IMPLEMENTING THE OPEN SCIENCE CLOUD THROUGH THE DATA COMMONS

EGI Internal note on the Open Science Cloud (v1.0), 20 August 2015

Nowadays, research practice 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, are foundational. Along with this, 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 benefits of research communities and the wider economy and society. This document presents a vision for establishing the Data Commons through an Open Science Cloud, one of the pillars of the Open Science Commons, as a network of service hubs.

The document analyses the current status and blockers for implementing an Open Science Cloud and discusses how such a cloud could strategically advance its competitiveness by providing research Data and community-specific tools as services through a platform that supports the participatory principle of **Open Science**.

We call these the "**Data Commons**", i.e. the possibility to **share data, the processing services and the applications, virtual laboratories and tools**, relying on existing federated data and storage facilities. By realizing the Data Commons, data, computing needed process it, and the tools are offered as a single research system.

INTRODUCTION

Thanks to the coordination role of the EC together with the Member States, the definition of research roadmaps has greatly advanced Europe in developing international research communities and fostering transnational and virtual access to research instruments and services.

Given the digitalisation and internationalisation of science, research communities depend on the availability of tools, storage and computing infrastructures to deposit the raw data produced, to further process and share them, and ultimately to enable a scalable access of existing curated data worldwide. These communities play the key role of "data factories", "data curators" and "data integrators" of existing research repositories to extract new knowledge.

In parallel to this, Europe has developed world-leading e-Infrastructures that federate ICT capabilities providing high throughput computing, high performance computing, storage, and generic cloud capabilities.

Unfortunately, to date all these **Commons** even if mutually dependent for their sustainability, are mostly developed in the context of different initiatives with a risk of duplication and inefficient use of public funding. Furthermore, open science is not coherently embraced as an overarching approach that would lead to a greater social value. The Open Science Commons vision address these issues as a new approach to digital research, tackling policy challenges and embracing open science as a new paradigm for knowledge creation and collaboration (http://go.egi.eu/osc).

Open Science Commons Vision

Researchers from all disciplines have easy, integrated and open access to the advanced digital services, scientific instruments, data, knowledge and expertise they need to collaborate to achieve excellence in science, research and innovation.

WHAT ARE THE DATA COMMONS?

The Data Commons are realized through an ensemble of open research data, tools, applications, virtual laboratories and the related knowhow that the researchers can freely share in a common space, which federates the data and hosts the computing and tools needed to extract knowledge from the data.

The Data Commons can be implemented as a federation of Cloud Hubs (in Europe and beyond) that provide all these capabilities in a federated, integrated way: a virtual space providing data, tools and processing capabilities, with the Hubs interconnected by a mesh of high-bandwidth links. The network of Hubs realizes the Open Science Cloud.

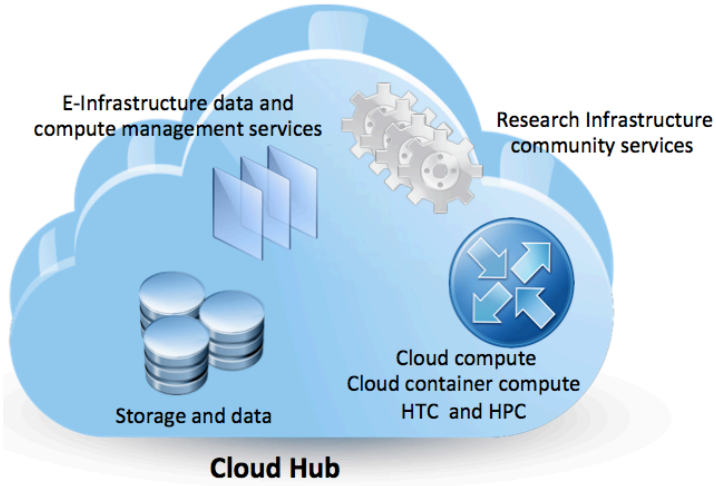

Figure 1. Functionalities delivered by the Cloud Hub

Why a federated cloud?

Cloud is a service provisioning paradigm that can support hosting for both data and software tools.

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, in addition the federation approach allows the implementation of hybrid models where private, community and public clouds can be integrated.

Why is the federated hub-based model a good approach?

