The Open Science Commons for the European Research Area

Tiziana Ferrari, Diego Scardaci, and Sergio Andreozzi

Abstract Nowadays, research practice in all scientific disciplines is increasingly, and in many cases exclusively, data driven. Knowledge of how to use tools to manipulate research data, and the availability of e-Infrastructures to support them for data storage, processing, analysis and preservation, is fundamental. In parallel, new types of communities are forming around interests in digital tools, computing facilities and data repositories. By making infrastructure services, community engagement and training inseparable, existing communities can be empowered by new ways of doing research, and new communities can be created around tools and data. Europe is ideally positioned to become a world leader as provider of research data for the benefit of research communities and the wider economy and society. Europe would benefit from an integrated infrastructure where data and computing services for big data can be easily shared and reused. This is particularly challenging in EO given the volumes and variety of the data that make scalable access difficult, if not impossible, to individual researchers and small groups (i.e. to the so-called long tail of science). To overcome this limitation, as part of the European Commission Digital Single Market strategy, the European Open Science Cloud (EOSC) initiative was launched in April 2016, with the final aim to realise the European Research Area (ERA) and raise research to the next level. It promotes not only scientific excellence and data reuse, but also job growth and increased competitiveness in Europe, and results in Europe-wide cost efficiencies in scientific infrastructure through the promotion of interoperability on an unprecedented scale. This chapter analyses existing barriers to achieve this aim and proposes the Open Science Commons as the fundamental principles to create an EOSC able to offer an integrated infrastructure for the depositing, sharing and reuse of big data, including Earth Observation (EO) data, leveraging and enhancing the current e-Infrastructure landscape, through standardization, interoperability, policy and governance. Finally, it is shown how an EOSC built on e-Infrastructures can improve the discovery, retrieval and processing

T. Ferrari • S. Andreozzi
EGI Foundation, Amsterdam, The Netherlands
e-mail: Tiziana.Ferrari@egi.eu; sergio.andreozzi@egi.eu

D. Scardaci (✉)
EGI Foundation & INFN Catania Division, Amsterdam, The Netherlands
e-mail: diego.scardaci@egi.eu

P.-P. Mathieu, C. Aubrecht (eds.), *Earth Observation Open Science and Innovation*,
ISSI Scientific Report Series 15, https://doi.org/10.1007/978-3-319-65633-5_3

capabilities of EO data, offering virtualised access to geographically distributed data and the computing necessary to manipulate and manage large volumes. Well-established e-Infrastructure services could provide a set of reusable components to accelerate the development of exploitation platforms for satellite data solving common problems, such as user authentication and authorisation, monitoring or accounting.

Creating the European Research Area: The "Open" Approach

The European Research Area (ERA) was endorsed by the European Council in 2000 (European Commission 2014a) as a way to build "*a unified research area open to the world based on the Internal Market, in which researchers, scientific knowledge and technology circulate freely and through which the Union and its Member States strengthen their scientific and technological bases, their competitiveness and their capacity to collectively address grand challenges*" (European Commission 2012).

Although several actions for the ERA implementation have been undertaken, such as the establishment of the European Strategy Forum on Research Infrastructures (ESFRI) (2017) or the development of an e-Infrastructure for connectivity, high performance, grid and cloud computing and data, the rapid growth of scientific data has highlighted the need for new methodologies.

In its Horizon 2020 consultation report on Open Infrastructures for Open Science, the European Commission concluded that "*open data e-Infrastructures increase scope, depth and economies of scale of the scientific enterprise. They are catalysts of new and unexpected solutions to emerge by global and multidisciplinary research. They bridge the gap between scientists and the citizen and are enablers of trust in the scientific process*" (Horizon 2020 consultation report n.d.), electing the *open* approach as a core aspect of the ERA.

This vision implies a European dimension beyond national and regional approaches, and an increase in capacities and capabilities. Research Infrastructures, including e-Infrastructures, are enabling instruments which provide (European Commission 2017) "*facilities, resources and services that are used by the research communities to conduct research and foster innovation in their fields. They include: major scientific equipment (or sets of instruments), knowledge-based resources such as collections, archives and scientific data, e-infrastructures, such as data and computing systems and communication networks and any other tools that are essential to achieve excellence in research and innovation.*"

Problems to Solve

Implementing such vision requires overcoming various barriers.

Lack and/or Incomplete Roadmaps for Research- and e-Infrastructures

The establishment of roadmaps at a national and European level is progressing at different speeds. Policies of access to the European research system are not homogeneous, services are not always available to researchers, and knowledge transfer is not yet a strategic mission of many public organisations. This represents a risk to the coherence of the European-scale initiatives. For example, with inconsistency in access policies, countries may find their national research communities excluded from crucial European-wide services. Shared understanding of the strategic objectives, the alignment of national strategies and proceeding faster in the development of the national policies are recognised by the European Council as necessary steps: "*the Member States should accelerate national reforms, where necessary, to boost the EU's potential in research, development and innovation*" (European Commission 2014b).

The lack of alignment in policies hinders the sustainability of e-Infrastructures of European scale. A consensus solution would offer them the certainty of a long-term commitment to allow for the construction of technical infrastructures and high-quality services.

Fragmented Solutions and Policies for Access to Data and Knowledge

Access to research data and existing bodies of knowledge has moved towards openness, but has not yet achieved it. Efforts to support Open Access are hugely beneficial, but the scientific publishing sector is still adjusting to the evolution towards openness. The resulting inconsistency restricts exploitation of research results.

Insufficient Cooperation Between Public and Private Sector

There is a general agreement amongst research communities and e-Infrastructure providers that working with the private sector is desirable. However, the mechanisms and models of engagement are not yet well understood, due to a range of technical and cultural factors. Steps such as SME instruments in Horizon 2020 have been broadly appreciated, but many challenges remain, due to different local strategies, restrictions in policies and insufficient European coordination of existing efforts.

Lack of National and European Organization Between All Stakeholders

While the vision of the e-Infrastructure commons (e-Infrastructure Reflection Group 2017a) has been embraced by many groups, the landscape remains fragmented and includes too many narrowly focussed services based on closed platforms that limit the portability of data, applications and knowledge. In addition, we are still missing a common body of knowledge and a coordinated broad programme for knowledge transfer, also including the private sector and the single researchers. This results in a barrier to entry for the emerging research infrastructures, skills and professions.

Today, services are provided by a broad range of sector-based, national and pan-European providers. Technical interoperability and service integration and management are becoming increasing important to ensure the support of the entire research lifecycle. Nevertheless return on investment for national and European funds dedicated to service support, is still sub-optimal due duplicated efforts, limited sharing of technical solutions and some lack of coordination.

As digital science services such as e-Infrastructures move toward sustainable operating models, the need for coordination and coherence is rapidly increasing.

Many Providers Without a Single Market

Despite some efforts, we lack a single portfolio of services to provide a "backbone" of European ICT capabilities. The existing offering is fragmented, with different policies of access and different channels for engagement with the user community. One of the main policy issues that are hindering the uptake of services at national level is the lack of a business model and procurement framework that allows international research collaborations to rely on research e-Infrastructure services in the long term.

