The Federated Scientific Data Hub

Prepared by IT department on behalf of CERN
edited by Bob Jones
15 February 2017

Executive summary

The paper proposes to establish a *federated scientific data hub* to deliver trusted data-intensive services in a federated environment where data complies with the "FAIR" principles (Findable, Accessible, Interoperable, Re-usable). It will leverage existing services across Member States and disciplinary, social and geographical borders.

The *data hub* will ensure the preservation of scientific data, together with the software, documentation and computing environment needed to process, (re-)analyse or otherwise (re-)use the data for the most advanced user communities that are willing to engage and contribute. Target data volumes in the exabyte range are envisaged with data management plans spanning at least one decade.

The *data hub* will offer production quality data-intensive services drawn from both public and private sources and presented through a common service catalogue. The services must conform to a minimum set of technical, legal and security requirements that ensure their suitability for inclusion. An initial set of interoperable services will include data management, compute (including HPC) and 'Human' services such as consultancy, s/w development, training etc. A combined service catalogue, service credit scheme and pay-for-usage model will make it easy to grow the *data hub* to include new funding streams, users and service providers.

The funding model will mobilise public and private sector funding with a mix of capital and operational budgets, involvement of commercial service providers as well as public-sector e-infrastructure operators and a broad spectrum of stakeholders. Separate models for capital investment, operational charges and charging structures for scholarly publishing will be combined to cover the whole lifecycle of the platform. Funding schemes that vary from country to country and between communities need to be carefully assessed and reflected in the selected models.

A multi-stakeholder governance model will be established to collaborate with and not replace the existing governance models for research infrastructures, scientific user communities as well as the e-infrastructure projects and initiatives that contribute services.

Contents

Introduction

Scientific excellence is the driving principle of the ESFRI Research Infrastructures. Ensuring the data produced by research infrastructures can be fully exploited to maximise their scientific output, impact on society and the economy is a growing challenge that the European Open Science Cloud (EOSC) and the underlying European Data Infrastructure (EDI) are foreseen to address.

The report of the Commission High Level Expert Group on the European Open Science Cloud[1] states that the EOSC aims to accelerate and support the current transition to more effective Open Science and Open Innovation in the Digital Single Market.

In the same report, Commissioner Carlos Moedas highlighted that "*The European Commission 'European Cloud initiative', issued in April 2016, set a very ambitious vision for the European Open Science Cloud; it drew a clear roadmap and set concrete commitments for the Commission to make this vision a reality by 2020*"

The Director-General of Research and Innovation, Robert-Jan Smits, has highlighted the need for a fast implementation mechanism with advanced research communities ready to take a leading role in the establishment of the EOSC.

This paper confirms the willingness of Europe's leading research organisations to engage and establish a working *federated scientific data hub*. This ambitious undertaking will deliver an operational, end-to-end example of Open Science and Open Innovation that can be expanded to encompass more research infrastructures, users and Member States.

[1] Realising the European Open Science Cloud, 2016, doi:10.2777/940154

Objectives

The objective for this *federated scientific data hub* is to provide confidence that a future full-scale EOSC together with the EDI will promote open science, be fit for purpose, inclusive, sustainable and fully inter-operable with existing systems and installed capacity. Implementing the vision of Open Innovation, Open Science and Open to the World[2] underpinning the EOSC requires:

- A sustainable long-term funding model which recognises the importance of stimulating innovation
- Understanding the focus, scale and distribution of investments needed by the stakeholders to support a wide range of sustainable science activities through a shared platform
- Recognition that the full value of scientific data will be realised by serving the interest of industries capable of leveraging that data.