A federation of hubs provides an organizational structure that meets European policies, regulations, restrictions and business models, which in some cases do not make the permanent relocation of data into centralized science-domain specific repositories possible, and/or into generic repositories (that integrate data and tools from multiple domains). Also, expertise about how to use data and tools from a specific research domain typically accumulates within the specific research community so this community-network is an ideal incubator for a hub and contributes to the implementation of hubs.

Within the federation data providers should always retain control to their data.

The federated approach allows the implementation of a multi-level governance model where different governing bodies of the Commons can be coexist and be integrated.

Who do we expect to operate the service offered by the hubs?

Cloud Hub services could provided in a coordinated fashion by multiple stakeholders, including research communities, research infrastructures and e-Infrastructures. A **federator** role needs to be established to ensure services are provided in a integrated way according to federated service management best practices and standards. A single interface needs to be provided to end-users.

Agreements with the data owners will be established, so that the Cloud Hubs can provide premium controlled access to quality data.

The Cloud Hubs do not duplicate the services of the institutional and disciplinary repositories, but rather replicate these data in an environment that can enrich the data itself with additional added value services and can provide scalable access where necessary by co-locating computing and data.

The value proposition of the Data Commons is the combination of services that augment the existing data infrastructures and allow extraction of knowledge and the generation of innovation from research data.

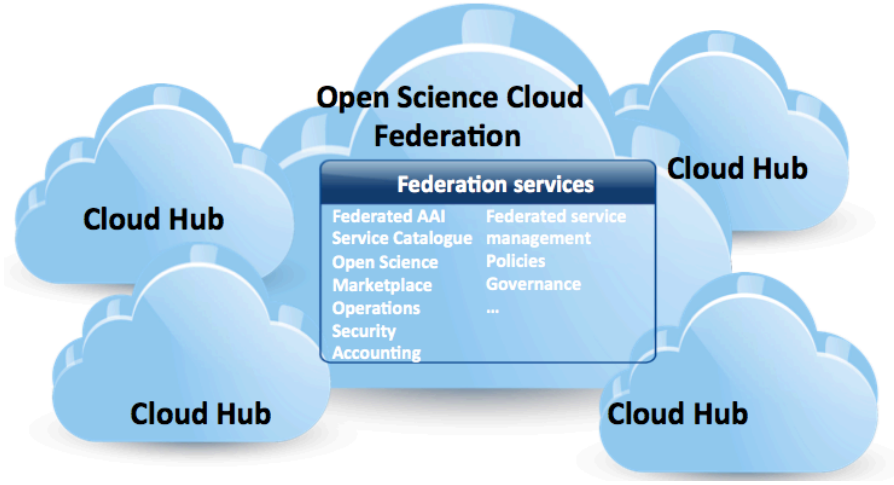

Figure 2. The Data Commons, based on a cloud federation of Cloud Hubs offering premium access to research data, tools and applications, and of community platforms to increasingly enrich this environment with new digital objects through a open, participatory model

How do the Data Commons relate to other Commons?

The Data Commons depend on existing Commons and augment them with new capabilities and a new way of sharing them.

- The data archiving organizations are the primary sources of the data replicated in the Data Commons. Agreements need to be established to ensure protection of Intellectual Property, and data access needs to be accounted for to ensure that data providers are properly acknowledged, that data access can be reported to funding agencies, and to foster a culture of open data sharing. The Communications Commons (GEANT) provide high performance connectivity between the Hubs.

- The **AAI services provided by e-Infrastructures** offer federated IdP, authentication and authorization for access to the Hubs.
- The **e-Infrastructure Commons** provide a federated platform for high throughput, high performance and cloud computing (EGI, EUDAT, PRACE…) that physically host the Hubs (provide the fabric layer).
- **Research Infrastructures** provide discipline-specific data repositories and tools for data manipulation (see Figure 2).
- The **Knowledge Commons** through a network of competence centres provide training, education, knowledge transfer programmes.