In addition, the lack of an established national Research- and e-Infrastructure component in some countries prevents the access even to the existing services. Such European-scale coherence must be achieved to make the ERA a reality. Insufficient competition in national research systems, barriers to pan-European cooperation and restricted circulation of and uneven access to scientific knowledge are also recognized as issues by the ERA implementation assessment.

The Open Science Commons

The Open Science Commons have been defined as an overarching policy designed to overcome the barriers preventing the implementation of the ERA. The Open Science Commons seek to encompass all the elements required for a functioning ERA: **research data, scientific instrumentation** (such as the Large Hadron Collider, the Copernicus satellites or Square Kilometre Array), **ICT services** (connectivity, computing, platforms and research-specific services such as portals), and **knowledge**. The Open Science Commons evolve from two preexisting and broadly accepted ideas:

- **Open Science** (also referred as Science 2.0 (European Commission, n.d.)) for the opening of knowledge creation and dissemination to a multitude of stakeholders, including society in general. Open Science supports multiple perspectives, from infrastructure-oriented views seeking to increase efficiency through better tools and services, to public-oriented views trying to ensure citizens have access to scientific knowledge (Fecher and Friesike 2014).

- A **Commons** to reinforce the need of sharing within a community in a way that allows non-discriminatory access, while ensuring adequate controls to avoid congestion or depletion when the capacity is limited (Frischmann 2013). The Commons concept is embedded in Open Science, which stresses cooperation to reduce barriers to collaboration, knowledge transfer and sharing of results. In the last decade, this has been driven by the digitalisation of the research process and by the globalisation of the scientific communities. Infrastructures often generate spill overs that result in large social gains and it is recognised that applying commons management principles maximises such benefits.

The Open Science Commons relies on **four pillars**, representing a wide range of groups, providers and community types:

- **Data**. The data that is the subject matter for research. It should be dealt with according to the principles of open access and open science, while maintaining trust and privacy for researchers.
- **e-Infrastructures**. The technology and technical services supporting researchers, building towards integrated services and interoperable infrastructures across Europe and the world.
- **Scientific instruments**. The equipment and collaborations that generate scientific data, from small-scale lab machines to global collaborations around massive facilities.
- **Knowledge**. The human networks, understanding and material capturing skills and experience required to carry out open science using the three other pillars.

Managing shared resources as a Commons maximises benefits for society. In the area of digital infrastructure, this has already demonstrated great benefits (the Internet itself is an example). Applying this principle to the Open Science process is expected to improve the stewardship from the funding agencies, in collaboration with the stakeholders, through mechanisms such as public consultations. This will increase the perception of shared ownership of the infrastructures. It will also create clear and non-discriminatory access rules together with the sense of shared ownership, which stimulates a higher level of participation, cooperation and social reciprocity.

The European Open Science Cloud

As part of the European Commission Digital Single Market strategy (European Commission 2015), the European Open Science Cloud (EOSC) initiative was officially launched in April 2016 by the European Commission. EOSC promotes not only scientific excellence and data reuse but also job growth and increased competitiveness in Europe, and drives Europe-wide cost efficiencies in scientific infrastructures through the promotion of interoperability on an unprecedented scale.

According to the first report of the High Level Expert Group on the European Open Science Cloud (EOSC) (European Commission 2016) appointed by the European Commission, EOSC has been defined as a support environment for Open Science aiming to "*accelerate the transition to more effective Open Science and Open Innovation in a Digital Single Market by removing the technical, legislative and human barriers to the re-use of research data and tools, and by supporting access to services, systems and the flow of data across disciplinary, social and geographical borders*". Indeed, the term "cloud" has been interpreted as "*a metaphor to help convey the idea of seamlessness and a commons*".

As guiding principles, the experts underlined how:

1. The EOSC must integrate with other e-Infrastructures and initiatives in the world, implementing a lightweight interconnected system of services and data, which follows the federated model.
2. Open refers to the accessibility of services and data under proper non-discriminatory policies ("*not all data and tools can be open*" and "*free data and services do not exist*").
3. EOSC should include all scientific disciplines.
4. The term cloud should not refer to ICT infrastructure but to a commons of data, software, standards, expertise and a policy framework relevant to data-driven science and innovation.

In the expert vision, the EOSC will be an accessible infrastructure for modern research and innovation implementing an internet of Findable Accessible Interoperable and Reusable (FAIR) data and services (Wilkinson et al. 2016). It should be based on standards, best practices and infrastructures, completed with adequate human expertise. The FAIR principles have to be supported, with a particular attention to the reuse of the open and sensitive data. The data needs to be sustained with a plethora of elements (standard, tools, processing pipelines, protocols) that make possible and simple their reuse, enabling a data driven knowledge discovery and innovation. Furthermore, the data science profession needs to be fostered to guarantee a professional data management and the long-term data stewardship.

In Europe, domain-specific European Research Infrastructures and cross-domain ICT e-Infrastructures, as well as other disciplinary and cross-disciplinary collaborations and services are already well established. These can be considered the ground for EOSC. However, the realisation of "*the ambition of increased seamless access, reliable re-use of data and other digital research objects, and of the collaboration across different services and infrastructures*" (that guarantees non-discriminatory access and reuse of data to both the public and private sector), requires further enhancements on this landscape with the aim to turn the "*ever increasing amounts of data [. . .] into knowledge as renewable, sustainable fuel for innovation in turn* *to meet global challenges*". The EOSC is the instrument defined by the European Commission to foster such evolution towards the realisation of the so-called Open Science.

This idea highlights the strong relationship between the implementation of the ERA through the Open Science, the Open Science Commons and the EOSC. In such context, the High Level Expert Group designed by the EC reported a list of key Open Science "trends" that should be taken into account in the EOSC design. They cover several aspects such as new modes of scholarly communications (e.g. applications, software pipelines and the data itself), new incentives to stimulate data publishing and tool sharing, fostering the emerging data science profession, the cross-disciplinary collaboration, support for innovative SMEs, the creation of an ecosystem, methodology and tools to reproduce of current published research, etc.

Open Science Commons for the EOSC

In this section, we propose the Open Science Commons as a possible approach to the implementation of the infrastructure and governance pillars of the EOSC, leveraging and enhancing the current Research Infrastructure landscape, through standardization, interoperability, policy and governance. The section discusses the current state of play and blockers for implementing an Open Science Cloud, and explains how the approach could strategically advance the competitiveness of research in Europe by providing research data and community-specific tools as services through a platform that supports the participatory principle of Open Science.

EOSC Architecture and Services

The Open Science Commons infrastructure comprises research data, processing services, applications, virtual laboratories and tools, relying on existing *federated* data and storage facilities from local, regional and international infrastructures, which can be organized as a federation of hubs, where each hub is a node of the EOSC providing certain capabilities in a standard and interoperable manner (Fig. 1). The cloud hub does not duplicate the data services of the reference institutional and disciplinary repositories, but rather make these accessible in an environment that can enrich the data itself with supplementary added-value services and can provide scalable access where necessary by collocating computing and data.

