A European Open Science Cloud

Abstract
This document outlines the position of EIROforum on a European Open Science Cloud. It explores the essential characteristics of a European Open Science Cloud if it is to address the big data needs of the latest generation of Research Infrastructures. The high-level architecture and key services as well as the role of standards is described. A governance and financial model together with the roles of the stakeholders, including commercial service providers and downstream business sectors, that will ensure a European Open Science Cloud can innovate, grow and be sustained beyond the current project cycles is described.

About the EIROforum
EIROforum partners are intergovernmental research organisations – CERN, ESA, EMBL, ESO, EuroFusion, European XFEL, ILL and ESRF – covering disciplines ranging from particle physics, space science and biology to fusion research, astronomy, and neutron and photon sciences. The partner organisations have a truly European governance, funding and remit, and in many cases share a global engagement. They are world leaders in basic research, as well as in managing and operating large research infrastructures and facilities. The EIROforum collaboration is helping European science reach its full potential through exploiting its unparalleled resources, facilities and expertise. By combining international facilities and human resources, EIROforum exceeds the research potential of the individual organisations, achieving world- class scientific and technological excellence in interdisciplinary fields. EIROforum works closely with industry to foster innovation and to stimulate the transfer of technology.

Prepared by CERN IT department on behalf of the EIROforum IT Working Group.

Executive Summary

EIROforum members and other Research Infrastructure operators face unsustainable demand for computing and networking services to deliver the promise of Open Science. They need more cost-effective approaches to collecting, processing, distributing and re-using the rapidly growing amounts of data being produced by their instruments.
This will require innovative ways of providing an integrated IT infrastructure and operations expertise needed to run applications.

Currently in-house resources, public e-infrastructure and commercial cloud services are not integrated to provide a seamless environment for data–intensive science. Existing services do not cover the full lifecycle of research from proposal submissions requesting access to Research Infrastructures, through to data acquisition, sharing and publication. Researchers are by-passing their in-house IT departments and publicly funded e-Infrastructures to make use of commercial cloud services that offer innovative, easy-to-use solutions and fill the service gaps. This *shadow IT* innovation represents an opportunity to introduce change but must be undertaken with full knowledge of the policy aspects including data protection, intellectual property rights and applicable legislation.
A European Open Science Cloud has the potential to provide the means to link such services together and increase scientific output.

The Helix Nebula initiative (HNI) has brought together more than 40 service providers, research organisations, data providers and publicly funded e-infrastructures. It has developed a hybrid cloud model with procurement and governance components suitable for the dynamic cloud market. A Pre-Commercial Procurement (PCP) is being negotiated to build a new form of *IT as a Service (IaaS)* platform using open source solutions in a federated Science Cloud.
Procuring cloud services from providers on a pay-per-usage model on the operations budget rather than the capital budget offers both flexibility and scalability.

E-infrastructure costs will become an integral part of the cost of doing science and, consequently, must be cost-justified in terms of benefits and impact. Moving to the cloud can enable more flexible pricing models such as per core/hour or per request/transaction or migration to Open Source Software (OSS) to control growing software licensing costs.
Most publicly funded research organisations lack detailed cost models inhibiting financial comparisons between traditional and cloud-based solutions.

RIs need to understand the benefits as well as the full costs of 'big data' services and be able to manage their own procurements in a competitive marketplace, migrate use cases and existing infrastructures to the cloud paradigm, and adopt an appropriate collaborative governance model. Services will be provisioned from commercial suppliers when they are not available in-house or can be delivered externally on better terms (i.e. at shorter notice, lower cost or better performance etc.). Publicly funded data centres will continue to guarantee long-term data preservation and service supplier independence. A market assessment of the public research sector and downstream business sectors that could build on the data produced by Research Infrastructures is needed to build confidence in the business model and justify investments in a European Open Science Cloud by the supply-side.

A significant difference compared to the current model is that funding agencies and research organisations will no longer provision services *exclusively* from their own in-house resources.

Stakeholders in the public and commercial sectors must not only invest in the building blocks for the development of e-Infrastructure listed in Table 1, but also in end-user facing services and in training the next generation of IT-savvy researchers. This will leverage the investments already made in the publicly funded e-infrastructures and commercial cloud services.
All stakeholder groups need to work together to ensure wide adoption of competitive, secure, reliable and integrated computing services.

Many research organisations that operate research infrastructures do not have the mandate to provide IT services to their users for the management and processing of their experimental data and will require assistance to bridge the gap from data to knowledge acquisition. The guiding principle is that funding from stakeholders like the EC and national funding agencies will be focused on innovation of services and uptake by new user communities and business actors while the operational costs will be borne by the operating organisations and the user communities**.**
The funding model for a European Open Science Cloud must be designed so that the services can be sustained by their operating organisations.

The EC's INFRASTRUCTURES 2016-2017 work programme foresees new e-Infrastructure for data and distributed computing and a pilot for the federation, networking and coordination of pan-European research infrastructures and clouds in general. Looking further ahead, the EC has taken steps to ensure funding for GÉANT over the full duration of H2020 by introducing 'Framework Partnership Agreements' (FPA). The FPA model represents a more long-term engagement that could encourage the integration of e-infrastructures co-funded via EC projects into the Research Infrastructures' computing models. The application of the FPA approach to a European Open Science Cloud could establish the basis for the European Research Area's digital commons and lead towards Science 2.0.
A European Open Science Cloud represents a strategic vision that can be a vector for introducing change in the service provisioning and computing models for the publicly funded research sector in the medium to long term.
A European Open Science Cloud has the potential to greatly improve the provisioning of IT services for Research Infrastructures to address their big data needs. It can encompass all the phases of the research lifecycle and offer a platform of joint innovation for the public and private sectors. It will significantly change the way IT services are procured, organised and funded. The key challenges are integrating frequently changing technologies, managing the complexity and identifying the optimal organisational and financial models. Researchers must be convinced that they will not lose control of their precious data. It is an ambitious undertaking requiring the active engagement of many stakeholders and careful planning of the technical, financial, legal and governance aspects. For it to succeed it must become a priority for all the actors involved with monitoring by the funding agencies and regular assessment by the user communities.
This position paper is a rallying call for adoption of such a strategic approach – within the EC and other funding bodies to work the operators of Research Infrastructures.

Table 1 – major stakeholder groups

National funding agencies • Policy makers • Third sector • Granting bodies
European Commission • DG CONNECT • DG RTD
Research communities • Thought leaders • Peers • Scholarly publishers
Research Infrastructures • Policy-makers • Operational staff • Data users
Public e-infrastructures • Service providers • Host organisations • Technology providers
Commercial cloud service providers
Independent Software Vendors
Open Source developer communities
Standards bodies

Table 2 - relevant EC co-funded projects

AARC	https://aarc-project.eu
Cloud for Europe	http://www.cloudforeurope.eu/
EGI	https://wiki.egi.eu/wiki/Main_Page/
EUDAT	http://www.eudat.eu
GÉANT	http://www.geant.net/
Helix Nebula	http://www.helix-nebula.eu
Indigo Datacloud	https://www.indigo-datacloud.eu/
OpenAIRE	https://www,openaire.eu
PICSE	http://www.picse.eu/
PRACE	http://www.prace-ri.eu/
SLALOM	http://www.slalom-project.eu/