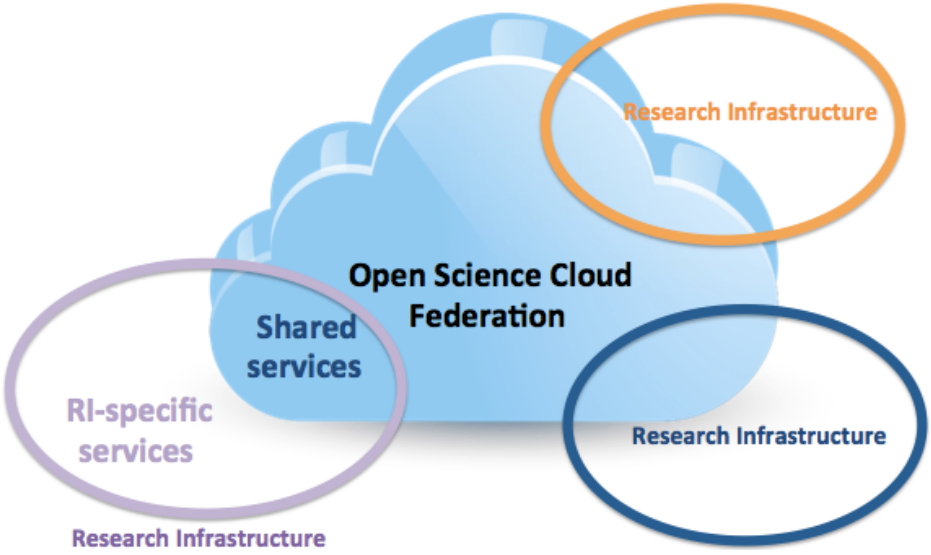

Figure 3. Complementarity of services offered by Research Infrastructures and e-Infrastructures and the Open Science Cloud through a system of Commons.

In other words, the Data Commons are part of a system of Commons that support altogether Open Science. In order to enable this, **interoperability** between the Commons is necessary to ensure portability of data and applications. **Open standards** are enablers of the Data Commons.

WHAT ARE THE CHALLENGES TO BE ADDRESSED?

- **CHALLENGE 1. Realizing a federated approach to research data**

The Open Science Cloud could provide a Marketplace making open research data, the related tools and knowledge discoverable. The marketplace would federate existing research data sets that are provided by archiving organizations that can ensure compliance to a set of quality standards defined by the marketplace. The Open Science 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 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.

The development of a Research Data Market and its business functions requires the development of a governance model that ensures a primary role of data archiving organizations as data suppliers. The marketplace would foster the reuse of research data.

Figure 4. Functions of the Open Science Marketplace

- **CHALLENGE 2. Offering of scalable access to and analysis of research data for reuse**

Making data findable is not sufficient. Also, researchers should not need to download locally large datasets before executing their workflows as this would make access to data less efficient and time consuming because of the lack of a shared approach. Research data must be made easy to access and reuse, this means scalable access – especially in case of big data that cannot be efficiently downloaded locally. **The Open Science Cloud could provide distributed data mirroring and caching capabilities based on federated IaaS cloud storage,** where research data can be temporarily stored for scalable access in agreement with the data providers, and provided with integrated computing platforms.

This service is not a duplication of existing data infrastructures, but rather provides the capability of 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.

How could premium access be offered? By implementing a **federation of large Cloud Hubs** connected by a broadband network infrastructure. 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.

- **CHALLENGE 3. Integrating (shared) tools and applications**

Knowledge cannot be extracted from data without the availability of specialized tools and applications (e.g. text mining). The Open Science Cloud 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 Open Science Cloud should provide PaaS and SaaS services that are community-specific and that could be dynamically deployed with a focus on the long tail of science. These services could be provided in the form of managed services by the Research Infrastructures.

- **CHALLENGE 4. Provide services for depositing data for resource-bound users**

Through virtual access the Open Science Cloud will federate infrastructures to provide services for the long tail of science that cannot benefit from these services at institutional and/or national level, but supports open research data.

- **CHALLENGE 5. Achieving integrated e-infrastructures**

The development of a Open Science Cloud would avoid duplication of provisioning of ICT services at national and European level. The Open Science Cloud should be developed as a federation of national Data Cloud Hubs, financially support by the Member States. The role of the EC 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 RIs, e-Infrastructures and data providers in matters concerning ICT provisioning. With EC coordination, Europe will develop an economically efficient system of tools that will accelerate the development of multidisciplinary science, open science and a sustainable system of integrated Commons.

The Open Science Cloud can be organised as an integration of existing e-infrastructures with an overarching governance and common agreed services.

The Open Science Cloud 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 existence of the national Cloud Hubs while the EC focuses on supporting the federating services.

- **CHALLENGE 6. Repeatability of digital research processes**

By increasingly sharing models and modeling 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.

- **CHALLENGE 6. Developing a commons-oriented governance**

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 Open Science Cloud 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.