In the proposed approach the Open Science Commons can be implemented as a federation of cloud hubs (in Europe and beyond), based on the cloud service provisioning paradigm. The cloud hub will provide various capabilities in a federated, integrated way: a virtual space providing data, tools, applications and processing, with the hubs interconnected by a mesh of high-bandwidth links to

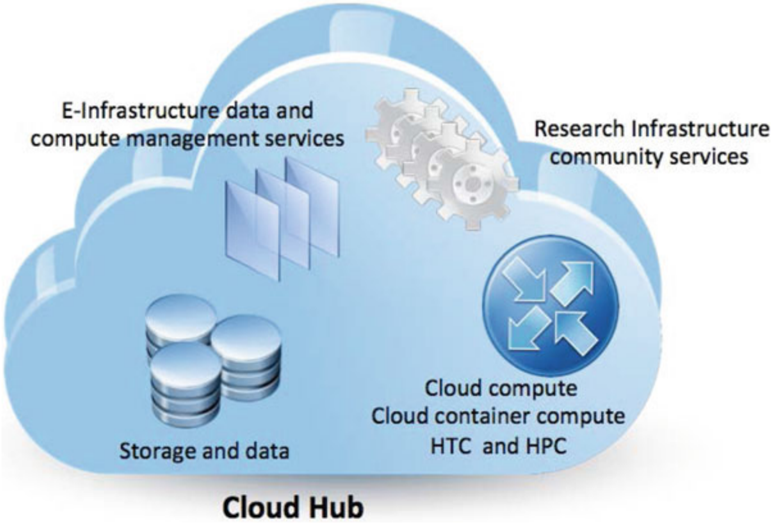

Fig. 1 Examples of functionalities delivered by a cloud hub of the EOSC

ensure efficient virtual access to public and managed access research data, which is provided as a service by the hub (DaaS). Within the cloud hub, the data provider always retains access control to data.

Being based on virtualization, clouds facilitate sharing, reuse and the combined offer of data and tools. Cloud federations enable “local hosting” and “control sharing” capabilities to respect ownership and allow accessibility for distributed communities of users. In addition, federation allows the implementation of hybrid models where private, community and public providers can contribute data and services in multi-supply environment.

Furthermore, the federation of hubs provides a multi-level governance model where different governing bodies of the Commons can coexist. This governance model meets the needs of European policies, regulations, restrictions and business models. By allowing distributed access to data, relocation into centralized repositories is no longer necessary. This greatly simplifies the integration of data and tools from multiple domains and regions. When expertise about how to use specific research domain data and tools is accumulated within the same research community, then the community becomes an ideal incubator for a hub and can contribute to the implementation of the EOSC federated infrastructure.

Realizing a Federated Approach to Research Data

The EOSC needs to aggregate offer and demand by exposing its assets via a marketplace to make research data, the related tools and knowledge discoverable, accessible and reusable. The marketplace would federate existing research data sets that are provided by data preservation organizations that can ensure compliance to a set of defined quality standards. The EOSC Marketplace should be open to any data provider that can ensure compliance to international data standards and best practices, as well as to European data regulations.

The Marketplace should be open for access to any research community that is willing to become a data provider. Through the marketplace, datasets and the associated metadata are discoverable. The marketplace provides information about intellectual property rights and access policies for reuse for research and commercial purposes when allowed.

Offering of Scalable Access to and Analysis of Research Data for Reuse

Making data findable is not sufficient. Local download of large volumes of data can be a huge barrier for downstream efficient analysis. EOSC should provide distributed data mirroring and caching capabilities based on federated cloud storage, where research data can be temporarily stored for scalable access in agreement with the data providers, and processed via integrated computing platforms.

This capability is not a duplication of existing data infrastructures, but rather provides efficient access to big data that is produced worldwide. The governance of the service would require an organization acting as a broker towards the data providers worldwide for the procurement of Data as a Service to the whole ERA.

A premium access could be also offered implementing a federation of large cloud hubs connected by a broadband network infrastructure for efficient replication. The network of Tier-1 hubs would be complemented by a network of disciplinary Tier-2 hubs, providing complimentary access to discipline-specific datasets. The cloud hub federation would be complemented by co-located services offering high throughput and high performance cloud computing.

Integrating (Shared) Tools and Applications

Knowledge cannot be extracted from data without the availability of specialized tools and applications (e.g. text mining). The EOSC would provide a library of community-specific applications and tools. This community platform should be open for publishing to any researcher. For greater specialization, the EOSC should provide PaaS and SaaS services that are community-specific and that could be dynamically deployed with a focus on single researchers or small research groups. These are provided in the form of managed services by the Research Infrastructures. By increasingly sharing models and modelling tools, researchers and research communities can capture the steps of the digital research processes they carry out for excellent science. With suitable abstractions and robust provenance capabilities, such models and tools would enable the repeatability, and therefore the incremental improvement of research practices and processes within and across research teams.

Provisioning of Services for Depositing Data for Resource-Bound Users

Through virtual access, the EOSC will federate e-Infrastructures to provide services for the long tail of science, citizen scientists, the general public and other stakeholders that cannot benefit from those at institutional and/or national level, but supports open research data.

EOSC Service Integration and Management

As the EOSC will involve multiple suppliers, the EOSC requires a service integration and management approach to managing suppliers and integrating them to provide a single customer-facing interface. The approach allows integrating interdependent services from various internal and external service providers into end-to-end services.

EOSC service integration and management plays different roles. It allows the end-to-end composition of services, the alignment of scope, value, service catalogue entries and their specifications across EOSC providers, the management of relationship and collaboration between the providers, and the EOSC standardisation guidelines. By doing so, the EOSC governance becomes accountable for the integrated services that are delivered and for the central point of control between demand and supply.

The EOSC service integration and management system includes the entirety of activities performed by service providers to plan, deliver, operate and control services offered to customers. These activities are directed by policies and need to be structured and organised by processes and procedures.

The processes include the management of Business Development and Stakeholders, Service Order, Service Portfolio, Service Level and Reporting, Customer Relationship, Supplier/Federation Member Relationship, Capacity, Service Availability and Continuity, Incident and Service Request, Problem, Configuration and Change, and finally Release and Deployment (Fig. 2).

EOSC Governance

Cloud hub services could be provided in a coordinated fashion by multiple stakeholders, including research communities, research infrastructures and e-Infrastructures. In this case, a *federator* role needs to be established to ensure services are provided in an integrated way according to a single community-defined lightweight service integration and management framework that defines a corpus of policies (the so-called “rules of engagement”), processes and tools for aggregating demand and supply from local, national and regional facilities.

