southbound interfaces are defined through the integration with the underpinning EGI Core Infrastructure Platform; northbound interfaces define how Cloud managed virtualised resources are made available to the users within the framework of the federation. Together, these form a set of federation requirements that each member resource provider must meet, irrespective of the individually chosen CMF.

The **EGI Federated Cloud** is a multi national cloud system that integrates institutional clouds into a scalable computing platform for data and/or compute driven applications and services. The EGI Federated Cloud brings together scientific communities, R&D projects, technology and resource providers to form a community that integrates and maintains a flexible solution portfolio that enables various types of cloud federations with IaaS, PaaS and SaaS capabilities. The collaboration is committed to the use of open source tools and services that are reusable across scientific disciplines. The EGI Federated Cloud provides the services and technologies to create federation of clouds (community, private or public clouds) that operate according to the preferences, choices and constraints set by its members and users.

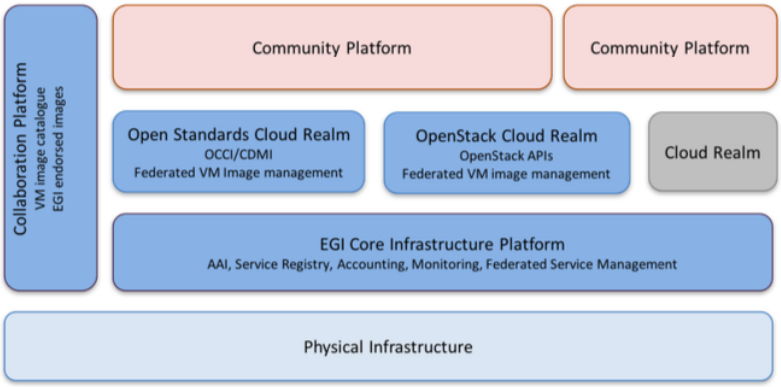

Figure 7: EGI federated cloud model

EUDAT

EUDAT Collaborative Data Infrastructure (CDI) has been established by EUDAT project [reference] and is now being further developed by EUDAT2020 [references]. EUDAT CDI is a cooperation of 35 partners including Europe's largest academic data centres and scientific repositories. EUDAT services are targeted to multiple large research communities as well as the individual scientists. EUDAT CDI delivers the data management services and solutions for long-term data preservation of massive scientific datasets as well as easy sharing and publishing the research results by long tail of scientists. It also provides solutions and services for efficient and reliable transfer of the data to, from and among EUDAT data centres. Auxiliary services include data discovery and exploration; meta data harvesting and meta data based search functionality. EUDAT services suite is presented in the Figure below.

EUDAT services:

- B2SAFE – fundamental service, it ensures long-term data persistency through automatic data replication and periodic integrity checks.
- B2STAGE - supports effective and reliable data ingest and staging to and from EUDAT premises to HPC and cloud computing facilities.
- B2SHARE - serves as an easy data sharing and publication platform as well as B2DROP that is a EUDAT's replacement of Dropbox
- B2FIND - enables data discovery and exploration based on meta data harvested from aforementioned EUDAT services as well as external meta data stores, e.g. user community driven.
- B2ACCESS – AAI solution for controlling the access to EUDAT, that enables authentication based on SAML, social and OpenID IdPs versus both web based (SAML) and non-web (X.509) EUDAT services. The EUDAT AAI landscape is complex: on the one hand there are various authentication and authorization methods already used by different communities, on the other there are different methods used by EUDAT services.
    - B2ACCESS works as a proxy that bridges various AAI technologies used on the one hand by the user communities (SAML/eduGAIN, OpenID Connect, etc.) or social identity providers (Google, Facebook, LinkedIn) and on the other hand offered by various EUDAT services, e.g. B2SHARE (web) and B2SAFE (non web) as well as external infrastructures. B2SAFE

performs token translation based on Unity [reference] and enables account synchronisation through a pull API. It also allows attribute management they can be added by user or community manager and pushed to or pulled from other services.

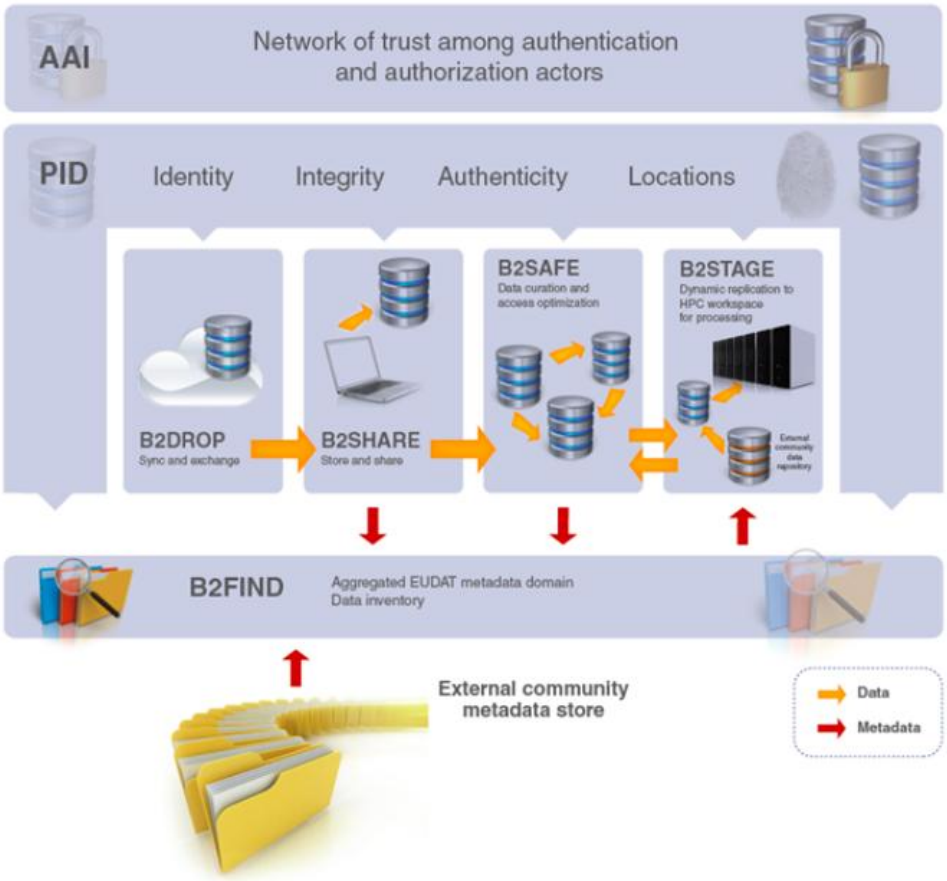

Figure 8: EUDAT services suite

PRACE

Partnership for Advanced Computing in Europe 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. The resources are provided in several forms such as Preparatory Access (short term access for code porting and tests), Project Access (long term research from 1 to 3 years for research groups) and SHAPE (SME HPC Adoption Programme in Europe). PRACE HPC resources are divided into Tier 0 (large machines from major project partners) and Tier 1 (smaller machines from other partners). PRACE manages a catalogue of services, both for administration and management as well as for end users, divided into the following categories:

- **Network** – the current PRACE network architecture is a 10Gb star topology,
- **Data** – main services provided in this category include GridFTP, GPFS and gtransfer,
- **Compute** – compute services allow users to submit jobs to the HPC centres where they have been granted access and include UNICORE, Globus GRAM5 and various schedulers available on particular sites,

- **Authentication, Authorization and Accounting** – user authentication is based on a central LDAP server which stores users DN's. Users are then provided services such as MyProxy and GSI SSH,
- **User Support** – these services offer direct support for users in terms of documentation and ticket based support system,
- **Monitoring** – distributed monitoring system is deployed between all sites, currently based on Icinga solution.

Helix Nebula

Helix Nebula is a partnership between major commercial and scientific institutions in Europe aiming at provision of sustainable computing Cloud service, focusing on such applications requiring access to large data sets from Large Hadron Collider (CERN), molecular biology (EMBL) and Earth Observation (ESA). Helix Nebula architecture is based on the idea of a marketplace which integrates consumers and suppliers. Various platforms are connected to "Blue Box" component based on Open Source SlipStream [reference] platform, which provides a unified interface to users in various access models (REST API, EC2, command line, web interface). "Blue Box" also provides federated identity management functionality, integrating security mechanisms from different suppliers using standards such as SAML, multi factor authentication and OpenID. The main goals of the Helix Nebula project include [reference]

- Establish a Cloud Computing Infrastructure for the European Research Area and the Space Agencies, serving as a platform for innovation and evolution of the overall federated cloud framework,
- Identify and adopt suitable policies for trust, security and privacy on a European level,
- Create a lightweight governance structure that involves all the stakeholders and can evolve over time as the infrastructure, services and user base grows,
- Define a funding scheme involving all the stakeholder groups (service suppliers, users, EC and national funding agencies) into a Public Private Partnership model, that delivers a sustainable and profitable business environment adhering to European level policies.