The *data hub* will be rapidly established to demonstrate these objectives can be achieved by aggregating existing services from a variety of sources and making them seamlessly accessible to end-users. The success of the *data hub* will be measured in terms of the engagement of stakeholders:

- Number of research infrastructures that use the *data hub* to provide access to the data they produce
- Quality and range of services that are offered to end-users
- Scale and diversity of user communities that make use of the services
- An increase in the adoption of open science policies by the stakeholders
- An increase in the sharing and re-use of research data
- Ability of the funding models to support the services beyond the set-up phase

The *data hub* will deliver trusted data-intensive services in a federated environment where data complies with the "FAIR" principles (Findable, Accessible, Interoperable, Re-usable) as part of a hybrid cloud that builds on the existing investments made in the public and private sectors. The *data hub* will ensure the preservation of scientific data, together with the software, documentation and computing environment needed to process, (re-)analyse or otherwise (re-)use the data for the most advanced user communities that are willing to engage and contribute. Target data volumes in the exabyte[3] range are envisaged with data management plans spanning at least one decade.

The *data hub* will be connected to the GEANT network and integrated with public e-Infrastructures and commercial services using a federated identity management scheme to form a data management and processing backbone. This integration will draw on the developments in member states as well as the projects supported by the DG CONNECT e-Infrastructure work programme and actively engage research infrastructures, which are today's big data scientific factories, through the ESFRI cluster projects supported by DG RTD.

[2] Open innovation, open science, open to the world. A vision for Europe; Directorate-General for Research and Innovation; June 2016, doi:10.2777/061652

[3] 1 Exabyte = 1000 Petabytes = 1 million Terabytes.

Background

The fundamental problem is fragmentation: fragmentation of infrastructure limiting the scalability, interoperation and resilience of essential services; fragmentation of funding streams that deter the combination of public and private resources into seamless services; fragmentation of datasets stunting the socio-economic impact of the knowledge they hold. Such fragmentation also limits the sustainability of the services and erodes the trust research communities have in their ability to securely maintain and preserve their data.

The EC and member states have made significant investments in e-infrastructures over more than a decade. This investment has resulted in a set of distinct e-infrastructures supported by different projects each serving selected research communities. The focus has been on innovation to satisfy the technical requirements while, sustainability planning, strategy formulation and business modelling remains underdeveloped. Current sustainability for many of these e-infrastructure projects depends upon structures that do not yet inspire trust from the user communities. To build trust in a federated scientific data hub it is essential that a viable plan exists through which end-users can have confidence in the long-term availability, reliability and durability of its data services. Consequently, the *data hub* cannot be approached as yet another e-infrastructure project if it is to attract and retain a critical mass of researchers.

Advances in networking, hardware architectures, virtualisation and automation have remodelled data centres so capacity is consolidated into large-scale sites offering more reliable, secure and cost effective services. The private sector has been quick to recognise the advantages this transformation can bring and is now benefitting from higher levels of service while saving energy and money. National governments and regions have also started to rationalise their IT installations. This transformation represents an opportunity for the public research sector to reduce fragmentation while improving service quality, cost effectiveness and impact.

Consider a small number of state-of-the-art, large-scale data centres offering innovative data-intensive services in a secure and trusted environment. This group of data centres distributed across Europe interoperate to collectively form a *data hub* offering a range of data-intensive services to Europe's research communities. All the data centres are linked by a high performance network that extends to scientific instruments and leading data providers around the world. The *data hub* provides guarantees of long-term data management supported by well-defined business models. Such a *data hub* would be a magnet for users, service providers, data and investment. This would be solid basis for building an Open Science Cloud capable of meeting Europe's ambitions.

Data-intensive services

The *data hub* must address four criteria for services in order to achieve reproducible inter-disciplinary research and bridge academia and industry.

- The entire data lifecycle has to be actively managed in order to make data findable and accessible.
- The approach to inter-operability must be vendor neutral and globally supported.
- The mechanism for making data secure and re-usable over time must be sustainable.
- The implementation must adhere to relevant standards, notably for security and interoperability.

These essential features will be published as a set of requirements for conformant services. Only services that conform to the technical, legal and security requirements will be eligible for inclusion in the service catalogue. The requirements and set of relevant standards for conformant services and their operators will be kept to a minimum to avoid stifling innovation. All services will be made available under equitable terms and conditions based on financial and legal policies that ensure service interoperability and encourage open science, respecting European legislation notably those addressing data protection (such as GDPR) and intellectual property.