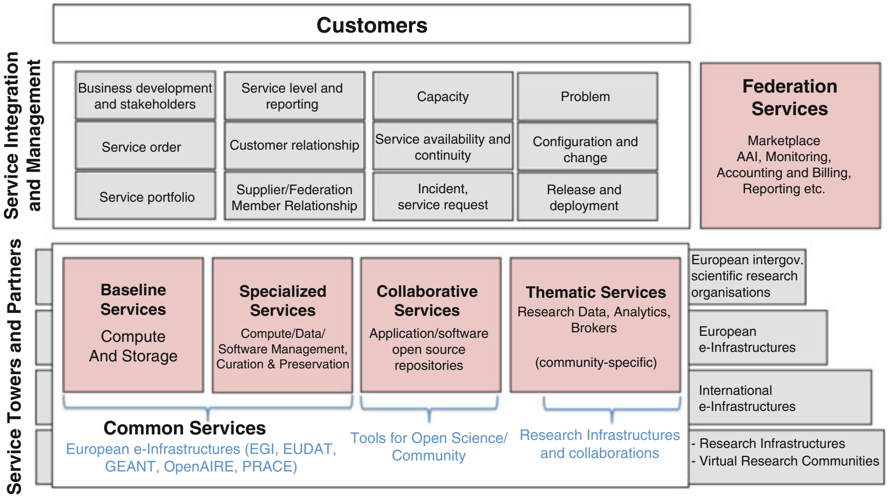
Customers
Service Integration and Management
Business development and stakeholders
Service level and reporting
Capacity
Problem
Service order
Customer relationship
Service availability and continuity
Configuration and change
Service portfolio
Supplier/Federation Member Relationship
Incident, service request
Release and deployment
Federation Services
Marketplace
AAI, Monitoring,
Accounting and Billing,
Reporting etc.
Service Towers and Partners
Baseline Services
Compute And Storage
Specialized Services
Compute/Data/ Software Management, Curation & Preservation
Collaborative Services
Application/software open source repositories
Thematic Services
Research Data, Analytics, Brokers
(community-specific)
Common Services
European e-Infrastructures (EGI, EUDAT, GEANT, OpenAIRE, PRACE)
Tools for Open Science/ Community
Research Infrastructures and collaborations
European intergov. scientific research organisations
European e-Infrastructures
International e-Infrastructures
- Research Infrastructures
- Virtual Research Communities

Fig. 2 The EOSC infrastructure includes a service catalogue (*bottom layer*) and a service integration and management services and activities (*top layer*) that ensure the delivery of services in a multi-supply environment. The federation services (*right*) support the service integration and management system, and enable the federation of distributed facilities. EOSC services (the "towers"), aggregated in cloud hubs and in the EOSC will be delivered by multiple suppliers: European intergovernmental organizations, European/international Research e-Infrastructures, and virtual research communities

The value proposition of the Open Science Commons is the combination of services and their integration layer, which altogether augments the existing data infrastructures and allows extraction of knowledge and the generation of innovation from research data. In such vision, open standards are considered as enablers of the Commons.

The realization of an EOSC must avoid duplication of provisioning of ICT services at national and European level and ensure efficient provisioning. National EOSC facilities are primarily financially supported by the European Member States. The role of the European Commission would be to ensure the persistency of the services that allow the national cloud hubs to operate as a federation, and to ensure the coordinated procurement, service provisioning and data brokering according to the requirements of the RIs. This would allow aggregation of demand across Europe, coordinated delivery and the development of economies of scale.

This would increase the coordination between Research Infrastructures, e-Infrastructures and data providers in matters concerning ICT provisioning. With European coordination, an economically efficient system of tools can be developed, which can accelerate the flourishing of multidisciplinary science, open science and a sustainable system of integrated Commons.

The EOSC can be organised as an integration of existing e-Infrastructures with overarching governance and common agreed services. The EOSC can be created based on both publicly funded and commercial providers as long as they are all based on open standards and remove the risks for artificial lock-in.

The role of national funding agencies is to ensure the sustainability of national cloud hubs, while European coordination is necessary to focus on supporting the service integration and management layer of the EOSC.

The development of a research data marketplace will promote the definition of governance involving data providers, archiving organizations, infrastructure providers, knowledge organizations, funding agencies, users and citizens. This EOSC governance will have to harmonize access, allocation of quotas, policies and acting as conflict resolution body. Besides being inclusive, the governance will need to reflect the federated nature of Europe and be inspired by the Commons principle.

EOSC and the e-Infrastructure Commons

The e-Infrastructure Reflection Group (e-IRG) (2017b) is a strategic body facilitating integration in the area of European e-Infrastructures and connected services, within and between member states, at the European level and globally. The mission of e-IRG is to support both coherent, innovative and strategic European e-Infrastructure policymaking and the development of convergent and sustainable e-Infrastructure services.

e-IRG has recently published a new version of its roadmap document. The e-IRG Roadmap 2016 is taking up the e-Infrastructure Commons concept, introduced in the previous version of the roadmap (2012) (e-Infrastructure Reflection Group 2012). The document intends to define a clear route on how to evolve the European e-Infrastructure system further. The new issue aims to turn the vision of the e-Infrastructure Commons, one of the pillars of the Open Science Commons, into reality by 2020 and has recommendations to all stakeholders to progress on the way towards the e-Infrastructure Commons. The key recommendations in the roadmap document encourage the research infrastructures and research communities to elaborate on and drive their e-infrastructure needs.

A concrete way towards the e-Infrastructure Commons, loosely integrating the different types of e-Infrastructures, is to use a marketplace with a proper governance including a representation of the users as a single point of access to all e-Infrastructure services and tools. The marketplace will act as a one stop-shop for EU researchers, i.e. a place where all e-Infrastructure services are accessible together, either directly or redirected elsewhere. The marketplace can make use of several technologies and services, such as cloud technologies, a searchable service catalogue and a common authentication/authorisation scheme. In this way, a standardised and single point of access to services will be achieved, without promoting monopolies, nor a single integrated provider. This has proven to be very difficult across different e-Infrastructure components. On the contrary, it will be open to new actors, encouraging cooperation, competition and innovation.

The e-IRG roadmap 2016 clearly identifies e-Infrastructures as the *seeds* to implement the EOSC and defines a clear roadmap towards 2020 to deal with the challenges described in the previous section.

The role of the e-Infrastructures to establish the EOSC is also described in the position paper *European Open Science Cloud for Research* written in collaboration by five leading European initiatives, EUDAT (2017), LIBER (2017), OpenAIRE (2017), EGI (2017a) and GÉANT (2017). In that paper, the relevance of the Open Science Commons for the EOSC has been underlined as *a key driver, not only of scientific progress, but also of economic and societal innovation*. The EOSC has been presented as a vehicle *to foster an open, collaborative platform for the management, analysis, sharing, reuse and preservation of research data on which innovative services can be developed and delivered*. Furthermore, the paper considers fundamental leveraging on the scientific infrastructures already available in Europe as outcome of decades of public investments. The EOSC could be developed by connecting the existing national and international infrastructures and services. Indeed, many of the resources and services needed for the Open Science Cloud already exist, but while technical challenges remain, *most of the barriers are ones of policy and concern funding, lack of interoperability, access policies and coordinated provisioning*. Then, the EOSC should address these issues and enhance both the service portfolio and the amount of available resources.