On the IaaS layer the components are divided into 3 categories: Processing (Virtual Machines), Storage (various technologies, support for Hierarchical Storage Manager and Information Lifecycle Management) and Network (VLAN, DMZ, GÉANT).

The Helix Nebula from its beginning envisioned possibility and need of integration with other e Infrastructures such as EGI, and procured interoperability requirements analysis deliverable D6.1 [reference]. Integration with Helix Nebula on technological level will involve integration with its central component called "Blue Box" which is an implementation of a Service Enabling Framework. The "Blue Box" acts as an intermediation layer between users and suppliers covering all functionality including identity management, monitoring, service catalogue, provisioning and others.

A.3. Description of AAI technical solutions deployed in the main e-infrastructures

eduGAIN and eduTEAMS

The eduGAIN service interconnects identity federations around the world, simplifying access to content, services and resources for the global research and education community. eduGAIN enables the trustworthy exchange of information related to identity, authentication and authorisation (AAI).

eduGAIN:

- Helps students, researchers and educators access online services while minimising the number of accounts users and service providers have to manage - reducing costs, complexity and security risks;
- gives service providers access to a larger pool of users internationally, and allows users to access resources of peer institutions or commercial or cloud services using their one trusted identity.

With eduGAIN participants from around 2,390 identity providers accessing services from around 1,520 service providers, eduGAIN has fast become the primary mechanism to interfederate for research and education collaboration around the world.

eduTeams

Collaboration is at the heart of research. Research teams can be created which cross borders and continents to bring the right skills together wherever they are based.

eduTEAMS, provided by GÉANT, gives the capability to build, manage and control these virtual teams. Built on top of eduGAIN, eduTEAMS aims to simplify the management of group and authorisation information. It enables the integration users from a wide range of environment, connecting them to specific services (such as instruments), and also to other generic services such as storage and compute provided by any e-infrastructure provider or even commercial entity.

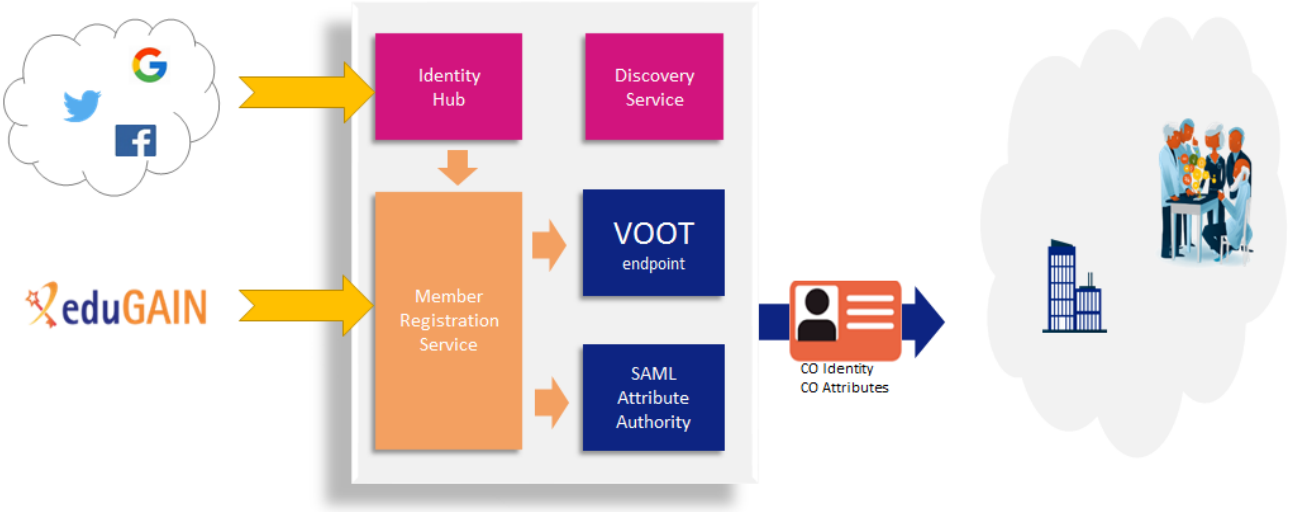

Figure 9: eduGAIN

Just like Identity Federations and eduGAIN, eduTEAMS aims at scalability and give this possibility of integration for services provided by any e-infrastructure, Research Infrastructure or long tail simple collaboration.

eduTEAMS works by allowing team administrators to add user identities into a virtual team. These identities can be from a range of organisations (both interfederated by eduGAIN and even third party identities).

Each identity can then be allocated a range of additional, team specific attributes. For example, a user from campus A and a team member using a Facebook identity can both be given specific roles within the team.

Team members can then be linked to services in groups.

New members can be allocated to a group and can inherit all the privileges and access of that group without having to manually add them to each service.

Equally when a member leaves their access to all services can be revoked with one instruction avoiding the need for the administrator to track and remove their access from each service individually.

This ability to add and remove users in a clear and consistent manner improves productivity and security of the team's data and services.

All users maintain their individual identities and do not need to be allocated or learn new usernames or passwords. Increasing security, reducing the overhead of managing teams and improving the user experience.

EGI CheckIn today:

- Available via eduGAIN
- IdP Discovery
- User Enrolment
- User Consent
- Support for LoA

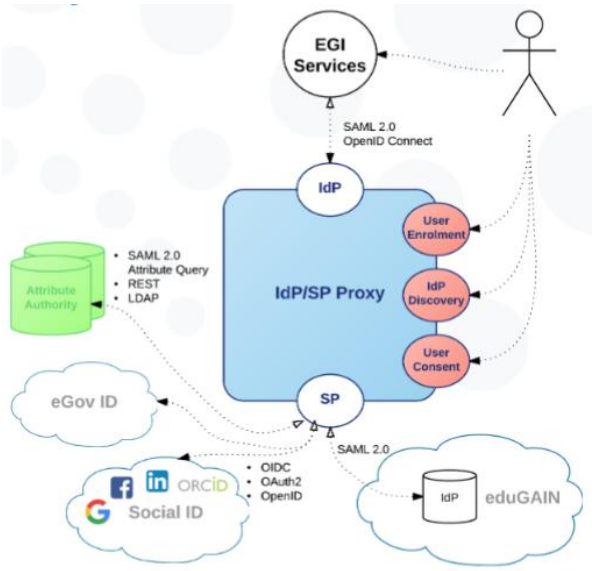

Figure 10: EGI CheckIn

- Attribute Aggregation - SAML2.0 Attribute Query, REST, LDAP
- Support for OIDC/OAuth2 Providers
  - Google, Facebook, LinkedIn, ORCID
- Support for OIDC/OAuth2 services
- Experimental support for eIDAS

EGI CheckIn is integrated in the AAI ecosystem:

- Integration with the ELIXIR AAI – it enable access for ELIXIR users to:
  - EGI configuration database (GOCDB)
    - Only for the members of the ELIXIR group: "Community:Compute:Grid site managers"
  - Virtual appliance registry (AppDB)

        - For the members of the “vo.elixir-europe.org”
        - Entitlements from ELIXIR AAI identify the members or the manager sof the VO, with different capabilities
    - Raising the LoA for the users holding the entitlements above
- Integration with EPOS
    - Integrated as an IdP for EPOS users (still in the test environment but can be moved to production upon request)
    - In the future CheckIn will reuse the Unity connector (see LToS use case) to allow EPOS users to access EGI services
- Integration with EUDAT
    - CheckIn will be integrated as both an IdP and an SP with EUDAT B2ACCESS to support the following cross-infrastructure use cases:
        - EGI users accessing EUDAT (web and non-web) resources using their EGI CheckIn ID
        - EUDAT users accessing EGI (web and non-web) resources using their EUDAT
- Integration with ARIA, structural biology
    - Integrated as an IdP in CheckIn
- First VO fully integrated: LToS
    - Integrated as an IdP for LToS users -> VO fully accessible through CheckIn
    - CheckIn has developed a Unity connector to get LToS VO membership information and generate entitlements (supported use case: AppDB uses vm_operator role of LToS users for AuthZ)