Contents

Executive Summary i
Contents iii
Overview 1
Sustainability in a world experiencing the data tsunami 1
Mind the Gap 1
Need for new way of procuring ICT services 1
Open Science requires an integrated approach 2
Hybrid cloud-based solutions 2
Background 2
Progress to date 2
Pre-Commercial Procurement 3
Challenges facing Research Infrastructure operators 3
Benefits of a hybrid approach for scalability 4
Commercial considerations 4
Supply-side 4
Demand-side 5
Procurement 6
The role of standards 7
Implementation: Scope 8
Federated Approach 8
Support services 9
Implementation: Connectivity 10
Transport of huge amounts of data and the lack of high-performance links 10
Identity management 10
Implementation: Open Data 11
Providing broader access to community-specific solutions 11
Data preservation 13
Reproducibility of research 14
Governance 14
Investment 17
Proprietary solutions are not solutions 17
Public investment 18
Investment in skills 18
Long-term strategic investment 19
Conclusions 20

Overview

- Sustainability in a world experiencing the data tsunami
- Mind the Service Gap
- Need for a new way of procuring ICT services
- Open Science requires an integrated approach
- Hybrid cloud-based solutions

Sustainability in a world experiencing the data tsunami

Traditional ways of meeting the growing demand for computing and networking services capable of addressing the 'Data Tsunami[1]' are seen to be unsustainable by funding agencies as well as the infrastructure operators such as GÉANT and EGI. The cost of collecting, processing, distributing and re-using the rapidly growing amounts of data produced by their instruments is a major concern for Research Infrastructure operators including the EIROforum members. A collaborative shift towards more cost-effective ways of generating and using scientific data and a greater role for the users of that data is required in order to develop a sustainable future for the evolution of Open Science.

Mind the Service Gap

Over the last decade, driven with sustained funding from the EC, the e-Infrastructure landscape across Europe has grown from regional prototypes to a set of pan-European production resources including EGI, GEANT, PRACE etc. This has resulted in a number of services within the context of each project but there is no common, overarching goal and so user communities must invest significant effort to bring these services together.

Currently in-house resources, public e-infrastructure and commercial cloud services are not integrated to provide a seamless environment for data–intensive science. Existing services do not cover the full lifecycle of research from proposal submissions requesting access to Research Infrastructures, through to data acquisition, sharing and publication. Researchers are by-passing their in-house IT departments and publicly funded e-Infrastructures to make use of commercial cloud services that offer innovative, easy-to-use solutions to fill-in the service gaps. This *shadow IT* innovation represents an opportunity to introduce change but must be undertaken with full knowledge of the policy aspects including data protection, intellectual property rights and applicable legislation.

A European Open Science Cloud has the potential to provide the means to link such services together and increase scientific output.

Need for a new way of procuring ICT services

Public research organisations have to find alternatives to the traditional route of purchasing and operating in-house IT equipment which requires capital investment on the physical infrastructure (servers, network, storage) needed to run an application as well as operations expertise. Cloud computing has the potential to reduce IT expenditure while at the same time improving the scope for innovative and flexible high-quality services. Procuring external cloud services from providers on a pay-per-usage model implies that infrastructure is no longer 'institutionalised' and the cost of cloud services can be found on the operations budget rather

[1] http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf

than the capital budget. There is 'elasticity' in cloud-based services and cloud-based infrastructure is inherently scalable.

Open Science requires an integrated approach

'Open Science' is still in its infancy - driven predominantly by the availability of enabling technologies and the opportunities for new ways of working rather than by demand from society at large, according to a recent consultation[2]. Lack of integration of the existing infrastructures (and, by inference, access to the data they carry) was seen to be a barrier to adoption of those technologies and working practices by 86% of the individual scientists who responded to the survey.

Hybrid cloud-based solutions

The Cloud for Europe project[3] has shown that uptake of cloud services by European Public Administrations is still very fragmented in terms of demand and procurement of IT services. The Helix Nebula[4] initiative, however, has demonstrated the potential of a hybrid model in which service providers, research organisations, data providers and publicly funded e-infrastructures are brought together. Building on that potential will allow us to support and transform publicly funded research into data driven knowledge which is of value to the wider research community and downstream industries.

Helix Nebula has already brought innovation to the relationship between suppliers and users and introduced a wider range of new players to the marketplace. This provides a platform onto which a European Open Science Cloud initiative[5] will add a further much-needed dose of innovation and accountability in the way technology is procured and deployed.

The goal of this position paper is to allow the EIROforum members to articulate their own expectations of the initiative by helping them to understand the new European Open Science Cloud and the way that it addresses the needs of the Infrastructure operators and users.

Background

- Progress to date
- Pre-Commercial Procurement
- Challenges facing Research Infrastructure operators
- Benefits of a hybrid approach for scalability

Progress to date

Milestones on the journey initiated by Helix Nebula have included:

- Creation of a vibrant public-private partnership of more than 40 organisations and companies.
- Development and continued monitoring of a strategic plan for cloud computing in the public research sector[6].

[2] Public consultation on "Science 2.0: Science in transition"

[3] http://www.cloudforeurope.eu/downloads

[4] http://www.helix-nebula.eu/helix-nebula-vision

[5] http://dx.doi.org/10.5281/zenodo.16001

[6] http://www.helix-nebula.eu/publications/deliverables/d92-strategic-plan-scientific-cloud-computing-infrastructure-europe-three

- Identification and evaluation (through testing and production use) of a hybrid cloud model meeting the needs of publicly funded research by linking commercial cloud services with e-infrastructures[7].
- Validation of inclusive procurement models that address many examples of key procurement barriers with a wizard tool allowing public organisations to analyse their procurement processes and determine a suitable procurement model for cloud services when their existing models are not a good match for the dynamic cloud market.[8]
- An inclusive, transparent and user driven governance structure capable of delivering on the initiative's objectives.

Hybrid clouds combine private infrastructure and operations with shared infrastructure and operations. A typical hybrid cloud use case would be the relocation of the presentation tier (user interface) and logic tier where the application knowledge is encapsulated to an off-site cloud and have them communicate with the database stored and managed within the organisation's own IT infrastructure. In order for the demand-side users to be encouraged to purchase cloud computing services, the services offered must be economically advantageous compared to other means of procuring IT services.

Pre-Commercial Procurement

Promotion of joint procurement has led to the creation of an expanding procurement network of publicly funded research organisations and establishment of a new Pre-Commercial Procurement (PCP), the Helix Nebula Science Cloud (HNSciCloud).

HNSciCloud is designed to pull together publicly-funded e-Infrastructures using open source solutions, to build a hybrid Infrastructure as a Service (IaaS) platform. It will host a competitive marketplace of European cloud players where they can develop their own services for a wider range of users beyond research and science including downstream business sectors that can make use of publicly funded research data.

The goal is to establish a sustainable European Open Science Cloud serving Europe's Research Infrastructures, communities and related business sectors and surpassing the capacity currently available via existing public e-infrastructures and the in-house facilities of research organisations. It will be based on the migration of *Infrastructure as a Service* into the more general *IT as a Service* consisting of software tools and applications and the platforms on which they run. Services will be provisioned from commercial suppliers when they are not available in-house or can be delivered externally on better terms (i.e. at shorter notice, lower cost or better performance etc.). Publicly funded data centres will continue to guarantee long-term data preservation and commercial service supplier independence.

Challenges facing Research Infrastructure operators

HNSciCloud will enable the federation, networking and coordination of existing Research Infrastructures and scientific clouds in preparation for what the 2016 INFRASTRUCTURES Work Programme calls the "European Open Science Cloud for Research". It brings Europe's technical development, policy and procurement activities together to remove fragmentation and support Research Infrastructure operators facing three key challenges:

- Empowering them to understand the benefits as well as the full costs of 'big data' services and manage their own procurements in a competitive marketplace
- Migrating use cases and existing infrastructures to the cloud paradigm

[7] http://www.helix-nebula.eu/publications/deliverables/d62-roadmap-the-integration-and-interoperation-of-commercial-cloud-e

[8] http://www.picse.eu/publications/deliverables/d-21-research-procurement-model

- Selecting an appropriate collaborative governance model that avoids the barriers that currently inhibit a 'joined up' way of working by involving the research user community, the research infrastructures and the research funding bodies.