The EGI Blueprint

The key role of the e-Infrastructures on creating the EOSC has been clearly introduced in the previous section. Now, as example, key services, platforms and tools of one of the main European e-Infrastructure, EGI, will be presented highlighting how these could be the ground for implementing the Open Science Commons and, then, the EOSC. In particular, already available features that could facilitate the building of the Cloud Hub model for EOSC previously defined will be underlined.

EGI, advanced computing for research, is a federated e-Infrastructure set up to provide advanced computing services for research and innovation. The EGI e-infrastructure is primarily publicly funded and comprises over 300 data centres and cloud providers spread across Europe and worldwide. EGI offers a wide range of services for compute, storage, data and support (EGI 2017b) and provides access to over 700,000 logical CPUs and 500 PB of disk and tape storage. Its principles are based on the Open Science Commons and its mission is creating and delivering open solutions for science and research infrastructures by federating digital capabilities, resources and expertise between communities and across national boundaries.

The EGI architecture is organised in platforms (Fig. 3):

- Core Infrastructure Platform, to operate and manage a distributed infrastructure;
- Cloud Infrastructure Platform, to operate a federated cloud-based infrastructure;

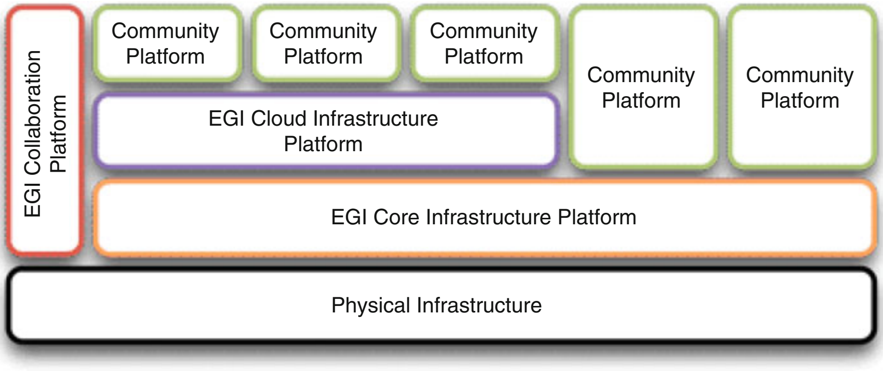

Fig. 3 EGI platform architecture

- Open Data platform, to provide easy access to large and distributed datasets;
- Collaboration Platform, for information exchange and community coordination, and
- 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.

In the remaining part of the section, the platforms that could provide functionalities useful to implement the EOSC are shortly introduced.

Core Infrastructure Platform

The **EGI Core Infrastructure** provides all the necessary operational tools and processes to operate and manage a large distributed infrastructure guaranteeing standard operation of heterogeneous infrastructures from multiple independent providers. This also includes:

- The **Authentication and Authorisation infrastructure** for homogeneous authentication and authorisation across the whole federation.
- A **Service registry** for configuration management of federated services.
- **Monitoring** tools, performing service availability monitoring and reporting of the distributed service end-points.
- **Accounting** for collecting, and displaying usage information.
- **Information discovery** about capabilities and services available in the federation.
- **Virtual Machine image catalogue** and distribution: allows researchers to share their virtual appliances for deployment in a cloud federation.

The federated environment of European e-Infrastructures, implemented in EGI through the Core Infrastructure platform, is a key enabler for distributed management and processing of big data and a fundamental baseline to implement the service integration and management system of the EOSC.

Collaborative Platform

It provides IT Infrastructure and services that facilitate collaboration between research communities. Its two main components are the Marketplace and the Application Database.

The Marketplace (EGI 2017c)

The marketplace has the ambition of becoming the platform where an ecosystem of EGI-related services, delivered by providers and partners, can be promoted, discovered, ordered, shared and accessed, including EGI offered services as well as discipline and community-specific tools and services enabled by EGI and/or provided by third parties under defined agreements.

The need of a Marketplace, making discoverable open research data and the related tools and knowledge, and acting as a one stop-shop for EU researchers, has been identified in the Service Hub model to build the EOSC and in the e-IRG roadmap 2016 introduced before. This should also act as a single point of access to all e-Infrastructure services and tools for all users. The EGI Marketplace could be seen as first test to implement such tool (to be extended both in term of features and coverage).

Application Database (AppDB) (EGI 2017d)

It is a tool that stores and provides information about:

- software solutions in the form of native software products and virtual appliances,
- the programmers and the scientists who are involved, and
- publications derived from the registered solutions.

Reusing software products registered in the AppDB, means that scientists and developers may find a solution that can be directly utilized on the infrastructure. In this way, scientists can spend less or even no time on developing and porting a software solution to the Distributed Computing Infrastructures (DCIs) and facilitate the reproducibility of experiments. AppDB, thus, aims to avoid duplication of effort across the DCI communities and to inspire scientists who are less familiar with DCI programming and usage. The service is open to every scientist interested in publishing and therefore sharing their software solution.

The AppDB can be considered as an example of a service providing a library of community-specific applications and tools that could enable the repeatability, and therefore the incremental improvement of research practices and processes within and across research teams.

The EGI Federated Cloud

EGI launched the production phase of a cloud federation to serve research communities in May 2014, the EGI Federated Cloud (EGI 2017e). It integrates community, private and/or public clouds into a scalable computing platform for data and/or compute-driven applications and services.

Its architecture is based on the concept of an abstract Cloud Management Framework (CMF) that supports a set of cloud interfaces to communities. Each resource centre of the infrastructure operates an instance of this CMF according to its own technology preferences and integrates it with the federation by interacting with the EGI Core Infrastructure platform.

This integration is performed by using public interfaces of the supported CMFs, thus minimising the impact on site operations. Providers are organised into realms exposing homogeneous interfaces and grouping resources dedicated to serve specific communities and/or platforms.

The EGI Federated Cloud is based on a hybrid model where private, community and public clouds can be integrated and already offers some of the facilities that a Service Hub should provide such as the virtualisation and the easy share and reuse of tools (Fig. 4).

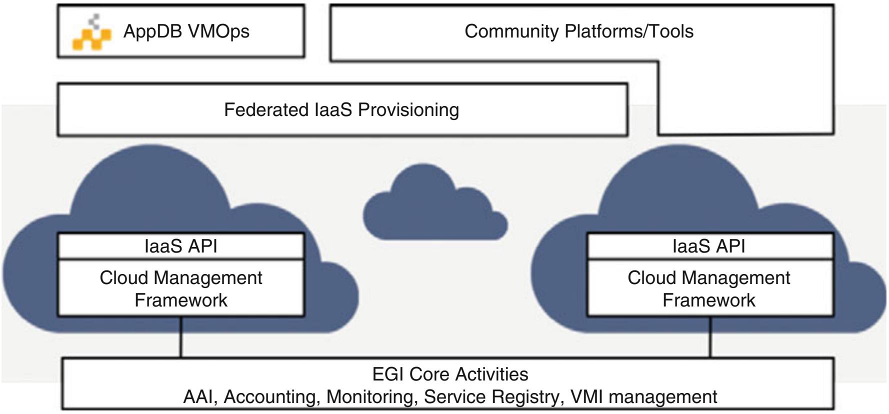
AppDB VMOps
Community Platforms/Tools
Federated IaaS Provisioning
IaaS API
Cloud Management Framework
IaaS API
Cloud Management Framework
EGI Core Activities
AAI, Accounting, Monitoring, Service Registry, VMI management

Fig. 4 EGI Federated Cloud architecture: each resource centre of the infrastructure operates an instance of a CMF according to its own technology preferences and integrates it with the federation by interacting with the EGI Core Infrastructure platform. Providers are organised in realms exposing homogeneous interfaces (Federated IaaS provisioning). Community platforms can exploit resources from one or more realms through such interfaces. The AppDB VMOps enables an automatic deployment of virtual appliances on all the resource centres supporting a specific community

The Data Hub and the Open Data Platform

The Data Hub provides easy and efficient access to large-scale datasets enabling sharing, discovering, and processing of data federated from different sources. The service offers a virtual access to files distributed across different types of storage and geographically distributed providers through homogenous and standard based interfaces (POSIX, CDMI, etc.).