INDIGO IAM

Authentication and authorization in INDIGO is managed by a separate service called IAM (Identity and Access Management). IAM manages identities, group and VO membership and authorization policies. To enable integration with various platforms and other AAI legacy systems, the IAM provides support for SAML, X.509 and OpenID connect user authentication. The IAM acts as an authoritative source for authorization attributes and policies, to enable consistent authorization across the infrastructures. IAM will redirect non authenticated users to their local, trusted Identity Providers and based on its decision a token will be generated which will be used by other INDIGO services. The INDIGO IAM service fully integrates with other AAI related projects such as AARC, to which some of the partners of the INDIGO consortium also participate. The **Identity and Access Management Service (IAM)** provides a layer where identities, enrollment, group membership and other attributes and authorization policies on distributed resources can be managed in an homogeneous way, supporting the federated authentication mechanisms supported by the INDIGO AAI. The IAM service provides user identity and policy information to services so that consistent authorization decisions can be enforced across distributed services. It provides the following functions:

- **Authentication**: The IAM supports the authentication mechanisms defined by the INDIGO AAI architecture (SAML, X.509, OpenID connect)
- **Session management**: the IAM provides session management functionality. Sessions are used to provide single sign-on and logout functionality to client applications and services.
- **Enrolment**: The IAM provides enrolment and registration functionalities, so that users can join groups/collaborations according to user-defined flows.
- **Attribute and identity management**: The IAM provides services to manage group membership, attributes assignment, to group/collaboration administrators and the ability, for users, to consolidate multiple identities in a single INDIGO identity.
- **User provisioning**: the IAM provides endpoints to provision information about users identities to other services, so that consistent local account provisioning, for example, can be implemented
- **Policy definition, distribution and evaluation**: the IAM provides tools and APIs to
    - define authorization policies on distributed resources

- define policy distribution flows so that policies can be imported from other IAM instances
- evaluate policies against a request context and issue an authorization decision

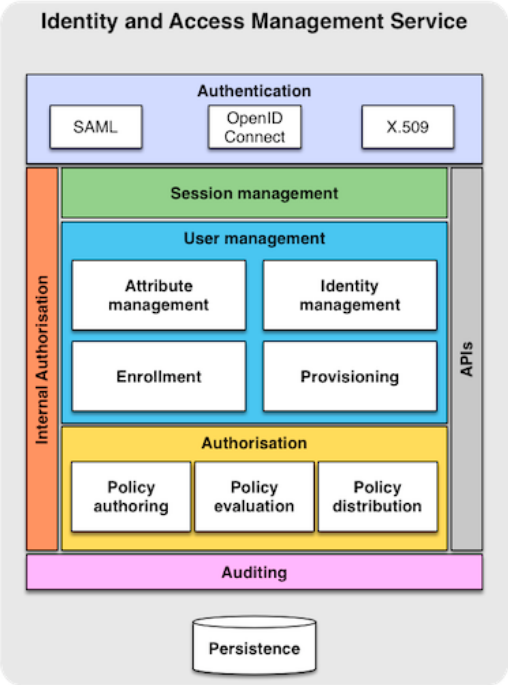

Figure 11: INDIGO IAM architecture

Integration in AAI ecosystem:

- with EGI
  - the EGI authentication and authorization architecture is currently based on the IGTF certification authorities for the provisioning of the personal user certificates. The authorization in EGI is commonly based on the user's membership in a Virtual Organization: user's attributes are managed by the Virtual Organization Management System (VOMS) that issues an attribute certificate (AC) describing the user membership. This information, which is included in the X.509 proxy used by the user to access EGI grid and cloud services/resources.
  - The INDIGO Identity and Access Management (IAM) can provide a set of advanced interoperable solutions: a login service (based on OpenID Connect) able to federate different IdPs and harmonize identity; a group membership service, that will allow to group users in organizations and decorate a user identity with additional attributes; an authorisation service, that will provide the tools to define and enforce authorization policies over the protected resources; support for controlled delegation of privileges across services and credential translation, through the integration with the INDIGO Token Translation Service (TTS).
- with EUDAT
  - INDIGO's AAI has to be integrated with EUDAT's AAI in order to enable INDIGO users to deposit and download the data to and from EUDAT services as well as publish, share and explore data through EUDAT facilities.
  - Integration can be performed in the following way:
    - The authentication is based on the INDIGO's Login Service that translates various users' credentials (SAML, OpenID Connect and X.509) to OpenID Connect. This service may be used as a proxy to EUDAT B2ACCESS service.

        - B2ACCESS is based on Unity. In the proposed approach Unity performs token translation from INDIGO's OpenID Connect to credential types supported by EUDAT services. OpenID Connect and OAuth2 could be used for authentication against web-based services (e.g. B2SHARE). SAML assertions could be used with B2DROP. Access to non-web services (B2SAFE, B2STAGE) requires translating to X.509 certificates (based on EUDAT CA) as well as account synchronisation through a dedicated mechanism that enables creating accounts in iRODS and GridFTP servers.
        - In the future, authentication with HTTP based REST interfaces of B2SAFE and B2STAGE (currently developed and piloted) can be performed through OpenID Connect and/or SAML translation.
- With PRACE
    - PRACE relies on a central LDAP repository which contains user Distinguished Name (DN) and certificate mapping. PRACE trusts all certificates that are issued by authorities, which are members of IGTF, such as EUGridPMA.
    - Integration of AAI between INDIGO and PRACE should be straightforward in terms of allowing users already having certificates trusted by PRACE to be able to access services using INDIGO interface using IAM token translation service.
    - In order to enable users who only use OpenID or other authentication methods, IAM would need to generate a certificate for users to present to PRACE systems. This would however require integration with PRACE LDAP credential database including generation or mapping of user DNs.
- With Helix Nebula
    - With respect to AAI, Helix Nebula expects that end users can rely on Single Sign On solution for authentication to the infrastructure, although specific technologies may vary between service suppliers. The "Blue Box" provides 3 main means for authentication including: SAML Federation, Multi factor authentication, EduGAIN and OpenID. The authentication and authorization is different for platform providers, who provision systems through the Blue Box and for regular users who just use the platform.
    - In terms of integration with INDIGO, the easiest way would be to establish trust between HN OpenID provider and the INDIGO IAM service. This would enable authenticated INDIGO users to automatically gain access to the "Blue Box" functionality, according to their authorization rights.

A.4. Descriptions of storage and data management technical solutions deployed in the main e-infrastructures

With respect to storage and data management functionalities provided by INDIGO-DataCloud project, the focus is to provide PaaS integrators and users with a unified interface enabling both basic and advanced data management scenarios. The project is developing a set of components called Unified Data Access (UDA) layer, which includes Onedata [reference], FTS 3 [reference] and DynaFed [reference]. The Unified Data Access layer will provide a PaaS level interface for end users, including REST APIs and the CDMI (Cloud Data Management Interface) [reference] standard protocol. The UDA layer will coordinate transparent migration of data across federated infrastructures and data life cycle management aspects.

UDA will support major transport protocols such as GridFTP and WebDAV for data transfer between the federated centres, as well as POSIX access for users to directly access remote data sets without the need for pre staging, either to the computational nodes or to their personal workstations. Advanced features of UDA include custom metadata definitions, Quality of Service and Data Lifecycle Management policies.

EGI integration

- From the point of view of the storage access and management, the EGI e infrastructure is providing two quite different approaches. For both approaches the authentication mechanisms is based on

X509 certificates.

- **EGI Grid Storage infrastructure**
  - The storage management of the EGI Grid infrastructure is based on services called "Storage Elements". There are different middleware solutions depending on the implementations of the hardware and software storage infrastructure at each centre. The Grid Storage Elements allow data access via Webdav protocol. Using this protocol the user's application is able to stream data and randomly read them at needs.
  - For data transfer, all the EGI Grid Storage elements exploit GridFTP, as EUDAT and PRACE are doing. The management of the QoS is done via Storage Resource Management (SRM) protocol. The INDIGO project will be able to access files on the EGI Grid Storage Elements in terms of both transfers and read access, exploiting the protocols supported by EGI. In this way the end user exploiting the INDIGO platform will be able to import/export data from/to EGI Grid infrastructure and process them using application and/or services deployed with the INDIGO PaaS Layer.
- **EGI Cloud Storage infrastructure**
  - The storage management in the EGI Fed Cloud environment is mainly providing:
    - Block Storage capabilities - useful to host services and legacy application running on virtual machines that need a real and local POSIX file system to host the data. In the EGI Federated Cloud this type of storage is actually managed through OCCI, and the users are able to exploit block storage resources together with computing resources.
      - For this type of storage, the INDIGO project enables seamless use of Block storage available at each centre, as the definition of the needed storage is one of the components that constitute each TOSCA Template that is submitted to the IaaS instance. So each of the computing resources will be automatically defined with CPU, RAM and storage requirements.
    - Object Storage capabilities - intended to let the application and the user exploit a public available Web Service endpoint where the data could be accessed worldwide on the network. In order to access to this storage in the EGI Federated Cloud, CDMI interface is suggested as the standard that each site should adopt, regardless of the technological implementation defined at each site.
      - The INDIGO project is building on top of the Standard CDMI implementation to support also advanced features such as QoS and Data Lifecycle Management.
      - Moreover, on top of the single IaaS instance INDIGO provides, , a federation layer that will ease data access for the end users. With these enhancements, users will be able to exploit with a single and simple interface all the storage available in the federation. The federated storage will be available both through the CDMI interface that will be mainly used for data management purposes but also with simpler protocols (POSIX, WebDAV) in order to let the legacy applications to exploit Cloud Storage resources.