We expect the scale and range of services being provisioned from commercial suppliers to gradually increase over time as the cloud market matures and Open Science becomes embedded in the research lifecycle. A significant difference compared to the current model is that funding agencies and research organisations will no longer provision services *exclusively* from their own in-house resources.

In an answer to a written question in the European Parliament about the current position regarding procurement of the European Science Cloud, Commissioner Oettinger stated that: *"The Commission has supported path finding studies on the use of hybrid models, bringing together public research organisations and e-infrastructures with commercial suppliers to build a common platform offering a range of services to research communities. This can be achieved by building on cloud technologies easily accessible to users and by promoting procurement of cloud services to encourage innovation on the supply side."* The role of Helix Nebula and the HNSciCloud in shaping that position is clear.

Benefits of a hybrid approach for scalability

If there is significant variation in demand, there may be an opportunity to reduce operating expenditure by matching the supply of resources to the level of demand. By employing a hybrid cloud model, an organisation can quickly and economically add resources as needed by bursting out of its private IT infrastructure to a commercial cloud processing and storage capacity. A cloud-bursting scenario can provide the benefits of cost savings, maximum utilisation of on-premises resources and rapid innovation, but also has its own set of challenges in ensuring the performance, agility, security and management aspects of a hybrid cloud infrastructure.

By intermixing private and public cloud infrastructures, organisations are able to use the hybrid model to leverage in-house and off-site resources. The hybrid model allows organisations to rely on the cost-effective commercial cloud for non-sensitive operations and on the private cloud for critical, particularly sensitive operations providing enhanced agility to move applications easily between the in-house and off-site resources taking into account aspects of policy, cost, security and availability.

Commercial considerations

- Supply-side
- Demand-side
- Procurement
- The role of standards

Supply-side

One important consideration is that this approach must generate benefit for the providers who have the responsibility of ensuring that they have the physical infrastructure to meet their users' demand and that their performance meets agreed service quality levels. Without an accurate view of future demand, planning for variable costs such as staff, replacement servers or coolers, and electricity supplies can all be very difficult, and optimising the distribution of virtual machines presents a major challenge. The more unpredictable and spikey the workloads, the greater the economic benefit of sharing the same services across diverse research communities in the public and private sectors. Analysis of the procurements made via Helix Nebula, suggests there is

insufficient installed capacity currently available in the European market to satisfy the exceptional demand that will be generated by the latest generation of research infrastructures. Significant investments by the supply-side, based on accurate future predictions of usage will be necessary. Consequently it is important that a market assessment of the public research sector and downstream business sectors that could build on the data produced by Research Infrastructures is performed (similar that performed by the UK government for public sector information[9]) in order to build confidence in the business model and justify investments in a European Open Science Cloud by the supply-side.

There are also licensing implications when transitioning from a scale-up architecture to a scale-out architecture: some applications are licensed per-instance or per-CPU, often over an annual term. In this instance, there can be significant cost implications of adding new instances to a pool of resources. In time, application vendors will follow infrastructure service providers in moving to more flexible pricing models such as per core/hour or per request/transaction. The alternative is to use Open Source Software (OSS) where the license cost issue is non-existent.

Demand-side

As identified in the GEANT Expert Group report[10], the user communities will increasingly be called upon to pay for the services they receive if e-infrastructures on which users can depend are to continue to survive. E-infrastructure costs will be an integral part of the cost of doing science and, consequently, e-infrastructure investments must make a substantial and sustainable impact in order to be justified in terms of costs and benefits.

A study of the cost effectiveness of European dedicated HTC and HPC computing e-infrastructures for research compared to equivalent commercial leased or on-demand offerings was performed by the eFISCAL project[11] in 2011. The conclusion was that the ratio of CAPEX (CAPital EXpenditure) to OPEX (OPerational EXpenditure) for e-infrastructures was 30%-70% and manpower accounted for approximately 50% of the costs (CAPEX+OPEX). A Total Cost of Ownership (TCO) study[12] was performed by SAP Research on specific CERN in-house services within the context of the Helix Nebula FP7 project.

Both of these studies indicated that most publicly funded research organisations lack detailed cost models for individual services. Financial comparisons between traditional and cloud-based solutions would need a set of guidelines for such organisations proposing which category of costs should be included or excluded. It is important to recognise that shifting the procurement of IT services to a pay-per-usage model will normally have a limited impact on TCO since the bulk of expenditure over the lifetime of an application is not related to the purchase of physical infrastructure. It is also the case that not all publicly funded research centres are in a position to make accurate estimations of the TCO of in-house IT services since some contributing costs are borne by different departments.

The adoption of cloud computing services by public research organisations requires additional justification in terms of the benefits of the new ways of working that cloud-based services enable. Research organisations justify their investments by the impact made in IT services on the end-user communities in terms of scientific output. To gauge this impact it is necessary to understand

[9] Market Assessment of Public Sector Information, May 2013, https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/198905/bis-13-743-market-assessment-of-public-sector-information.pdf

[10] http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/geg-report.pdf

[11] http://www.efiscal.eu/files/deliverables/D2%203%20Executive%20Summary%20-%20Computing%20e-Infrastructure%20cost%20calculations%20and-business%20_models_vam1-final.pdf

[12] http://www.helix-nebula.eu/publications/deliverables/d73-costing-exercise-comparing-in-house-vs-cloud-based-operation-the-cern

the needs and activities of the end-users. Factors such as pattern of demand and transitional costs need to be included in any financial analysis of a potential cloud computing solution.
A European Open Science Cloud will need to perform IT capacity planning for all engaged research communities on a regular basis. As an example, the WLCG project has a Computing Resources Scrutiny Group[13] which reviews the computing resources for the LHC experiments on an annual basis.

Procurement

The EC-funded 'Procurement Innovation for Cloud Services in Europe' (PICSE[14]) project is identifying barriers to procurement of cloud services by public research organisations and is developing a new procurement model to overcome them. With the advent of cloud computing, the delivery of ICT services is going through a fundamental change. However, while cloud technology service options continue to evolve, procurement processes and policies of public research organisations have remained firmly rooted in historical practices that are no longer effective. In order for public research organisations of all sizes to take advantage of the best the cloud market has to offer, a more flexible and agile procurement model must be identified and implemented.
PICSE has contacted a number of public sector organisations and initiatives (including CERN[15], Cloud for Europe[16], DG DIGIT[17], ECMWF[18], EMBL[19], ESA[20], ESRF[21], Europeana[22], GRNET[23] and Umeå University[24]) to discuss their current practices.
The main challenges identified that need to be addressed in the procurement of cloud services can be summarised as follows:

- As with all purchases of new technologies, procuring innovative services requires new skills and competences.
- Organisational/cultural barriers to cloud adoption are very important, especially when the organisation is purchasing cloud for the first time.
- Financial issues may arise due to the new way to evaluate costs in moving to the cloud.
- Legal/organisational issues may be encountered due to the cloud service deployments particularities e.g. applicable law, data location restrictions, data protection, etc.
- Security, including network security, data protection, privacy, data and service portability, interoperability are all elements to be considered when identifying the cloud solutions to purchase.
- Vendor lock-in (dependency on the vendor) and vendor viability are aspects that have to be considered.