The technology behind the Data Hub service is the Open Data Platform, implemented in the EGI-Engage project (2017), aiming at overcoming the technical barriers that are still faced to federated data on cloud across multiple storage providers. Its design emerged from the analysis of several user communities' requirements, including some of the major Research Infrastructures on the ESFRI roadmap, with focus on open data management. It allows the integration of various data repositories available in a distributed infrastructure, offering the capability to make data open, and link them to key open data catalogues following respective guidelines, such as the OpenAIRE (2017) open access infrastructure. The core enabling technology of ODP is Onedata (CYFRONET, n.d.), a data management solution that allows a seamless and optimised access to data spread over a distributed infrastructure (see Fig. 5).

The Data Hub service can help on implementing the Cloud Hub based architecture for EOSC dealing with the offering of scalable access to and

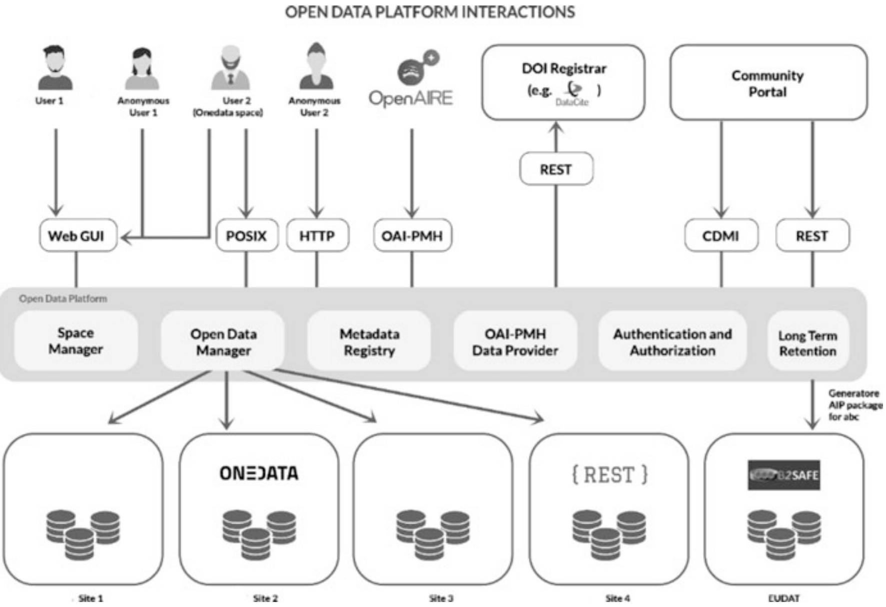

Fig. 5 Open Data Platform architecture, showing its modular backend at the *bottom* of the picture, and how it integrates with different data management systems in its distributed configuration at different data centres ("sites")

analysis of research data for reuse. Indeed, it can not only make data and metadata findable through its catalogues and metadata, but can also provide the distributed data mirroring that, integrated with the IaaS features offered by the Federated Cloud, allows an efficient and scalable processing of the datasets reducing or removing the needs of data movements.

The future evolution of the Data Hub foresees the introduction of a smart caching mechanism that could cleverly move portions of dataset before the users' request dramatically decreasing time to access. These caching mechanisms will take into account factors like the number and type of data analysis applications that are currently running in the infrastructure, the popularity of datasets and the catalogue of recently generated data. Such factors will drive the geo-replication of the data, implementing a network of distributed data hubs that minimizes the need of data transfers, by coupling storage and computing resources and caching heuristics.

EO Data Exploitation via the e-Infrastructures and EOSC

The new generation of Earth Observation (EO) satellites from the Sentinel missions that ESA developed for the Copernicus programme, is generating large amounts of data that are not easily integrated into processing chains outside the Copernicus ground segment. Very often, public and private institutions aiming to deliver end-user services based on EO data do not possess the computing power, the storage capacity or the software technology to cope with these new data flows. Handling an increasing volume of EO data is one of the main challenges for the community and hybrid clouds, coupled with big data management solutions and applications, are seen as a potentially efficient solution.

E-Infrastructures can improve the discovery, retrieval and processing capabilities of Copernicus data through their capabilities for the management of big data. Indeed, they offer virtualised access to geographically distributed data and the computing necessary to manipulate it, and to manage large volumes of different types of data. These solutions can be reused by any scientific discipline facing the challenge of integrating large datasets including earth observation. Furthermore, e-Infrastructures also provide users with high-throughput access to data sets, without the necessity of prestaging data on local storage, thus enabling full support for hybrid cloud scenarios. This scenario gives users the freedom to mix computational and data storage resources from various infrastructure providers, including their own private ones.

In addition, well-established e-Infrastructure services complete the set of reusable components that could accelerate the development of exploitation platforms for satellite data solving common problems, such as the user authentication and authorisation, the monitoring, the accounting, etc.

This already promising scenario needs enhancements to overcome the barriers previously described and, the implementation of the EOSC vision introduced in this chapter draws a clear roadmap to enhance the e-Infrastructure services in the next years. However, the current maturity of the e-Infrastructure services already allows their use for the EO data exploitation. This is detailed in the next section.

Infrastructure Services for EO Data Exploitation

The design of appropriate solutions for managing and exploiting large amount of EO Datasets needs expertise from different sources, such as:

- Data consumers, to help defining and designing real added-value services to be integrated with the existing EO Exploitation Platforms;
- ICT experts and platforms operators, with advanced knowledge on EO systems, to develop and provide hosting platforms for these services and to offer general solutions;
- E-Infrastructures, to supply the computing and storage resources needed for data access and exploitation and provide the tools to manage the datasets in a distributed environment

The joint expertise from these three sectors can allow the creation of an integrated environment for fast development and prototyping of services for exploitation platforms and scientific applications.