EUDAT integration:

- EUDAT delivers the data management services and solutions for long-term data preservation of massive scientific datasets as well as for easy sharing the research results by long tail of

scientists. It also provides solutions and services for effective and reliable transfer of the data to, from and among EUDAT data centres. Auxiliary services include data discovery and exploration service, meta data harvesting and meta data based search functionality and AAI.

- The EUDAT project focuses on the domain of the registered data. A fundamental feature of EUDAT is to ensure physical datasets availability by their redundant storage in multiple data centres. Datasets' referability is ensured by assigning persistent identifiers (PID). PIDs are immutable over time despite possibly changing physical location of the dataset replicas. This basic functionality is provided by B2SAFE service. For large datasets EUDAT provides means to effectively and reliably ingest and stage data from EUDAT data centres (B2STAGE service). EUDAT is also addressing the need of easy sharing of the datasets that are results of the scientific work. These data can be deposited and accessed by the web portal as well as addressed by using PIDs (B2SHARE). Datasets can be explored using B2FIND solution.
    - INDIGO software stack will be integrated with several services of EUDAT including B2SAFE, B2STAGE and B2SHARE. INDIGO's AAI solution has to be integrated with EUDAT's B2ACCESS.
        - **B2SAFE and B2STAGE.** For *large datasets transfers* to and from B2SAFE service (export, import) INDIGO software may use commands for direct interaction with iRODS servers. Alternatively, iRODS APIs can be used, e.g. python irodsclient [33] that is most interesting. It is a Python binding to iRODS, which supports most of the functionality of i commands except sophisticated database queries, zone management and iRODS federation based file replication configuration. It is considered as pre alpha. In addition, GridFTP client functionality can be exploited for *efficient and reliable transmission* of the massive data set to and from EUDAT through GridFTP endpoints (B2STAGE) running in front of iRODS. These GridFTP servers can be accessed using standard GridFTP clients (e.g. UberFTP [34], globus url copy [35], GridFTP GUI [36], GridFTP GUI Client [37]) run by INDIGO users within their virtual machines or in workstations. EUDAT's GridFTP endpoints can be also integrated into the INDIGO workflow by using GlobusOnline service in an interactive way or through the Globus Online Transfer API [38] (used e.g. by Globus Online Python client library [39])

PRACE integration:

- PRACE offers only limited data storage capacities for users (especially in case of Tier 1 systems) and only for limited time frame, after which users have to move their data somewhere else. This could be a potentially interesting aspect to allow users of INDIGO's Unified Data Access layer to transparently migrate data between for instance EGI or EUDAT infrastructures to PRACE systems only for duration of computations.
- The main issue often faced by PRACE users is the need to transfer files from their local machines (either desktop or their facility local storage) to PRACE machines for computation and then get the results back. Currently PRACE supports 2 main ways to achieve this:
    - direct GridFTP
    - gtransfer itself is based on GridFTP, however it provides an easier to use interface. Both methods require the user to have a valid X.509 certificate.
- Integration with respect to data management will require establishment of trust between the INDIGO IAM service and PRACE authentication based on LDAP directory. INDIGO users with access to PRACE resources could benefit from the integration of the storage management of INDIGO and PRACE in 2 ways:
    - By directly mounting their virtual folders managed by the INDIGO Unified Data Access layer component called Onedata and transparently accessing data on PRACE machines without the need for pre staging,

- By using the FTS3 service for managed data transfer to and from PRACE machines.

Helix Nebula integration:

- In terms of data management, an interesting use case for integration with Helix Nebula would be integration and harmonization of QoS and SLA with respect to data management, focusing on accounting and billing. INDIGO-DataCloud develops a negotiation service for SLA, including data aspects, and this could be used by "Blue Box" billing functionality to enable HN users to request data storage on INDIGO supported infrastructures (e.g. EGI) with data management functionality.

A.5. Description of computing technical solutions deployed in the main e-infrastructures

The computation management functionality in the INDIGO-DataCloud project is controlled by a set of components, which coordinate both Cloud and Grid based deployments and execution of user services and jobs. The most user-facing component is Orchestrator, which accepts inputs defined in the TOSCA Simple Profile in YAML Version 1.0 [reference] language, and prepares the virtualized infrastructure based on virtual machines or lightweight containers for the execution of user's jobs or services. By default, Cloud infrastructures based on industry standard Cloud Management Frameworks including OpenStack and OpenNebula are supported, extending them with several features, as required by the end users. These include, among others, support for batch systems for job submission, fair-share scheduling and spot instance detection and hardware specific support such as GPGPU's or InfiniBand. Furthermore, both OpenStack and OpenNebula are being extended by INDIGO to support lightweight containers in addition to hypervisor based virtualization.

Integration of the EGI e-infrastructure can be at different levels, in order to allow end users to deploy in a transparent and powerful way both services and applications in heterogeneous environments such as Grid and Cloud.

- **Grid Job submission**
  - The scientific communities that need to submit jobs to EGI Grid infrastructure will take advantage of the advanced functions developed by INDIGO project in the area of the "Scientific Computational Portal as a Service": the user shall be able to request the creation of a portal to submit his/her Grid jobs; INDIGO PaaS will instantiate a customized service taking care of its high availability and scaling in order to face unpredictable workloads. The complexity of the job submission to the Grid will be drastically reduced for the end users that will be able to manage their jobs in a simple and easy way.
- **Application/Service deployment on Cloud**
  - The scientific communities that need to run a long running service or an application on EGI Federated Cloud will be able to deploy them in a simple and transparent way, using both APIs and web user interfaces, thanks to the functions provided by the INDIGO platform.
  - At the IaaS level, INDIGO will provide important extensions of the OCCI functionalities, already used in the EGI Federated Cloud, in order to address critical gaps like the possibility to run lightweight containers using the standard OCCI interface and support for essential network orchestration (creation of private networks, security groups, etc.). INDIGO also provides enhancements for the resource schedulers currently available in the major Cloud Management Frameworks (i.e. OpenStack and OpenNebula): requests will be scheduled in a smarter way allowing pre-emption, prioritisation and optimal usage of resources (e.g. for interactive usage patterns, fair share, etc.). The IaaS orchestration functions will be enhanced as well, supplying the CMF with orchestration

service endpoints compatible with the TOSCA template standard.

Some of the solutions interesting from the point of view of both user communities and e-infrastructures, that for sure worth testing also from an integration point of view are:

Infrastructure Manager

- Applications and services require customized computational environments that can be provisioned from multiple sources (e.g. on-premises Clouds, public Clouds, virtualization platforms, container orchestrators, etc.). However, the use of these platforms requires users to have non-trivial skills. For that, IM is a tool that **deploys complex and customized virtual infrastructures on multiple back-ends**. The IM automates the Virtual Machine Image (VMI) selection, deployment, configuration, software installation, monitoring and update of virtual infrastructures. It supports a wide variety of back-ends, thus making user applications Cloud agnostic. In addition **it features DevOps capabilities, based on Ansible** to enable the installation and configuration of all the user required applications providing the user with a fully functional infrastructure. It is a service that features a **web-based GUI, a XML-RPC API, a REST API and a command-line application**.
- The main goal of the IM is to provide a set of functions for the effective deployment of all the required virtual infrastructures required to deploy an application or service in a Cloud environment, either composed by VMs or by Docker containers. The IM considers all the aspects related to the creation and management of virtual infrastructures:
    - The **software and hardware requirements specification** for the user applications, using a simple language defined to be easy to understand by non-advanced users who just want to deploy a basic virtual infrastructure, but with enough expressivity for advanced users to set all the configuration parameters needed to get the infrastructure fully configured.
    - **The selection of the most suitable Virtual Machine Images (VMI)** based on the user expressed requirements.
    - **The provision of Virtual Machines on the Cloud deployments (or Docker containers in Kubernetes, for example)** available to the user, including both public IaaS Clouds (Amazon Web Services, Microsoft Azure, etc.), on-premises Cloud Management Platforms (OpenNebula, OpenStack, etc.) and Container Orchestrators (Kubernetes).
    - **The contextualization of the infrastructure at run-time** by installing and configuring all the required software that may not be available in the images (either VMIs or Docker images).
    - **The elasticity management**, both horizontal (adding/removing nodes) and vertical (growing/shrinking the capacity of nodes). The IM supports the TOSCA Simple Profile in YAML Version 1.0 for infrastructure description.
- The IM is currently being used in production in different areas. It is the component used for virtual infrastructure provision in EC3 to deploy elastic virtual clusters on multi-clouds. It is used to automate the deployment process of complex infrastructures for medical services (TRENCADIS) and for educational organizations (ODISEA).
- The service evolved in the INDIGO-DataCloud project (https://www.indigo-datacloud.eu/). It is used by the INDIGO Orchestrator to contact Cloud sites to finally deploy the VMs/containers.
- New features added:
    - Support for TOSCA 1.0 YAML specification with the custom node types described in https://github.com/indigo-dc/tosca-types/blob/master/custom_types.yaml
    - Support for the Identity and Access Management Service (IAM).
    - Support for the Token Translation Service (TTS) to support IAM authentication on OpenNebula Clouds.
    - Improvements to access OpenStack Clouds that support IAM
-

Udocker

At present the usage of containers is very limited in multiuser environments such as High Performance Computing mainframes, Linux Clusters or Grid infrastructures. The main limitation is that execution of containers using Docker or LXC requires having root access privileges in the host system. A tool that supports the execution of containers in user space is ***udocker***.

Containers – a few facts:

- "Linux Containers" is a technology provided by the Linux kernel, to "contain" a group of processes in an independent execution environment: this is called a "container".
- Advantages of Containers versus classical Virtual Machines:
  - Containers are much more light-weighted than a Virtual Machine
  - They provide an enormous simplification of the software deployment processes
- The most extended software to build containers is called Docker
- Docker is optimized for the deployment of applications as containers

One can generate a completely tailored "docker image" of any Linux Operating System, with all the required libraries, compilers, source codes,... which later on can be run as a container

Adoption of docker is being very slow in computing farms or interactive Linux-system shared by many users:

- The typical situation is that docker is not installed, and one cannot run containers without support from the system software.
- The main issue is that docker needs root permissions to run a container.
- Even though the user, within the context of the container is completely isolated from the rest of the machine, it raises all the alarms among security people
- A user with access to docker can own the hosting system

The INDIGO-DataCloud project has developed a solution – **udocker:**

- It is a tool to execute content of docker containers in user space when docker is not available
- enables download of docker containers from dockerhub
- enables execution of docker containers by non-privileged users
- It is executed under the regular user id (no root privileges needed anymore).
- privileges are not used in any step not for running not for installing
- It can be used to execute the content of docker containers in Linux batch systems and interactive clusters managed by others
- Acts as a wrapper around other tools to mimic docker capabilities

A.6. GÉANT Connectivity / Network management & monitoring services

GÉANT develops the services its members need to support researchers, educators and innovators - at national, European and international levels. Our portfolio of advanced services covers connectivity and network management, trust identity and security, real-time communications, storage and clouds and professional services.

The connectivity services support the NRENs in delivering world-class network facilities to the research and education community. The GÉANT network offers the high speed and huge capacity essential for the world's biggest science projects, setting the standard for speed, service availability, security and reach, delivering levels of performance that commercial network providers cannot provide.

Connectivity

The GÉANT and NREN networks underpin the work of a wide range of e-infrastructure and scientific

research projects by providing a high performance, reliable and cost-effective communications platform across the research and education (R&E) community. Service options cover IP, dedicated private connections, virtual private networks and roaming options.

GÉANT IP provides general-purpose IP transit for national research and education networking (NREN) organisations and other approved research and education partners and providers. Its core function is to provide a private service for IP (internet protocol) traffic that is separated from general-purpose access to the internet. Working at speeds of up to 100Gbps, GÉANT IP provides core connectivity that supports inter-NREN connectivity.

Specialist Connectivity Services

Many performance-critical services require guaranteed performance levels and additional security that is difficult to achieve through shared IP services. In particular, applications such as data centre backup and replication, real-time mission-critical services and broadcast quality video need the guaranteed bandwidth and low latency that only dedicated circuits separated from general IP traffic can offer.

Point-to-point services provide dedicated connectivity between two sites over the existing infrastructure without the cost and difficulty of building and managing a dedicated physical network. This type of connectivity can provide fixed latency between collaborating institutions, a high level of security and, if needed, guaranteed bandwidth of up to 100Gbps. Furthermore, the possibility of providing a L2 Ethernet channel end-to-end allows the use of network and transport protocol other than the classic TCP/IP, enabling users to experiment with new ways of using network connectivity. Good examples of such advanced use of the service are the SMARTfire and the InfiniCortex projects, using experimental streaming transport protocol and the InfiniBand network stack, respectively. There is also the long-lasting use in Radioastronomy and for the ITER project (linking the Cadarache facility in France to the Elios supercomputer in Rokkasho, Japan).

VPN Services Many projects may require teams across Europe to be able to collaborate effectively with enhanced privacy. By creating a Virtual Private Network (VPN), all sites on the VPN can communicate without the need to arrange separate physical networks, while benefiting from the privacy and security of a private infrastructure. GÉANT can provide VPNs between many sites over great distances within Europe and reach the USA (via Internet2 and ESnet), Canada (via CANARIE), and Asia.

The GÉANT Multi-Domain Virtual Private Network (MD-VPN) provides an end-to-end international network service that enables scientists all over Europe to collaborate via a common private network infrastructure. The MD-VPN service can be used for connectivity between clusters, grids, clouds and HPC (high-performance computing) centres, allowing them to form virtual distributed resources for third-party research projects. MD-VPN offers fast delivery of VPNs to end users and so can be used in a variety of ways, from a long-term infrastructure with a high demand for intensive network usage to quick point-to-point connections for a conference demonstration.

Open Interconnectivity As international and public-private partnerships grow in importance within the R&E sector, so the high- performance, flexible and neutral interconnection points provided by the GÉANT Open service can offer new opportunities. Users can connect their own circuits – at 1Gbps, 10Gbps or 100Gbps – and can then request interconnections with any other participant.

GÉANT Testbeds Service The GÉANT Testbeds Service (GTS) delivers integrated virtual environments as 'testbeds' for the network research community. GTS is designed for researchers of advanced networking technologies to help support testing and development over a large-scale, dispersed environment. GTS can support multiple projects simultaneously and isolates them from each other and from the production GÉANT network to provide security and safety. This facility is leading the way in providing facilities to help develop the next generation of internet services.

Network management & monitoring services

perfSONAR can be used by GÉANT, national research and education networking (NREN) organisations,

campuses and major projects for quick and easy performance troubleshooting. It provides easy, transparent end-to-end monitoring, giving access to network measurement data from multiple network domains. It can operate at local level or around the globe and is scalable to provide at-a-glance information about multiple network paths simultaneously. By tracking performance across and between domains it is now possible to identify and rectify any potential performance bottlenecks, helping research teams focus their efforts on their research and allowing NRENs to identify where investments in new capacity will provide the best return. With more than 1,400 measurement points across the globe, it is now far easier for NRENs and research teams to accurately measure network performance and ensure it meets their research needs. The development of perfSONAR is the result of the combined work of GÉANT, Internet2, ESnet and Indiana University.

eduPERT is part of GÉANT's commitment to helping network users get the best performance from their connections. Performance Enhancement Response Teams (PERTs) provide an investigation and consulting service to academic and research users on their network performance issues.

eduroam provides 50 million students and researchers with access to thousands of wi-fi access points in over 70 countries using a single, secure login facility - making international collaboration much easier. Over 5 million international logins a day are enabled by eduroam.

A.7. National and regional resources centres inputs

The centres answers to the questionnaire are presented here in the alphabetical order.

CNRS-CC-IN2P3; France

The CC-IN2P3[47] is a national HTC centre mostly dedicated to the High-Energy Physics community.

The challenges faced by the CC-IN2P3 are:

- a common/standardized API for authentication and authorization
- a common/standardized API for accessing computing and storage resources

It will be important to rely on a federated authentication infrastructure, adopted by everyone.

A well-designed network infrastructure and smart processes to manage data placement are also necessary in order to use computing resources from everywhere.

Traceability must be guaranteed for security purposes respecting local state laws.