[13] http://wlcg.web.cern.ch/collaboration/management/computing-resources-scrutiny-group
[14] http://www.picse.eu/
[15] http://home.web.cern.ch/
[16] http://www.cloudforeurope.eu
[17] http://ec.europa.eu/dgs/informatics/identity_en.htm
[18] http://www.ecmwf.int/
[19] http://www.embl.de/
[20] http://www.esa.int/ESA
[21] http://www.esrf.eu/
[22] http://www.europeana.eu/
[23] https://www.grnet.gr/
[24] http://www.umu.se/

- Dynamic and changing cloud services must be monitored to ensure proper performance and benefit realisation. Service level agreements (SLAs) must be drafted and managed diligently, an area where the EU SLALOM project has begun working[25].
- Vendor contract negotiation is complicated and critical. There are no standard contracts for cloud. The SLALOM project is finalising a cloud service contract template with equitable terms and conditions for suppliers and customers.
- Contract termination conditions need to be carefully evaluated. Porting data to another cloud or non-cloud solution may involve high costs. Cloud escrow is also missing.

These challenges have an impact on all the steps of the procurement process. There is a clear impact on skills and knowledge required. IT managers within public research organisations should have a clear understanding of the new technology being purchased.
Functionally similar to financial market brokers, cloud brokers match provider supply with consumer demand. This model benefits all parties: experiencing more predictable demand, cloud providers can better optimize their workflow to minimize costs; cloud users access cheaper rates offered by brokers; and cloud brokers generate profit from charging fees. Including such brokerage models in a European Open Science Cloud could reduce the risks that arise from market instability. The adoption of a hybrid cloud model will also help to reduce the impact of market instabilities on a European Open Science Cloud.

The role of standards

Standards improve transparency and comparability for service users. They open up new markets for suppliers and offer equal access conditions, particularly for small and medium-sized companies. Standards also improve the quality, security and sustainability of products and services and adoption of suitably defined standards exposes the supplier's unique selling propositions. Open standards can be adopted to provide interoperability between parts of the infrastructure, portability from one cloud service provider to another and trust in the integrity (provenance, reliability, etc.) of the infrastructure that has been built.
Emerging cloud standards for application orchestration provide template-driven descriptions of applications as a transparent way of abstracting the relationships between cloud applications and services and the underlying platform or infrastructure. One example of this is TOSCA (Topology and Orchestration Specification for Cloud Applications) from OASIS[26], selected by the Horizon 2020 EC co-funded Indigo Datacloud[27] project. This gives suppliers and users interoperable descriptions of cloud-hosted services and applications, including their components, relationships, dependencies, requirements, and capabilities. TOSCA has the potential to expand customer choice, improve reliability, and reduce cost and time-to-value, facilitating the agile, continuous delivery of applications (DevOps) across their entire lifecycle.
Portability is another significant property since prospective users want to avoid vendor lock-in when they choose to use cloud services. Users need to know that they can move their data and applications between multiple cloud service providers at low cost and with minimal disruption. Portability through the appropriate standardisation of APIs, data models, data formats and vocabularies will help automate business processes surrounding cloud computing procurement, enable straightforward technical integration between the client and provider, and allow for flexible and dynamic application deployments across multiple clouds.

[25] http://www.slalom-project.eu/
[26] https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=tosca
[27] https://www.indigo-datacloud.eu/

Trust and confidence in cloud computing services relies on work such as ENISA's CIIP[28] (Critical Information Infrastructure Protection) initiative which defines appropriate strategies, policies and specific measures for protecting information on the cloud. The underlying cause of many of the risks and challenges associated with cloud computing is that the user passes over responsibility for data and for applications to the cloud service provider and the provider has a multi-tenant environment in which resources are shared. In addition to the many economic and technological advantages that cloud computing offers to research communities, there are also significant security benefits in migrating applications and usage to the cloud, as noted by ENISA. The shared resources available in clouds also potentially include rare expertise, shared best practices and advanced security technologies, beyond the means or abilities of the vast majority of SMEs, many larger companies and even many government bodies, to provide for their in-house systems.
A truly interoperable cloud will encourage adoption by users, safe in the knowledge that they can change providers, or use multiple providers, without significant technical challenges or effort. This will expand the size of markets in which cloud providers operate.

Implementation: Scope

- Federated approach
- Support services

Federated Approach

A European Open Science Cloud should offer an initial portfolio of services corresponding to the list of e-Infrastructure services documented by eIRG in its blue paper of 2010[29] with the technical characteristics identified by the High Level Expert Group on Scientific Data in their "Riding the Wave" report from the same year[30].

Implementations for the majority of the foreseen services already exist at varying levels of maturity. The key challenges are integrating frequently changing technologies, managing the complexity and identifying the optimal organisational and financial models. Researchers must be convinced that they will not lose control of their precious data. The data centres operated by public research organisations can provide such guarantees. They can rapidly expand the available capacity by making use of commercial service providers offering commodity compute and data services as part of the hybrid cloud model. By keeping a "safe copy" of the research data, the public research organisations can also insulate the researcher communities from changes in service provider and technology.

A European Open Science Cloud should take a bottom-up approach to implementation, starting with IaaS. Integration should start with a common catalogue of services and a federated identity management system offering a single sign-on facility to access services across all suppliers. Starting bottom-up is essential to get the core technical, financial, and policy principles right. IaaS can be introduced without impacting higher-level user-facing services that will require a significant software investment. It also represents a strategy with lower risk because the IaaS market is more mature than the PaaS and SaaS markets.

[28] https://www.enisa.europa.eu/activities/Resilience-and-CIIP
[29] http://e-irg.eu/documents/10920/238805/e-irg_blue_paper_2010
[30] http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf

The services of a European Open Science Cloud will need to be integrated with a range of resources currently operated by public organisations to form a hybrid cloud solution.
Realisation of the benefits of a hybrid cloud is inhibited by many barriers related to procurement, trustworthiness, technical standards and legal terms of reference, risk of vendor lock-in and so on. The overall challenge is to overcome these barriers in order to boost productivity by stimulating all stakeholder groups to work together to ensure wide adoption of competitive, secure, reliable and integrated computing services.

In order for a European Open Science Cloud to be deployed rapidly, it is essential to build on the existing infrastructures. This requires an agreed overarching architecture and the commitment of the service operators to make a European Open Science Cloud a priority. There must also be agreement by all the stakeholders on the governance structure and financial model to ensure a Open Science Cloud can grow, innovate and be sustained.

The EGI Federated Cloud[31] is an example of an inter-disciplinary approach to infrastructure implementation allowing data sharing and collaboration between research communities.
It is a grid of academic private clouds and virtualised resources, built around open standards and focusing on the requirements of the scientific community. Technical consistency in the service delivery between participating suppliers is ensured by use of recommended publicly defined interface specifications such as OCCI[32], CDMI[33] and OVF[34].

The experience gathered by EGI in managing its federated infrastructure[35] will be directly relevant and provide insight into making a larger portfolio of capacity style HPC services for data centric applications accessible to its existing user-base. Working with commercial cloud service providers will inject the innovation potential created by the uptake of cloud computing in research and business sectors.
The complementary expertise developed by PRACE and related projects in efficient parallel programming paradigms and optimising software for a range of architectures is also directly relevant to a European Open Science Cloud and application/service developers. The HPC capability services offered by the PRACE centres should be integrated to form part of the overall ecosystem. This will require the PRACE HPC centres to participate in the federated identity management scheme and data sharing services described below.

Support services

Support services will also be required to ensure the operational staff in the public research organisations can resolve end-user support issues as quickly and as efficiently as possible. Similarly, security response services will be necessary to handle incidents that may affect the platform. The publicly operated infrastructures that are part of the hybrid cloud already have user-support and Computer Security Incident Response teams (CSIRTs) in place but they do not fully interoperate and all cloud services supported by a European Open Science Cloud, whether operated by commercial service providers or public organisations, will need to be integrated into these structures.