E-Infrastructures are the foundation of such an environment and could become an attraction pole to merge these sectors and enable such potential. By bringing these individual groups together, the fostering of new knowledge through the sharing of specialised expertise will be tremendously accelerated. This will ultimately improve innovation capacity leading to fast development and prototyping of services able to deal with so extensive sources of information through the integration of e-Infrastructure services, EO exploitation platforms and scientific applications. E-Infrastructure services could also evolve according to the specific needs identified by both data consumers and EO platform operators. This will facilitate the exploitation of EO data with the final aim to facilitate the creation of applications to address societal challenges, enabling policymakers and authorities, including environmental agencies to develop long-term strategies as well as react efficiently to sudden crisis situations.

Furthermore, the adoption of the e-Infrastructures for EO Data exploitation will facilitate the market uptake of the satellite data, including those coming from the Copernicus programme, and will contribute to the creation of a European solution for exploiting Earth Observation data, fostering EO services within the public and private sectors. About the latter, widely diverse and small companies throughout Europe could benefit from the easier data access to develop new products and services.

The following image shows the role of the e-Infrastructure services for the EO Data exploitation in the European Scenario (Fig. 6).

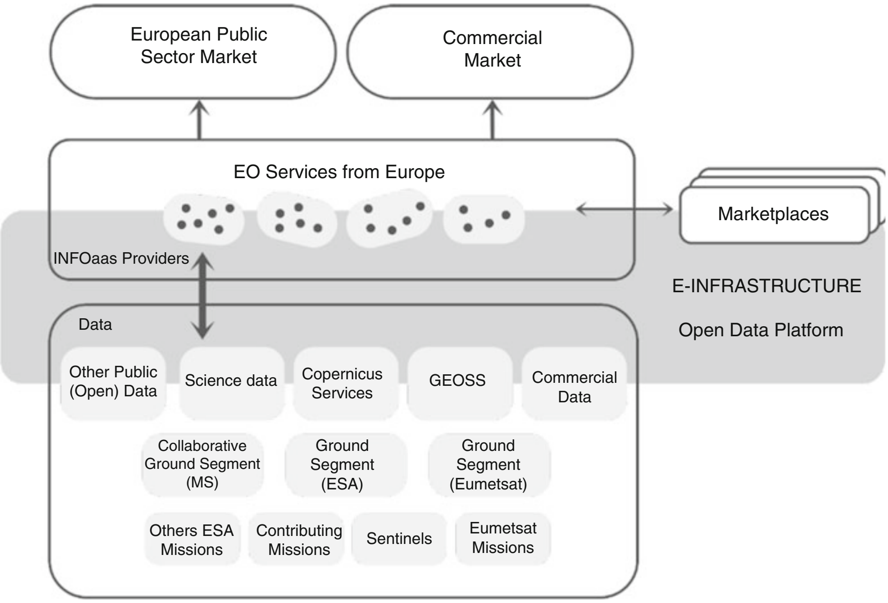

Fig. 6 Role of European e-Infrastructure services for EO data exploitation

An Example: e-Infrastructure Services to Implement the ESA Generic Exploitation Platform Open Architecture

As example, this section shows how the e-Infrastructure services could facilitate the implementation of a EO exploitation platform based on the generic Exploitation Platform Open Architecture (ESA 2016) defined by ESA with the aim to harmonize the architecture of such platforms (in particular the ESA Thematic Exploitation Platforms—TEPs) and make the adoption of shared solutions easier.

ESA Exploitation Platform Open Architecture

ESA defined the concept of a General Exploitation Platform Architecture (Fig. 7) to harmonize the different Exploitation Platforms and maximize the reuse of technology and components, thus reducing the cost required to develop, maintain and operate them. This architecture includes a set of common components, to be reused by the single platforms and tailored to the particular platform needs.

- **User access portal**: it is the interface for the final users and the system operators. It includes services used by the other components, such as authentication, accounting, monitoring and collaboration tools and implements a *single access portal*, *collaboration tools*, *documentation and support tools*, *services mar-*

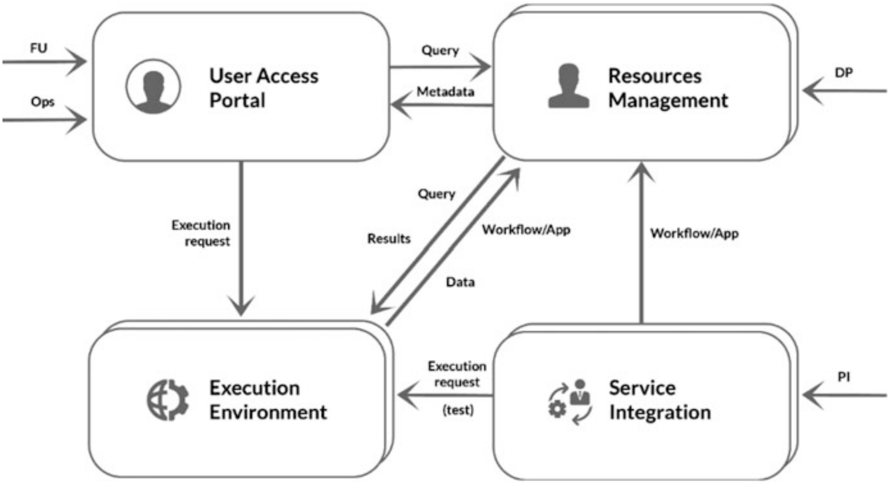

Fig. 7 ESA Exploitation Platform Open Architecture (ESA 2016)

ketplace and *operator interface* functionalities. It interacts with the Resource Management component for getting information about the data, application and workflow resources required for a given processing service and with the Execution Environment for executing processing services.

- **Resource management**: handles the resources available in the platform and implements *data discovery*, *data management* and *processing services management* functionalities. It provides the User Access Portal and the Execution environment components with interfaces to:
  - Perform query to retrieve data products, data results, application or workflow packages and the related metadata,
  - Publish new data or applications provided by a data provider or a scientific user.
- **Service Integration**: provides the programmatic interface with a framework to integrate applications, algorithms and/or software into the platform as a new processing service. It implements the required *development and integration system* functionality. It can send requests to execute an application to the Execution Environment for testing purposes.
- **Execution Environment**: provides the platform with an environment to run processing services. It implements both the *on-demand processing* and *massive processing* functionalities. It receives the application and data required to run the processing from the Resource Management component. It can also perform queries to the Resource Management component to identify the resources required by the processing that are not directly specified in the execution request.

Infrastructure Services to Implement the ESA Exploitation Platform Open Architecture

The following table shows the relations between each of the ESA Exploitation Platform components and the EGI services highlighting the features of such services that support the implementation of the TEPs functionalities.

TEPs macro-components	EGI services	Functionalities
User access portal	EGI core platform services: AAI infrastructure, accounting and monitoring.	Single access portal: AAI infrastructure Services: accounting and monitoring
Resource management	EGI Data Hub & Open Data platform	Data discovery: query to the Data Hub metadata catalogue Data management: registration and setup of EO data products into the Data Hub metadata catalogue and related pre-staging through the smart cache
Service Integration	N.A.	N.A.
Execution Environment	EGI Federated Cloud: OCCI and Open Stack standard interfaces Open Data platform	On-demand and massive processing: (1) standard OCCI and Open Stack interfaces to access a public cloud to run to run workflows, apps or bulk processing, (2) list of computing resources close to the pre-staged data

This mapping demonstrates that several TEP functionalities, described in the ESA Exploitation Platform open architecture, could already benefit from the EGI services. EGI already validated this statement having supported the integration of the Geohazard Exploitation Platform (GEP) (2017) within its FedCloud in the last years. Such result could be advertised to other ESA TEPs that could also benefit from the EGI services. This fits very well with the ESA vision of reusing technology and components to reduce the cost. The integration of the Hydrology TEP is also already planned.