Solutions that are interesting for CC-IN2P3 are mainly cloud (virtualization on demand) and storage resource solutions.

The CC-IN2P3 is using the Grid Middleware (gLite) and Dirac to provide resources to different communities of users around the world.

CNRS-IDRIS; France

IDRIS is a French national HPC centre offering resources based on technical and peer review scientific selection processes to users belonging to French scientific organisations.

One main issue is that the total amount of resources requested during the calls for proposals is always significantly higher than the available resources.

The interoperability issues are those of the users: how to better interact between the different infrastructures they use (Tier-2, Tier-3, other Tier-1, Tier-0):

- access, authentication and authorization issues (AAI)
- network protocols and services
- data transfer

[47] https://cc.in2p3.fr/en/

- data interoperability and reusability
- compiler, libraries, tools, batch managers
- workflow management systems and portals across different systems.

IDRIS is one of the partners of the Pico2 pilot project within the EOSC pilot to develop seamless data frameworks between Tier-1 and Tier-2 infrastructures, based on iRODS and multipoint dynamic VPN technology.

Another direction for IDRIS concerns the possibilities offered by new lightweight virtualisation technics. Some experiments are planned inside the PRACE project.

IDRIS experiments with some workflow managers (Unicore, Nice, SynfiniWay, SysFera DS) as well as distributed file systems (Avaki, Andrew File System, GPFS) since many years. This being said, the workflow managers and portal technologies tested so far have not convinced users of IDRIS. Nevertheless, there are still communities (in particular in life sciences) for which they know that this is not only a potential benefit for users, but a real requirement to attract them and to offer to not technical people a way to easily access to significant computational and data management resources, that they will not use otherwise.

This appears to be a critical point, especially for some communities (e.g. life sciences) who would need interfaces that allow also non-expert users to easily access the resources.

CNRS-IN2P3-IRES; France

CNRS-IN2P3-IRES is a Tier 2 French regional centre.

The main challenges identified are:

- standardized way (like SAML2) for users authentication for each service (grid, cloud, iRODS)
- French grid/cloud infrastructure is using certificates issued by a specific CA, which is not installed by default on users' operating system. Having certificates released by well-known CA (like TCS e-Science) would make access to infrastructure easier.

The infrastructure is connected to several federations (France Grilles, EGI, IFB) and provides strong expertise on OpenStack, which can be shared with EOSC partners, as well as the computing resources.

The centre also tests OpenIO storage solutions and propose iRODS storage infrastructure, which can be opened for tests in the framework of the EOSCpilot project.

The DIRAC interware is a very impressive tool that permits to address different types of computing resources (Cloud, HTC). Container support in DIRAC would be of great interest to easily use HPC resources. OpenStack is a promising cloud framework that can be used to federate distributed infrastructure resources. It provides a convenient way to provide HTC and HPC resources to users and administrators. Additional services (like SlipStream or components from the INDIGO DataCloud project) may be used to simplify the deployment of infrastructures across the sites.

Common Data Centre Infrastructure (CDCI) for Astronomy Astroparticle Physics and Cosmology, Université de Genève, Switzerland.

The CDCI is attached to the Astronomical Observatory of the University of Geneva and emerged from the ISDC Data Centre for Astrophysics (http://isdc.unige.ch/). The ISDC started its activities as the INTEGRAL Science Data Centre and is now actively contributing to several other space missions and ground-based projects with a prime scientific focus on high-energy astrophysics. The answers to the survey from this site are different probably because this centre is dedicated to a specific scientific community. For this reason the input is also presented in section 3 that gathers communities' inputs. The CDCI explains why generally the scientific collaborations use their own resources in the context of working with the CDCI.

Other services are not considered due to:

- intrinsic rules of collaborations around the projects and the associated privacy / data property issues,

- absence of a clearly defined framework for funding the cloud-based computing by funding agencies,
- high price of CPU time and storage at the national academic science cloud services (e.g. SWITCHengines) and
- absence of the guarantee of long-term commitments by the cloud service providers (the astronomical projects and space missions are developing on the time scale of decade(s)).

A standardized / transparent / stable in time approach to provision of cloud computing resources, with pricing models fitting well into standard funding scheme of space science and astronomical projects (through national science foundation(s), space agencies, European Southern Observatory, etc.) would increase proliferation of the use of cloud computing for the CDCI services. The services (access to astronomical data and data analysis pipelines) currently run locally, but the CDCI plans to use cloud computing to cope with peak loads (e.g. at the moments of re-processing of many-year data sets of space missions, before public data releases etc.).

The CDCI also plans to use cloud services to deploy data analysis pipelines for astronomical “big data” projects currently at development stages. The main interest will be to use EOSC to provision “on-demand” increases of CPU power and storage for peaks of data processing activities for all CDCI astronomical data projects and for deployment of data analysis pipelines in response to queries for the astronomical data products received through the web data analysis interface. Another particularly important aspect addressed by the CDCI infrastructure is a long-term preservation of data together with the full data analysis software system. CDCI’s activity is related to the data of space missions and astronomical observatories, but the problem is generically important and should be addressed across science disciplines. A dedicated study of reliable long-term (decades time scale) preservation of data analysis software systems together with the full raw data sets would be interesting to address within a broader context of the EOSC.

Best solutions: the CDCI is currently exploring the deployment of data analysis pipelines using the Docker-type container approach, which appears promising for interoperability and which also provides a solution for long-term preservation of data analysis software.

DESY; Germany

The most important challenges identified by DESY are:

1. AAI/IAM: an AAI solution which provides access to computing, storage/data, and web-services likewise is essential. Preferable would be a Single Sign-On solution, allowing federated access at the European level. Certificates are not an option for some of the scientific communities. For some large highly distributed communities, the currently used mechanism of providing external accounts doesn’t scale any more e.g. for the XFEL at DESY.
2. HDF5server and HDF product designer: HDFgroup has some services like HDF5server and HDF product designer, which allow native and distributed HDF5 access and registries of schemas for HDF5. Since HDF5 is the most widely used format (and the only de facto standard for binary data) it would be nice to support and provision services based on these. There are only few attempts out there, but HDF has demonstrated its capabilities in a cloud environment.
3. High performance data transfers between laboratories, serving the same communities, including the necessary AAI (see 1). Taking the amount of data that will be produced midterm the currently available access mechanisms are no longer sufficient.
4. Interoperable vocabulary and standardized mechanisms must be available through web portals to specify the quality of the data storage from high performance scratch disk to long term archiving. Due to the high number of scientists and the diverse requirements from the different scientific communities, interacting with computer centre stuff to discuss data storage qualities and capabilities it no longer an option. Provisioning of resources should follow the definition of requirements from users, potentially supported by intelligent systems translating from "user space language" to "storage provider language".
5. Providing access to private and experiment data through modern Web 2.0 and cloud mechanisms

will be more and more required. The 'afs' approach is no longer sufficient. Scientists must have the ability to share their data, including mass data, with individuals and groups. Synchronization with local and mobile devices is a must. DESY would suggest continuing its efforts in evaluating the INDIGO-DataCloud AAI solutions (OpenID Connect and IAM) and the interoperability work in terms of Quality of Service in storage. In case, the EGI propagated Onedata[48] storage federation solution turns out to be used by e-Infrastructures, DESY would suggest continuing its efforts to make the dCache Quality of service capability available through the federated Onedata system. In the area of Cloud Management Frameworks we would suggest to provide the results of INDIGO-DataCloud in making Containers management in OpenStack as well as investigating in the INDIGO improved scheduler and pre-emptibable instances for OpenStack. Their best solutions: In the area of Computing we are operating OpenStack and we try to get as many INDIGO extensions integrated (like the TOSCA translator), and in terms of data federations we are running DynaFed, a solution provided by the INDIGO-DataCloud project. In terms of data management, we are running dCache as a highly scalable storage backend, serving close to 20 PBytes, to support preconfigured or flexible storage qualities in conjunction with ownCloud/nextCloud as a sync and share interface to store, retrieve and share data.

France Grilles; France

France Grilles (http://france-grilles.fr) is the French National Grid Initiative since 2010. Structured as a Scientific Interest Group, it provides a portfolio of services on a distributed e-infrastructure for the storage and analysis of scientific data.