31 https://www.egi.eu/infrastructure/cloud/
32 http://occi-wg.org/about/specification/
33 http://www.snia.org/cdmi
34 http://www.dmtf.org/standards/ovf
35 https://wiki.egi.eu/wiki/Fedcloud-tf:UserCommunities

Implementation: Connectivity

- Transport of huge amounts of data
- Identity management

Transport of huge amounts of data and the lack of high-performance links

In order for a European Open Science Cloud to operate effectively, it is necessary to assure there is sufficient network capacity to permit data ingress from the Research Infrastructures.
GÉANT[36] is the high bandwidth pan-European research and education backbone that interconnects National Research and Education Networks (NRENs) across Europe and provides worldwide connectivity through links with other regional networks. The GÉANT network is the primary means of connecting the research organisations and universities to the commercial providers. The Helix Nebula initiative has already demonstrated that it is possible to make practical use of the data centres of commercial cloud service providers over the GÉANT network. GÉANT Open[37] is a service allowing NRENs and approved commercial organisations to exchange connectivity for the public Research and Education sector with NOC support, SLA monitoring, a defined policy[38] and cost model[39]. Commercial cloud service providers are expected to add the cost of connection and usage of GÉANT Open to the price of the cloud services delivered to the research community. Commercial cloud service providers will also want to offer the same types of cloud services to customers from business sectors and will have to integrate alternative network providers which will allow the stakeholders to compare the efficiency and cost effectiveness of all the network services provided by the different suppliers.
The opening up of a European Open Science Cloud to users beyond the publicly funded research sector is essential if it is to attract investment from the private sector and support a vibrant innovation cycle. Looking further into the future, a European Open Science Cloud could be more closely linked to the data acquisition and real-time requirements of Research Infrastructures. For example, European XFEL, ILL and ESRF together with Eurofusion sites all require online or rapid feedback in order to prepare for the next experimental run. This implies important increases in network capacity. Similar real-time needs will also be important for the applications of new detector techniques addressed by the ATTRACT consortium[40].

Identity management

eduGAIN[41] is an international inter-federation service interconnecting research and education identity federations. It enables the secure exchange of information related to identity, authentication and authorisation between participating federations. eduGAIN provides an infrastructure for establishing trusted communications between Identity Providers (IdPs) and Service Providers (SPs) in different participating federations. End-users authenticate at IdPs and obtain access to services delivered by SPs.
Federated identity management is also gaining traction in business sectors as shown by the rising popularity of Universal 2nd Factor (U2F) as an authentication standard created by the FIDO (Fast

[36] http://www.geant.net/
[37] http://www.geant.net/Services/ConnectivityServices/Documents/GEANT%20Open%20Service%20Brief.pdf
[38] http://www.geant.net/Services/ConnectivityServices/Documents/GN3PLUS13-1439-12_geant_open_exchange_production_policy_v4_3.pdf
[39] http://www.geant.net/Services/ConnectivityServices/Documents/GEANT%20Open%20Service%20Description.p
[40] http://www.attract-eu.org/
[41] http://services.geant.net/edugain/Pages/Home.aspx

IDentity Online) Alliance[42] an industry group established to standardize authentication technology and devices that can simplify and strengthen two-factor authentication for businesses and consumers. So it will be essential for eduGAIN to ensure it can engage with commercial IdPs and SPs to avoid isolating the research and education community. The recently started AARC (Authentication and Authorisation for Research and Collaboration)[43] H2020 project intends to further develop eduGAIN and it is essential that a primary goal of this project should be to ensure eduGAIN can support a European Open Science Cloud in production usage.

Implementation: Open Data

- Broader access to community-specific solutions
- Data preservation
- Reproducibility of research

Providing broader access to community-specific solutions

Providing access to third-party open data requires appropriate management structures for data as well as the connectivity allowing interchange of the data itself.
The value chain for information can be considered in three layers – data providers, value-added providers and downstream users. The Global Earth Observation System of Systems (GEOSS[44]) is an example of a common infrastructure provided by a community of data providers. The 'GEOSS Portal' is a single Internet access point for users seeking data, imagery and analytical software packages relevant to all parts of the globe. GEOSS does not offer to host datasets or guarantee that they are always available but simply makes them accessible from their original sites.
GEO has a working group which has recently defined three conditions for legal interoperability among multiple datasets from different sources to exist[45]:

- use conditions are clearly and readily determinable for each of the datasets,
- the legal use conditions imposed on each dataset allow creation and use of combined or derivative products, and
- users may legally access and use each dataset without seeking authorization from data creators on a case-by-case basis, assuming that the accumulated conditions of use for each and all of the datasets are met.

Similarly, fourteen research infrastructures in the biological, biomedical and environmental sciences developed commonly agreed principles of data management and sharing. The document[46] produced by the BiomedBridges project makes key recommendations on how data management and sharing via the research infrastructures can be supported and encouraged:

1. The RIs encourage data sharing and reuse and support the notion that public funders should encourage Open Access to data from publicly funded research where possible.

[42] https://fidoalliance.org/

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

[44] http://www.earthobservations.org/geoss.php

[45] https://www.earthobservations.org/documents/dswg/Annex%20VI%20-%20%20Mechanisms%20to%20share%20data%20as%20part%20of%20GEOSS%20Data_CORE.pdf

[46] Principles of data management and sharing at European Research Infrastructures, February 2014, http://dx.doi.org/10.5281/zenodo.8304

2. Some data may only be shared under certain conditions and with appropriate safekeeping mechanisms in place, such as personally identifiable data, data subject to ethical or legal restrictions, or restrictions for intellectual property protection.
3. To encourage data sharing, systematic reward and recognition mechanisms are necessary.
4. Proposals for publicly funded research at RIs should include a data management plan concerning the deposition of data in long-term archives that addresses specific resources and activities (including standardisation of data production and curation/annotation).
5. Funding for tools and activities connected to data deposition must be available.
6. Systems, services and resources must be in place to facilitate straightforward data deposition by researchers, including support concerning the necessary data use agreements and consent forms for data with data protection or intellectual property requirements.
7. Systems are also needed to capture and track data provenance and use.
8. To ensure necessary trust by data providers or depositors, RIs must guarantee high standards of security and traceability.

The UK is ranked top of 86 countries by the Open Data Barometer[47], which measures a country's readiness to secure benefits from open data, its publication of key datasets and evidence of emerging impacts from open government data. The 2015 Open Data Institute report "Open data means business: UK innovation across sectors and regions[48]" provides convincing arguments for learning from the private sector when it comes to managing the sharing of public sector data, highlighting the role of value-added providers. The UK's central repository of public sector open data, data.gov.uk, contains nearly 15,000 datasets published with an Open Government License. Examples include geospatial/mapping data (OpenStreetMap[49]), transport-related data (Traveline[50]), demographics/social data (Office for National Statistics[51]) and business data (Companies House[52]).
Best practices include the adoption of Open Data Certificates[53] and the use of Creative Commons[54] public domain licence (CC0) and attribution licence (CC-BY). The Creative Commons attribution and share-alike licence (CC-BY-SA) is also used, but may limit a company's ability to use that data for commercial products and services by requiring them to also attach the same open licence to the data they derive.
Some data can never be "open" in the literal sense and specific authorization may be required (e.g. for medical patient data). However, the "FAIR" principles of Findability, Accessibility, Interoperability and Reusability[55] should still be respected and form the basis for a European Open Science Cloud data policy.
OpenAIRE[56] is a network of Open Access repositories, archives and journals that support Open Access policies. OpenAIRE is a network of more than 580 data providers, integrating more than