An Example: Integration of the Geohazard Exploitation Platform Within the EGI Infrastructure

The Geohazard Exploitation Platform has been developed by Terradue on behalf of ESA. Terradue has integrated such platform with the EGI infrastructure developing the needed extensions in their workflows to use the EGI Federated Cloud interfaces and tested in several resource centres belonging to the EGI infrastructures.

Terradue has extended their workflows to be compatible with the following EGI components:

- AAI;
- Federated Cloud: Implemented an OCCI driver;
- AppDB: registering the virtual machine images that compose the platform.

The workflows have been successfully tested together with EGI Operations and the resource centres providing resources.

Furthermore, Terradue and EGI have agreed to two VO Service Level Agreements (EGI 2016) for the provisioning of EGI resource to support both the Geohazard and Hydrology TEPs. Thanks to such agreements, the GEP will be able to access European cloud compute resources. In total, the six data centres offer more than **360 virtual CPU cores**, **800 GB of memory** and **10 TB of storage**. The GEP just started to use the EGI Federated Cloud resource in production.

Conclusions

Nowadays, the implementation of the European Research Area (ERA), as depicted by the European Council, cannot be considered fully achieved. The realization of an open and integrated environment for cross-border seamless access to advanced digital resources, services and capabilities fostering the reuse of research data and services, is being accelerated by the European Open Science Cloud initiative of the European Commission. Open Science is seen as the natural paradigm to foster and drive such developments. It can remove the barrier between adjacent communities, enable multidisciplinary collaborations, reinforce the need of knowledge sharing and allow non-discriminatory access.

This chapter illustrated the benefits of a Commons approach to Open Science, and in particular the advantages of the Commons for the implementation of the European Open Science Cloud infrastructure and governance.

We presented a possible approach to implement the EOSC via the Open Science Commons. In our approach, the EOSC architecture will rely on a federation of cloud hubs, where a cloud hub provides data, services and capabilities in a standard and interoperable manner. The hubs adopt the cloud service provisioning paradigm to facilitate sharing, reusing and the combined offer of data and tools through the virtualisation. In addition, the federation of hubs provides a multilayer organizational structure that meets European policies, regulations, restrictions and business models and allows the creation of a community-network able to bring together the different kinds of expertise available in each hub. The multi-supply environment of the EOSC can be regulated by a corpus of policies, processes and tools defining the EOSC service integration and management system owned, maintained and developed by the EOSC according to the Commons governance model. The EOSC cloud hub services will be contributed by multiple stakeholders: data providers, European Research Infrastructures, e-Infrastructures, research collaborations and local, regional and national facilities.

The exploitation of Earth Observation data will directly benefit from the EOSC and the adoption of the Open Science Commons, leveraging technologies, services and resources provided in the context of existing European e-Infrastructures. EOSC and e-Infrastructures can become an attraction pole to design and develop appropriate solutions for managing and exploiting large amount of EO datasets. This will allow the creation of an integrated environment for fast development, prototyping and provisioning of services for exploitation platforms and scientific applications.

References

CYFRONET. One Data web site. https://www.onedata.org/

EGI (2016) VO SLA Terradue and VO OLAs Terradue. https://documents.egi.eu/document/2763

EGI (2017a). https://www.egi.eu/

EGI (2017b) EGI Service catalogue. https://www.egi.eu/services/

EGI (2017c) EGI Marketplace. http://marketplace.egi.eu/

EGI (2017d) EGI AppDB. http://appdb.egi.eu/

EGI (2017e) EGI Federated Cloud. https://www.egi.eu/federation/egi-federated-cloud/

EGI-Engage project (2017). http://cordis.europa.eu/project/rcn/194937_en.html

ESA (2016) ESA Thematic Exploitation Platforms architecture. http://go.egi.eu/EP-OpenArchitecture

EUDAT (2017). https://eudat.eu/

European Commission (2012) Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, A Reinforced European Research Area Partnership for Excellence and Growth, COM (2012) 392 final. http://ec.europa.eu/euraxess/pdf/research_policies/era-communication_en.pdf. Accessed 15 Feb 2017

European Commission (2014a) Communication from the Commission to the Council and the European Parliament, European Research Area, Progress Report 2014. COM (2014) 575 final. http://ec.europa.eu/research/era/pdf/era_progress_report2014/era_progress_report_2014_communication.pdf. Accessed 15 Feb 2017

European Commission (2014b) Conclusions on progress in the European Research Area, Competitiveness Council meeting, Brussels, Feb 2014. http://www.consilium.europa.eu/uedocs/cms_data/docs/pressdata/en/intm/141120.pdf

European Commission (2015) Open Science at the Competitiveness Council. http://ec.europa.eu/digital-agenda/en/news/open-science-competitiveness-council-28-29-may-2015

European Commission (2016) First report of High Level Expert Group on the EOSC. https://ec.europa.eu/digital-single-market/en/news/first-report-high-level-expert-group-european-open-science-cloud

European Commission (2017) European Charter for Access to Research Infrastructures. https://ec.europa.eu/research/infrastructures/index_en.cfm?pg=access

European Commission. Science 2.0: science in transition. http://ec.europa.eu/research/consultations/science-2.0/background.pdf

European Strategy Forum on Research Infrastructures (2017) ESFRI web site. http://www.esfri.eu/. Accessed 15 Feb 2017

Benedikt Fecher, Sascha Friesike (2014) Open science: one term, five schools of thought. http://book.openingscience.org/basics_background/open_science_one_term_five_schools_of_thought.html

Frischmann BM (2013) Infrastructure: the social value of shared resources. Oxford, Oxford University Press

GEANT (2017). http://www.geant.org/

Geohazard Exploitation Platform (2017). https://geohazards-tep.eo.esa.int/

Horizon 2020 consultation report (n.d.) Open Infrastructures for Open Science. http://cordis.europa.eu/fp7/ict/ e-Infrastructure/docs/open-infrastructure-for-open-science.pdf. Accessed 15 Feb 2017

e-Infrastructure Reflection Group (2012) e-IRG roadmap 2012. http://e-irg.eu/documents/10920/12353/e-irg_roadmap_2012-final.pdf/4c5cab85-dca4-49b7-b7f0-66c2bdd057af

e-Infrastructure Reflection Group (2017a) e-Infrastructure Commons. http://knowledgebase.e-irg.eu/e-infrastructure-commons

e-Infrastructure Reflection Group (2017b). http://e-irg.eu/

LIBER (2017). http://libereurope.eu/

OpenAIRE (2017). https://www.openaire.eu/

Mark D. Wilkinson et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. http://www.nature.com/articles/sdata201618