The biggest challenge for France Grilles has been to achieve usability and usage. Thank to grid technology, it was possible from early days to achieve interoperability of distributed resources within the framework of virtual organizations. The real challenge was to provide the services on top of the resources to hide the grid complexity from end users. This was achieved through the deployment of user-friendly services like Dirac and environments like VIP on top of multinational (Biomed) or multidisciplinary (France Grilles) virtual organizations. The biggest challenge for EOSC will be to achieve usability and usage. There is a major gap today between e-infrastructures and the vast majority of users: e-infrastructures provide services but the vast majority of potential users either ignores the existence of these services or ignores how to use them. In FP7 and Horizon 2020, EC has funded the construction of links between ESFRIs and e-infrastructures, the so-called "competence centres". ESFRI competence centres are completely failing at building links between e-infrastructures and user communities. Moreover, these competence centres are thematic by nature. As a consequence, the multidisciplinary applications, the transmission of knowledge across disciplines has been lost. Users have fled from e-infrastructure conferences that are now focussed only on technical and technocratic issues. The EC must fund again an intermediate layer of engineers who are knowledgeable about the EOSC catalogue of services and who work with all user communities. Interoperability of infrastructures will be achieved through the deployment of scientific applications across disciplines. France Grilles has been very successful at building interoperability across user communities and reaching out to the long tail by building a human network of engineers and scientists willing to share their expertise and know-how. It is of utmost importance for EOSC adoption to create such a human network across disciplines. This human network will build bridges across disciplines and as a consequence across infrastructures.

GRICAD; France

GRICAD is a local centre offering computing and storage resources to users in the area of Grenoble.

The challenges they face are:

- authentication and authorization management,
- access policies which must be homogenized between infrastructures,

[48] https://onedata.org/#/home

- network efficient connection,
- knowledge of existing infrastructure in order to guide the local users to the right resources,
- Financial: who funds the infrastructure, for whom?
- Data transfer.

Some solutions are interesting:

- distributed iRODS storage
- eduGAIN authentication
- connection with grid and cloud computing type resources as EGI
- monitoring of network connection (perfsonar)

Some of them are already been tested as iRODS and will be extended with the Pico2 pilot.

Notebooks server[49] is a very interesting way to make use of infrastructures easy to projects from the long tail of science.

INFN-Padova; Italy

This site is a Grid and Cloud resources provider. It represents an implementation of a unique cloud service, called "Cloud Area Padovana" (CAP), which encompasses resources spread over two different sites: the INFN Legnaro National Laboratories and the INFN Padova division. The "Cloud Area Padovana" project leverages on the long-standing collaboration between these two INFN sites, located about 10 km from each other. In particular, they have been operating a WLCG Tier-2 Grid facility distributed between the two sites [1], for both ALICE and CMS experiments, for many years. On the same site we find not only the Tier 2 LNL-PD grid site but also a small cloud infrastructure, part of the EGI FedCloud. So the administrators have a long experience in making different environments interoperate.

Our contact explains: As EGI FedCloud site manager and private local cloud provider via CAP, I am currently facing the issue of merging the two resources, i.e. adding the FedCloud resources to CAP and installing FedCloud components (especially keystone-voms and occi interface) without interfering with the local CAP deployment and configuration. A common AuthN/AuthZ mechanism would also be desirable (e.g. as the one provided by INDIGO-DataCloud IAM). I am currently testing Oneata as tool for facilitating shared storage within international geographically distributed user communities. I would like to have a solution that allows users to run jobs both on cloud and HTC resources with the same interface. We do not have HPC resources at our site.

INFN-ROMA; Italy

INFN-Roma is an ATLAS Tier 3. Again we have here an example of coexisting grid and cloud environments, as we can find not only a Tier 3 level grid-site but also one of the three partners on an implementation of a very innovative metropolitan area distributed cloud based data centre, based on OpenStack cloud management framework, the RMLab. This infrastructure puts together resources present in two INFN sites, Roma2 and Roma3, and a national laboratory, National Laboratory of Frascati, linked through a distributed layer 3 private network connection configured by GARR, the Italian NREN.

Those responsible of the site face as the currently most important challenge is the possibility to automatically scale out the infrastructure based on specific workloads and to be "multi tenant" ready (support of different type of analysis with different distros / libs). Investigating a distributed Docker based solution for data analysis might be a good starting point. INFN-Roma is now experimenting with an automation / scale out solution based on Ansible and VMWare. A poster that might give you a better insight has been provided and is available in the file repository[50].

[49] A notebook interface is a virtual notebook environment used by developers.

[50] https://repository.eoscpilot.eu/remote.php/webdav/WP6%20-%20EOSC%20Interoperability/Task%206.1%20e-infrastructure%20gap%20analysis%20%26%20interoperability%20architecture/Input-infrastructures/Input-INNFN-ROMA-Poster_2017-05_WorkshopCCR-Caruso.pdf

Jülich Supercomputing Centre; Germany

For the Jülich Supercomputing Centre (JSC) the different granting policies for HPC and Cloud are a major issue, i.e. HPC resources are usually allocated to scientists who passed a scientific evaluation (peer review) while Cloud and distributed (non-HPC) resources are offered via pay-per-use, pre-allocation for communities or on a voluntary base.

A major technical issue is the transfer of metadata between different "services" or interface to data storages, not only the metadata describing the data but e.g. access rights as well. An example is data sharing through a web service/interface where the metadata describing the data is often stored in a database that is part of the web service and thus not available on the filesystem (or via a POSIX interface), which are currently the standard data access methods on HPC systems. A better integration of AAI on the resource is needed instead of translating/mapping an authentication token to a user account.

UNICORE File Transfer Protocol (UFTP) is a solution to experiment. Unity is an AAI solution for federated infrastructures that integrates several identity providers and services with different authentication methods (SAML, OAuth, X.509...). It is the base of the EUDAT B2ACCESS[51] and is used in the Human Brain Project (HBP[52]) project to bridge the HBP-Collaboratory Cloud service (for data management, resource and process sharing) and the HPC infrastructure.

The HBP Collaboratory is a good example how Cloud and HPC resources could be accessed through the same interface.

KIT; Germany

The Karlsruhe Institute of Technology (KIT) hosts the Steinbuch Centre for Computing (SCC).

At KIT the users face real challenges in actively working with data. There are infrastructures and sites available, where such data can be stored to (e.g. EUDAT, HDF in Germany, KIT for HPC users at HLRS in the state of Baden-Württemberg) but:

- network connections are not powerful enough for this burst-like data transfer
- AAI: HPC centres are not connected to Federated IDM systems like eduGAIN, B2ACCESS, etc.

A lot has been done for LHC and WLCG in a production environment:

Analyse what is good and not so good (e.g. in terms of certificates) and then improve the services.

AAI and federated identity management is the most important key factor. KIT gives the example of the Baden-Württemberg state-wide operational identity federation.

The KIT points out the issue of long-term availability of services, infrastructures and resources and comments the fact that a project based funding is not sufficient. This comment is taken into account in the analysis.

Mésocentre Clermont Auvergne (MCA); France

MCA is a local centre serving users from all scientific disciplines and is a node of *France Grilles*, the French national grid infrastructure.

The challenges are connected to limited human resources (to install OpenStack solution cloud, to support usage).

Cross border interoperability of MCA computing resources has been successfully achieved thanks to the grid technology.

Data interoperability is emerging especially interoperability of data collected from multiple observatories.

[51] https://eudat.eu/services/b2access
[52] https://www.humanbrainproject.eu/)

For data interoperability, GAFA[53] technologies are important, for data mining approaches of unstructured data, tools like Elastic Search should be explored.

To foster interoperability, it's important to have a team of engineers, who know how to use different resources, share their expertise and build bridges between communities.

PSMN; France

The *Pole Scientifique de Modélisation Numérique* (PSMN) is the local computing centre of the *Ecole normale supérieur* (ENS) Lyon.

The main challenge is to authenticate and authorize external users. Access to national and European id federations is essential.

Unistra; France

The survey was completed by two infrastructures located in Strasbourg: the HPC centre of the University and the grid and cloud centre operated by the *Institut Pluridisciplinaire Hubert Curien* (IPHC). They shared the calls for attributing computing hours since 2014. Interoperability between these two infrastructures is limited: different software environments, different authentication, and no shared file systems. Common authentication and common front-ends for users are the most important challenges to answer user's needs.

Solutions like web portals to submit jobs and mechanisms to integrate HTC jobs into HPC workflows are of great interest for them as well as common file system tool to facilitate the distribution of the files between the two infrastructures.

They already collaborate with the University of Freiburg and with the KIT.

A.8. E-Infrastructures inputs

EGI e-infrastructure

EGI (egi.eu) is a federated e-infrastructure set up to provide advanced computing services for research and innovation. The EGI federated e-infrastructure is publicly funded and provides compute and storage resources to support research and innovation. The federation is governed by the participants represented in the EGI Council and coordinated by the EGI Foundation. The Council participants are organisations representing national e-infrastructures and two European Intergovernmental Research Organisations.