47 http://barometer.opendataresearch.org/report/analysis/rankings.html
48 http://theodi.org/open-data-means-business-uk-innovation-sectors-regions
49 http://www.openstreetmap.org/
50 http://www.traveline.info/
51 http://www.ons.gov.uk/
52 https://www.gov.uk/government/organisations/companies-house
53 https://certificates.theodi.org/
54 https://creativecommons.org/
55 http://datafairport.org/
56 https://www.openaire.eu

10 million Open Access publications, related to about 25,000 organisations and 45,000 projects from 3 funders. OpenAIRE is contributing to the Linked Open Data movement, and has recently launched the DLI Service[57], for Data Literature Interlinking. A European Open Science Cloud will be interfaced as a content provider to this resource and as a consumer of service APIs which will allow others to build integrated data discovery and analysis services.
The Zenodo digital repository powered by Invenio and operated by CERN as part of OpenAIRE has been extended with important features that greatly improve data sharing and it has become very popular with researchers from many disciplines around the world. In particular, Zenodo now offers persistent identifiers for data objects so datasets and software from the popular GitHub code repository as well as publications can be cited and includes interfaces permitting metadata to be harvested.
EUDAT[58] is developing a collaborative data infrastructure (CDI) for European research communities. The B2services suite currently consists of the B2SAFE service for implementing data management policies within the EUDAT CDI, the B2STAGE service which provides tools and API's to interact with the EUDAT CDI, the B2SHARE data repository service to store and share research data, the B2FIND service for finding research data, the B2DROP service as EUDAT's DropBox-like service to synchronise and exchange data within a trusted environment.
Metadata and indexing facilities across the set of services from OpenAIRE repositories and EUDAT data services as well as engaged cloud service providers are seen as being particularly relevant.

Data preservation

Data centres operated by the group of publicly funded research organisations and related third parties provide compute and storage services to the research community as well as access to scientific datasets and publications.
Next generation "data factories", including the Research Infrastructures on the ESFRI roadmap, are characterised by data volumes that can extend from multiple PetaByte to several ExaBytes and even beyond (such as the SKA[59]) serving up to several thousands of researchers around the world, as well as many more potential users via Open Access.
Data preservation – for current and future re-use and sharing – is a fundamental component of on-going data management plans and there is common agreement on the OAIS model (ISO 14721) together with closely related standards (ISO 16363 and 16919). This approach focuses almost exclusively on management of repository data and additional capabilities are needed to satisfy the key use cases driving data (knowledge) preservation, sharing and re-use in a multi-disciplinary environment. These additional capabilities require a good understanding of who will re-use the data ("the consumers") together with knowledge capture from the Open Scientists who are "the producers" (OAIS terms) of the data.
Preservation policies implemented in a measurable and certifiable manner across shared e-infrastructures together with domain and institutional repositories would stimulate much wider re-use of data through the captured and preserved knowledge, as well as the capability to preserve and re-use data and knowledge for significantly longer periods of time. This translates to a larger return on investment for the funding agencies, together with associated scientific, educational and cultural benefits.

[57] https://www.openaire.eu/dliservice

[58] http://www.eudat.eu

[59] In the preprint "Imaging SKA-Scale data in three different computing environments" Richard Dodson (ICRAR) et al., 2 November 2015, the authors compared commercial cloud services (AWS), a cluster and a HPC installation on performance, usability and cost for SKA image workloads and rated the cloud service services favourably.

Reproducibility of research

Federated cloud-based services will improve reproducibility and transparency (serving Responsible Research & Innovation principles, as envisaged by the OpenAIRE & FOSTER[60] report[61]), facilitating wider access for the knowledge-based industries, and letting the free flow of ideas and knowledge speed up innovation and delivery of added value to the marketplace. The RDA Reproducibility Interest Group defined a set of high-priority services for reproducibility of Open Science, as follows[62]:

1) Persistent linking and availability of data and code (via repositories or other mechanisms) used in the generation of published research results, with the publication itself;

2) Development, encouragement, and adoption of meta-data standards for data and code, especially for those linked to publications;

3) Development, encouragement, and adoption of data and code publication, authorship, and citation practices, especially for those linked to publications;

4) Development and adoption of appropriate tools and computational infrastructure that enable: the sharing of research workflows and permit replication of computational scientific findings; the persistent linking of all digital scholarly objects used to generate research findings such as datasets in repositories; and versioning of digital scholarly objects to ensure persistent reproducibility.

To support reproducible science a European Open Science Cloud will need to integrate a network of Zenodo-like repository services and link them to the computing services to ensure that registering and storing research outputs becomes a simple and standard operation at the end of the compute cycle. In addition, this will enable users to analyse to re-analyse the registered data with the referenced codes and extend it with their own software directly contributing open science workflows.

Governance

- Shadow IT and the changing role of IT departments
- Inclusive governance structure
- End-users and procurers at the heart of the decision making process

Disruptive technologies such as cloud offer a myriad of possibilities but come with new pressures for service provisioning. Cloud technology is more accessible to users meaning they are more knowledgeable about what products and services they need and due to the rapidly growing and easily accessible cloud services market, they have alternatives to their traditional supplier for acquiring them. Around the world, IT departments are being by-passed as users procure their own cloud services directly. This a growing tendency by individuals and workgroups to sign-up for commercially operated cloud services without any involvement from their IT departments which creates serious risks for public organisations. The risks from such shadow cloud services include issues with data security, transaction integrity, business continuity and regulatory compliance. Consequently the role of service provisioning for IT departments has to change to become more of a broker for technology and services. In this new role it is important for the IT department to know what is available on the market, how well it works, to be able to assess providers, validate security, understand service levels and ensure policies and legislation are respected. So there is an urgent need to organise the introduction of commercial cloud services

[60] https://www.fosteropenscience.eu/project/

[61] https://www.fosteropenscience.eu/sites/default/files/pdf/927.pdf

[62] https://rd-alliance.org/sites/default/files/case_statement/RDA-ReproducibilityIG-Revised-2_0.pdf

in the public research sector in a consolidated and secure manner. Forming a network of public research organisations that can procure cloud services will attract the interest of service suppliers as well as funding agencies. The majority of this procurement funding will be directed to service providers and the approach has the advantage of permitting the procuring organisations to choose which services and providers receive these funds and thus represents a change to the established funding model for public sector IT services. Bringing together the public and private sector in the innovation cycle will strengthen Europe's global competitiveness and encourage the creation of new and sustainable jobs and the promotion of growth.

The introduction of procurement of pay-per-use cloud services by funding agencies and research organisations on behalf of their end-users represents a significant change to e-Infrastructures and will impact the governance model. Currently publicly funded e-Infrastructures are supplier driven while a European Open Science Cloud puts procures and users at the heart of the decision making process. It will be necessary to establish an inclusive governance structure where all the stakeholders are represented and avoid a monopoly of any procurer, supplier or research community. The governance principles have to ensure the interests of both public and private participants are met and that a European Open Science Cloud becomes sustainably attractive and beneficial for all stakeholders from both sectors.

A European Open Science Cloud will be a cornerstone of an open science commons and its governance model needs to take into account the realities of the public research sector with the following objectives:

1. Enable integration of existing e-Infrastructures with commercial cloud computing effectively and efficiently
2. Ensure alignment with the Digital Single Market, foster coherence, equitability and inclusiveness
3. Ensure participation of all stakeholders and fair balance of their needs and interests
4. Ensure transparency, openness and responsiveness
5. Ensure value for money and fair incentives and returns
6. Continuously manage legal and ethical compliance and other risks
7. Ensure accountability and responsibility of stakeholders and decision makers
8. Manage the identity and brand of a European Open Science Cloud and ensure sustainable innovation and growth.