- **CHALLENGE 7. Intellectual property rights and data regulations**

Regulations need to be adjusted to allow the "free movement of research data" in Europe and beyond to be shared in the Cloud Hubs, and to foster its reuse by the private sector, to generate revenues from it, and ensure the long term sustainability of quality, curated, preserved data.

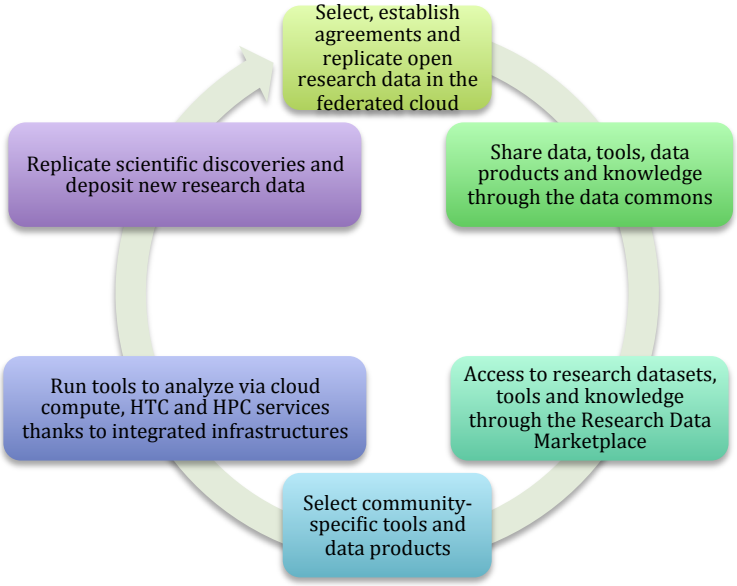

Figure 5. The Open Science Cloud can realize the data commons by simplifying the sharing of data, tools and knowledge in the context of a federated governance, shared policies and data regulations.

How Existing Initiatives Address These Challenges

EGI launched the production phase of a cloud federation to serve research communities in May 2014. Since then a tremendous number of use cases has emerged from the long tail of science, Research Infrastructures, international research collaborations and flagship EC initiatives requiring scalable cloud access to (big) open data from multiple data providers. To face these needs and host the large datasets major infrastructure investments are needed.

The open data platform on cloud being implemented in EGI-Engage aims at overcoming the technical barriers that are still faced to federated data on cloud across multiple storage providers. The EGI-Engage Competence Centres involving some of the major Research Infrastructures on the ESFRI roadmap are providing cloud requirements for a cloud federation, and will help developing the data commons. Specifically, EGI is collaborating with the life science and the astronomy and astrophysics communities to prototype the concept of data commons on the cloud federation of EGI. The EGI Application Database (http://appdb.egi.eu/) is a community platform that allows researchers to share their virtual appliances for deployment in a cloud federation.

In a recent article that appeared on Nature in July 2015 named "Data analysis: Create a cloud commons", Lincoln D. Stein et alt. encouraged major funding agencies to "ensure that large biological data sets are stored in cloud services to enable easy access and fast analysis"[1]. In addition, the "US National Cancer Institute has several pilot projects exploring the practicalities of sharing and analysing genomic data on clouds" and "the NIH and other funding agencies are already discussing a variety of 'biomedical commons' concepts".

The data commons approach is already being implemented at a national level by various infrastructures like CANFAR for astronomical data, NECTAR in Australia and the Open Science Data Cloud in the US.

[This paragraph will be further extended with information about other initiatives active or interested in developing the Data Commons]

CAN THE DATA COMMONS BE GLOBAL?

Research has no boundaries and research collaborations require integration of data repositories that are distributed worldwide. A global Data Commons can be realized by connecting Cloud Hubs through **international data peering points** that allow exchange of data across domains. The global Data Commons can be implementing by federating Data Commons initiatives worldwide. Examples of Data Commons are the NSF funded project Open Science Data Cloud and NeCTAR, the Australian research cloud. The challenge is about linking the existing commons together and broadening the spectrum of disciplines served.

DISCLAIMER. The present note is based on the EGI experience gathered in federating e-Infrastructures worldwide and in analysing requirements of Virtual Research Communities and more than eight Research Infrastructures on the ESFRI roadmap for the delivery of cloud-based services across different disciplines. The present paper only reflects of the views of the authors and aims at sharing ideas for the development of a European vision of the Open Science Cloud.

[1] http://www.nature.com/news/data-analysis-create-a-cloud-commons-1.17916