The challenges identified by EGI are:

Policy Challenges:

- The access policies applicable by the various providers of the EGI Federation, at national and local level, differ greatly:
- Local/National infrastructures are not always open to any researcher from any disciplines. In many cases these are primarily devoted to selected disciplines and/or projects, or can only be allocated to researchers affiliated to specific projects/institutions, preventing them to support researchers from other domains and organizations. In some cases infrastructures that are open to all disciplines, offer different service levels depending on the discipline.
- Where existing, the process to apply for access varies depending on the provider. This can be a problem in a federated environment, as it requires ad hoc activities depending on the services of interest to a research community. In some cases providers don't have a defined order management process with

[53] Google, Apple, Facebook and Amazon.

named individuals in charge of it. Some of these processes have a long lead time, which make them unsuitable for short term projects and collaborations. Some processes can be incompatible, making it impossible to combine services from different suppliers.

- We lack a suitable procurement framework that allows Europe-wide research collaborations and projects that want to acquire services from Research e-Infrastructures, to easily place service orders that span local, regional, national and organizational domains. On the supply side, Research e-Infrastructures lack the legal frameworks that allow them to respond collectively to tenders.

- Business models of services providers (e.g. free at point of use, pay for use, policy-based) are very different.

- We lack solutions for hosting privacy sensitive data in a secure manner in a distributed multi-supply environment.

Technical Challenges:

- Lack of an AAI infrastructure that allows the support of multiple protocols for user identity management, authentication and authorization. However, EGI-Engage and the AARC project and new services being introduced by e-Infrastructures, are quickly advancing the state of art and we expect progress.

- Different data and compute management protocols exposed by services. However, regardless of the heterogeneity of protocols and interfaces exposed by services, EGI has considerable experience in community-specific and/or general service access solutions that allow achieving interoperation across heterogeneous services. These have been very successful in achieving interoperation at international level, expanding to service providers not only across Europe but also from other regions. An area that still requires investigation and the definition of an interoperations blueprint is hybrid cloud federation involving commercial and publicly funded clouds.

Solutions proposed by EGI that the project could experiment with:

- Solutions such as Onedata and iRODS for federation of data across scientific domains and within one domain, to enable ease of discoverability, access and bringing data near to computing. A piloting activity involving different data providers and compute infrastructures would be beneficial.

- Brokers for instantiation and management of virtual machines (VMs) taking into account data locality.

- Container Management in a multi-provider environment (like a Kubernetes Federation) to enable a single platform for running container-based applications on the infrastructure.

- AAI: Researchers using a single digital identity through the EGI AAI CheckIn service to access resources federated by different data providers and compute infrastructures.

Solutions recommended by EGI:

- CheckIn: Identity Provider/Service Provider Proxy for “linking” of multiple service providers to multiple federated identity providers and aggregating user information from different community-managed attribute providers.

- APEL accounting of the usage of heterogeneous resources of different types (e.g. computing, data etc.).

- Workload managers e.g. DIRAC.

- AppDB VMOps: GUI dashboard with unique interface for IaaS providers of a cloud federation.

- AppDB Cloud MarketPlace and CloudKeeper: a browsable catalogue of Virtual Appliances (Virtual Machine Images with associated extra software) for communities to publish and promote their software and a tool for distributing securely and automatically to IaaS clouds to enable the portability of workloads

across providers.

- IaaS Provisioning tools (e.g. IM, Terraform or OCCOPUS) that unify and simplify the access to heterogeneous cloud frameworks using a single infrastructure description and a single workflow for managing resources at all providers. These make use of a variety of standards for consistent access and management of different clouds (e.g. OCCI, TOSCA).

- DNS as a Service for managing names of services across a federation and allowing easy named-based discovery for applications.

EUDAT e-infrastructure

EUDAT is a pan-European collaborative data infrastructure. Its consortium includes research communities, national data and HPC centres, technology providers, and funding agencies. EUDAT aims to build a sustainable cross-disciplinary and cross-national data infrastructure providing a set of shared services to access and preserve research data.

The interoperability challenges are at different levels, from technical to organisational, from customer towards service provider point of view, below a short list of items that come to mind:

- AAI interoperability. Each (e-)infrastructure, service provider, community has adapted different kinds of methods and technologies for IdM, authentication and authorisation, creating more or less different user domains. To allow access to a multitude of (e-) infrastructures and services, these user domains should be bridged across community, (e-)infrastructures and service domains, bridged across different AAI technologies and methods. This is not just a technical challenge, but to maintain a certain level of assurance (LoA) across the different IdM domains, also an organisational challenge.

- Interoperability on the level of security. When providing access across (e-) infrastructures a minimum agreed level of security should be applied and adhered to across the (e-)infrastructures. For this already good initiatives are developed, for example via the WISE community, and via the Security for Collaborating among infrastructures (SCI) framework, of which recently version 2 has been released.

- Standardisation and interoperability between API for access and data transfers. Within the (e-) infrastructure and services different types of API's are supported for access and data transfers. To enable easy access and data flowing across services and (e-) infrastructures a number of API's should be standardised. To make efficient data transfers possible, API and protocols for data transfers should support third party transfers. This does not mean a single API or single protocol. Multiple API can be adopted, but a limited number should be adopted across the (e-) infrastructure and services. For sustainability, these API should be preferably based on HTTP and on open and/or well adopted standards.

- Interoperability in service and resource provisioning. Service and resource provisioning within a distributed infrastructure is in itself already challenging. To plan and to organise service and resource provisioning across e-infrastructures becomes even more challenging. Not just for the service providers but also for the community, that is requesting the service and resources. To plan and to organise work within a distributed environment provides additional managerial and communication overhead. Having tools like a central service catalogue in which available services can be promoted, having service requests procedures which support distributed service requests and resource provisioning, tools for managing projects with stakeholders across communities, (e-) infrastructures and service providers, having an operational organisation which collaborates across (e-) infrastructures and service providers.

Solutions proposed by EUDAT that the project could experiment with:

- B2DROP – secure and trusted data exchange service for researchers and scientists to keep their research data synchronised and up-to-date and to share with other researchers

- B2SHARE –a user-friendly, reliable and trustworthy data repository service for researchers, scientific communities and citizen scientists to store and share small-scale research data from diverse context

- B2SAFE – is a robust, safe and highly available service which allows communities and departmental repositories to implement data management policies on their research data across multiple administrative domains in a trustworthy manner
- B2FIND – is a simple, user friendly metadata catalogue of research collections stored in EUDAT data centers, community and other data repositories.
- B2HANDLE – is a robust service for managing persistent identifiers, it enables people to register data, making it possible to refer to or cite research data providing long term references.
- B2NOTE – is a simple, user friendly service to manage annotations on data sets and data objects
- B2ACCESS – is an easy-to-use and secure identify management system to provide access to EUDAT services, supporting multiple IdM protocols.
- SPMT (Service Portfolio Management Tool) – is an easy-to-use service to manage services descriptions through the life cycle of a service, from initial concepts of a service until it's being deprecated.
- DPMT (Data Project Management Tool) – is an easy-to-use service to manage projects in a distributed environment crossing borders, service providers and administrative domains.

Solutions recommended by EUDAT for the project:

As described at point 1, providing access to services and data across (e-) infrastructures, services and different administrative domains is one of the main interoperability issues.

To solve these issues within EUDAT, we developed the concept of credential conversion (e.g. token translation) to support multiple methods for user authentication and service integration. For this EUDAT developed the B2ACCESS service to enable access to the EUDAT Collaborative Data Infrastructure (CDI). The B2ACCESS service has been in production since October 2015. To enable access between EUDAT, EGI and PRACE infrastructure and services EUDAT started integration work between EGI and PRACE AAI systems to enable access across these e-infrastructures. EUDAT has also started to pilot the use of the B2ACCESS service with the EPOS community to provide an EPOS unified IdM domain and bridge access to services between EPOS and CDI. To foster this collaborative work, it would be of great interest to extend this to other communities, infrastructures and services.

PRACE e-infrastructure

PRACE has been contacted but did not answer.

ANNEX B. 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 delivers sustainable, integrated and secure distributed computing services to the national research communities and their international collaborators. The federation is coordinated by a National Coordinating Body providing a single point of contact at the national level and has official membership in the EGI Council through an NGI legal representative.
Virtual Organisation (VO)	A group of people (e.g. scientists, researchers) with common interests and requirements, who need to work collaboratively and/or share resources (e.g. data, software, expertise, CPU, storage space) regardless of geographical location. They join a VO in order to access resources to meet these needs, after agreeing to a set of rules and Policies that govern their access and security rights (to users, resources and data).