In addition a European Open Science Cloud would become a critical ICT infrastructure for the European Research Area and would need to be protected by identifying vulnerabilities and ensuring an operational security plan is in place to minimize the detrimental effects of disruptions. The governance structure is composed of several bodies as shown in Figure 1.

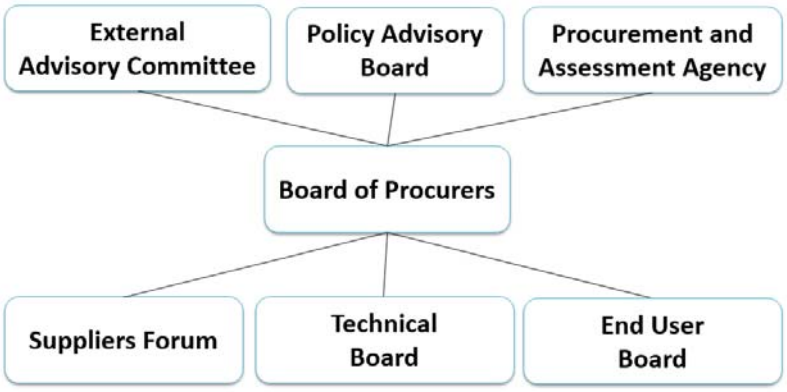

Figure 1 Governance Structure

Each body in the governance structure has a specific role and composition:

- *Board of Procurers* – this grouping of all procurers (research organisations, funding agencies etc.) is the ultimate decision making body of a European Open Science Cloud.
- *Policy Advisory Board* – experts addressing legal, contractual and ethical aspects to ensure that a European Open Science Cloud is compatible with European legislation. It would ensure the application of best practises for the contractual aspects of delivering cloud services including service level agreements implementing recognised policies for trust, security and privacy notably for data protection; certification requirements; a code of conduct; and terms and conditions that respect European legislation.
- *Procurement and Assessment Agency* – one or more organisations commissioned by the Board of Procurers to perform the joint procurement and centralised billing of services on behalf of all procurers as well as gather data necessary to measure a set of agreed Key Performance Indicators (KPIs). Having an organisation to oversee the procurement process, certify and enrol service providers as well as handle the contractual arrangements between suppliers and procurers with centralised billing would simplify the operation and expansion of a European Open Science Cloud.
- *End-User Board* – grouping of end-users from engaged research communities including the *long-tail of science* to provide a consultative opinion on the relevance and added value of delivered services. End-users contribute applications software, data and publications. Responsibility for all data that is made available, linked or accessed via the services provided by the project remains with the data providers and must have been obtained in accordance with the laws and regulations in operation in the country in which the data provider resides. This includes any requirement for approval from an appropriate ethics committee or other regulatory body.
- *Suppliers Forum* – consultative forum open to all cloud service suppliers (commercial and publicly funded) who want to enter into a dialog with the procurers and end-users and provide input on all aspects of a European Open Science Cloud.
- *Technical Board* – grouping of technical experts to assess the technical maturity and suitability of services, including security aspects.
- *External Advisory Committee* – grouping of external experts from the public and commercial sectors that will provide advice to the Board of Procurers on the state and future directions of a European Open Science Cloud.

The details of the appointments, voting rights and procedures of the various bodies remains to be defined, together with how to handle the situation where a participating organisation is both a service supplier and procurer.

The relationship of a European Open Science Cloud to H2020 assumes that the e-infrastructure programme includes two parallel tracks, *production* supporting the sustainability of pan-European e-infrastructures and *innovation* representing changes to the production services. The *production* track builds on national structures to ensure the long-term operation, maintenance and evolution of a set of services provided to a wide range of user-groups across the borders of individual member states. The *production* track delivers a portfolio of horizontal core networking, compute and data services that provide the backbone of a European Open Science Cloud through the integration and consolidation of e-infrastructure platforms and the creation of a common service catalogue. The *innovation* track is organised as short-term, competitive cycles where the best proposals are developed in to prototypes that are assessed against agreed criteria to become candidates for inclusion as production services. A service lifecycle manages individual services starting from their conception, development in the *innovation* track, transition into the *production* track, operation and eventual retirement. The transition from prototype service to production service is a decision that involves the stakeholders represented in the governance structure described above. It is expected that Research Infrastructures, including ESFRI projects, will become stakeholders of a European Open Science Cloud as procurers and end-users. The Internet2 NET+[63] initiative contains many of the aspects necessary for the governance of a European Open Science Cloud and can be a good source of inspiration.

Investment

- Proprietary solutions are not solutions
- In-house investment
- Investment in skills
- Long-term strategy

Proprietary solutions are not solutions

The cost of providing licenses to popular propriety software packages for the users of Research Infrastructures continues to increase. As an example, between 2008 and 2014, CERN's spending on software doubled without any significant increase in the number of licenses. Moving to a cloud model where software licenses are rented on a pay-per-use basis may help stem this increase. But some proprietary software packages have an effective monopoly in the research domain and their market dominance can offset any potential savings.

It is essential that there is appropriate investment in open source solutions in key domains so they can be supported by multiple providers. We must leverage the richness in the diversity of European suppliers and to match it with the expertise available in production e-Infrastructures, demonstrating the technical feasibility of interoperability between these players.

The European Technology Platform for High Performance Computing project[64] published a Strategic Research Agenda for achieving HPC leadership in Europe[65] which specifically highlights the upcoming big-data challenges for leading research activities and the relevance of cloud services:

[63] http://www.internet2.edu/vision-initiatives/initiatives/internet2-netplus/

[64] http://www.etp4hpc.eu/

[65] http://www.etp4hpc.eu/wp-content/uploads/2013/06/ETP4HPC_book_singlePage.pdf

*"Europe is in a unique position to excel in the area of **HPC Usage and Big Data** owing to the experience level of current and potential users (and the recognition of the importance of data by such users as CERN, ESA, and biological data banks) and the presence of leading ISVs for large-scale business applications. Europe should exploit that knowledge to create competitive solutions for big-data business applications, by providing easier access to data and to leading-edge HPC platforms, by broaden the user base (e.g., through Cloud Computing and Software as a Service (SaaS), and by responding to new and challenging technologies."*

There is no clear business case for purely commercial HPC services at the scale of PRACE tier-0 installations but smaller-scale commercial 'HPC in the cloud' offerings are starting to appear on the market. This will help address the shortfall between supply and demand for capability HPC services as seen as PRACE[66] where typically only one third of the requests can be satisfied. The use of capability HPC services by the commercial sector, in particular SMEs, is being investigated by the EC funded Fortissimo project[67]. This will make hardware, expertise, applications, visualisation and tools available and on a pay-per-use basis. In parallel, the UberCloud Marketplace[68] is offering on-demand access to HPC services for individual engineers and scientists.

Public investment

The steps described above will need considerable public investment as well as investment from commercial service providers to bring the platform together. In order for the research community to be able to benefit fully from the existence of a European Open Science Cloud, it has to expand beyond the basic IaaS level and provide higher-level services that are closer to the needs of the daily work of a researcher. The HNSciCloud PCP project provides a vehicle for joint investment in IaaS services and a similar approach should be envisaged for higher-level software services. The natural follow-on step for successful PCP projects is to procure at a larger scale with PPI co-funded projects that could significantly increase the capacity and impact of a European Open Science Cloud.

This will take a sustained investment by all the stakeholders in both the public and commercial sectors, not only in cloud technology, supporting infrastructure and strategic software but also in end-user facing services which will simplify access to a European Open Science Cloud.

Significant investment in software capability will be absolutely essential to obtain the best performance from current and future computer and storage architectures. Many sciences today benefit from commodity CPU and disk storage but there are significant architectural changes in modern CPUs (memory layout, I/O paths, accelerators, vector units, etc.) which means it will be necessary for science to invest heavily in software and training to be able to migrate application codes and programmers and fully exploit these new technologies. This investment in software is essential to maintain European competitiveness in this area, and should include coordination of existing expertise to the benefit of diverse communities.

Investment in skills

The design, creation and operation of e-infrastructure services are essential tools in the development of skills and competencies for the European market. The ability to fully exploit the potential for knowledge and job creation that is locked-up in the datasets and algorithms at the centre of Open Science will require the nurturing of a new generation of data scientists with a

[66] PRACE annual report 2014, May 2015, ISBN 978902169416

[67] http://www.fortissimo-project.eu/project/the-project.html

[68] https://www.theubercloud.com/store/

core set of ICT skills. The EIROforum organisations have core competences in training and education which can contribute to this activity. A European Open Science Cloud can build on this and similar initiatives to help train the next generation of IT-savvy researchers, and also improve outreach to the general public.

Long-term strategic investment

A European Open Science Cloud must leverage the investments already made in Europe for the publicly funded e-infrastructures and commercial cloud services. Through Horizon 2020, the EC and national funding agencies have recently confirmed their commitments to GÉANT, AARC, EGI, OpenAIRE, EUDAT and PRACE. In order to ensure full synergies, DG CONNECT foresees that e-infrastructure projects will be grouped into clusters of related projects. This new phase of funding for the clusters of e-infrastructure projects offers the EC a window of opportunity and a means to focus on establishing a European Open Science Cloud. In parallel DG RTD intends to fund a pilot action that will encourage the uptake of a European Open Science Cloud by the Research Infrastructures. Close coordination between DG RTD and DG CONNECT funded projects will facilitate the establishment of a European Open Science Cloud.

The financial plan for a European Open Science Cloud should be designed so that the services can be sustained by their operating organisations according to a continuum of funding models ranging from sponsored resources for peer-reviewed scientific cases to communities who would pay for the services they receive. Additional resources will be required in order for these services to be expanded and to serve a wider range of users. The European Commission together with regional, national and thematic funding agencies will need to become stakeholders and contribute to the expansion of European Open Science Cloud. The guiding principle is that funding from such stakeholders will be focused on innovation of services and uptake by new user communities and business actors while the operational costs will be borne by the operating organisations and the user communities.

Below is a non–exhaustive list of areas where funding agencies can contribute to the creation of a European Open Science Cloud:

- Development of new services to be deployed on the e-infrastructure. Significant effort will be required to co-develop scalable services that can operate in a distributed virtual environment and serve a wide range of users.
- Financial incentive scheme to increase adoption of services by users including 'long-tail of science' research groups and SMEs.
- Engaging the use of the services by new research communities (e.g. curation of data-sets, connection of identity federations, deployment of community specific services etc.)
- Development of training and educational activities building on the cloud services to maximise their impact. This can also include expansion of services to support for volunteer computing so that researchers can build citizen-cyberscience communities and further engage the general public in science.
- Organisation of user forum events as well as outreach and dissemination to a range of audiences and production of material for policy related activities.
- International collaboration (beyond Europe) through interoperation with equivalent structures in other regions of the world.
- Expansion of the network of service providers across the European member states to address national and thematic needs.

Many research organisations that operate research infrastructures do not have the mandate to provide cloud services to their users for the management and processing of their experimental data. This represents a gap in the scientific lifecycle and a missed opportunity to highlight the results and impact of public funded research. These research organisations will require assistance to bridge this gap by supporting their users so they can make use of cloud services to manage and process their experimental data.
The European Commission's INFRASTRUCTURES 2016-2017 work programme foresees a pilot action addressing the federation, networking and coordination of pan-European research infrastructures and clouds for the purpose of increasing research and science data availability and use. It also foresees Data and Distributed Computing e-infrastructures for Open Science which should cooperate with the pilot action. The combined focus of these funding calls should provide an incentive for the existing e-infrastructures and Research Infrastructures to work together to form the basis of a European Open Science Cloud. Looking further ahead, the EC has taken steps to ensure funding for GÉANT over the full duration of H2020 by introducing 'Framework Partnership Agreements' (FPA). The FPA model represents a more long-term engagement that could encourage the integration of e-infrastructures co-funded via EC projects into the Research Infrastructures' computing models who need to plan for future decades[69]. The application of the FPA approach to a European Open Science Cloud could establish the basis for the European Research Area's digital commons and lead towards Science 2.0[70].

Conclusions

Cloud computing represents a paradigm shift in the way IT resources are provisioned for research communities. Traditionally the IT departments of research organisations have developed and operated in-house the services that their users required. But commercial cloud services are a disruptive technology with easy-to-use commodity services made available often on a 'freemium' basis to users at a global scale. Consequently the role of IT departments is changing as users by-pass their traditional service provision channels to get the on-demand services they want and thereby introducing shadow IT services that are outside the policy and security boundaries of research organisations. This is impacting data intensive science and how e-infrastructure services are used by researchers and judged by funding agencies.
This wave of change is taking place within the broader context of Open Science bringing ever-greater transparency, accessibility and accountability, wherein stakeholders in the research process increasingly expect to be able to access and reuse the outputs of taxpayer funded research.
From the grassroots, Open Access first emerged from the High Energy Physics scholarly research community[71], who saw benefit in no longer waiting for traditional publication schedules before sharing research findings (and, subsequently, data and software code). Top-down, governments and other funders see openness as a catalyst for increasing public and commercial engagement with research, bringing about both societal and commercial benefit.
This new reality represents a threat to the established service procurement and delivery models but also an opportunity. In an era of rationalisation and budget concentration, all means of optimising service delivery and reducing operational costs must be considered.

[69] EIROforum discussion paper: Long-term sustainability of Research Infrastructures, http://www.eiroforum.org/downloads/20150325_discussion-paper-research-infrastructures-sustainability.pdf
[70] http://ec.europa.eu/research/consultations/science-2.0/background.pdf
[71] Open Access: Unlocking the Value of Scientific Research, Richard K. Johnson (SPARC), March 2004, http://www.sparc.arl.org/sites/default/files/media_files/OpenAccess_RKJ_preprint.pdf

The EIROforum members have extreme IT needs that increase with the progress of the research infrastructures they operate while the budget envelope for IT remains, at best, unchanged.
Cloud computing and the cloud services market did not exist when the computing models for many ESFRI research infrastructures were conceived.
These computing models must evolve to become more agile and opportunistic, capable of using IT capacity in whatever form it is delivered, be it in a grid, cloud, HPC or even a volunteer structure.
We expect commercial cloud services to play an increasing role in these computing models.
Commercial sectors are investing heavily in cloud services leading to a rapid expansion of the market and a breath-taking rate of innovation that the publicly funded research sector cannot match but can leverage and so profit from such advances.
A European Open Science Cloud represents a strategic vision that can be a vector for introducing change in the service provisioning and computing models for the publicly funded research sector in the medium to long term.
A European Open Science Cloud has the potential to greatly improve the provisioning of IT services for Research Infrastructures to address their big data needs. It can encompass all the phases of the research lifecycle and offer a platform of joint innovation for the public and private sectors. It will significantly change the way IT services are procured, organised and funded. The key challenges are integrating frequently changing technologies, managing the complexity and identifying the optimal organisational and financial models. Researchers must be convinced that they will not lose control of their precious data. It is an ambitious undertaking requiring the active engagement of many stakeholders and careful planning of the technical, financial, legal and governance aspects. For it to succeed it must become a priority for all the actors involved with monitoring by the funding agencies and regular assessment by the user communities.

This position paper is a rallying call for adoption of such a strategic approach – within the EC and other funding bodies to work with the operators of Research Infrastructures.