The services will include those components necessary to support what the Research Data Alliance (RDA) refers to as the 'virtual layer for management of the complete lifecycle of scientific data'[4]. An initial set of interoperable services will include data management, compute (including HPC) and 'Human' services such as consultancy, s/w development, training etc. The consultancy and training services are essential if the *data hub* is to broaden its user base. Support for uploading, porting and execution of user application codes and datasets will be an essential feature of the platform. A common metadata catalogue will include all datasets accessible via the data services together with a search/discovery facility. The services must offer a secure user Authorisation and Authentication mechanism and be accessible via the GEANT network. A centralised help desk linking together the help desks of the service providers, will also be available for users. A network of Computer Security Incident Response Teams (CSIRTs) linking all service providers will ensure operational security of the *data hub*.

These services will be drawn from both public and private sources and their level of usage will be monitored by the platform. Users can apply for grants corresponding to service credits which can then be used to pay for any conformant services in the service catalogue. Service credits will be allocated to end-users from the funds attributed to the *data hub* from multiple sources. Service providers set their own prices which must be visible in the service catalogue and will be paid according to the credits consumed. The service credit model will build on the experiences gathered from a pilot implementation of a commons credit model sponsored by the National Institute of Health in the USA[5]. The combination of a service catalogue, service credit scheme and pay-for-usage model will simplify growing the *data hub* to include new funding streams, users and service providers.

[4] Recommendations for Implementing a Virtual Layer for Management of the Complete Life Cycle of Scientific Data; Tobias Weigel, Peter Wittenburg; Jan 23, 2017; doi:10.23728/b2share.a921cfe6422544ec96302f60dece7393

[5] https://datascience.nih.gov/BlogCommonsCreditsModelPilot

Data preservation

Funding agencies have recognised the need for preservation and sharing of data with requirements on data management plans, preservation of data, reproducibility of results and sharing of data and results becoming increasingly important and in some cases mandatory. The business case for data preservation in scientific, educational and cultural as well as financial terms is increasingly well understood and funding beyond the standard lifetime of projects is required to ensure this preservation. Yet successful data preservation can only be performed if one understands the motivation for such preservation – who will be the eventual re-users of the data, what is or will be the knowledge base of these re-users and what types of re-use are desired, for example for scientific, commercial, educational or cultural reasons. Analysis of the needs of the research communities leads to four broad use-cases:

- **Bit Preservation**
  Bit preservation as a basic service on which higher level components can build.
- **Preserving data, software and know-how**
  Preserving data, software, and know-how in the research communities where they are produced is the foundation for the long-term data preservation strategy to ensure reproducible science: Data preservation alongside software evolution to accelerate access to knowledge.
- **Share data and associated software with (larger) scientific community**
  This brings additional requirements including storage for the released data, compute resources to process it and raises accessibility and intellectual property issues. Formalisation and simplification of data formats and analysis procedures together with documentation targeted at the specific consumer communities all become essential.
- **Open access to the general public**
  Engaging citizen scientists and the general public requires making a sometimes simplified and reduced set of the data available with associated education and outreach material as well as the continuous effort to provide meaningful examples and demonstrations.

Significant progress has been made in solutions for long-term data preservation to enable future re-use[6]. Continued investment in data preservation is needed to prevent data becoming unusable or lost and the investment in its production/acquisition squandered. Some of this investment can be more cost effective if performed centrally, e.g. by providing bit preservation services for multiple disciplines or projects, whilst important elements need to be addressed on a case-by case basis.

Each data centre that participates in the *data hub* will be certified as a trustworthy data repository according to the ISO 16363 standard. Compute services may be co-located within the data centres of the *data hub* or accessible over the network. A key concept in this model is that data does not leave this environment. Rather, clients access data within this combined *data hub* and compute cloud from externally and arrange for the processing. Small-scale data sets will eventually be downloaded for analysis locally, but organized and large scale processing would be done in this environment. Having all of the data virtually co-located in this manner will open the way to radically new analysis models while continuing to support existing approaches. Any data generated on external resources will be copied back for long-term storage to the *data hub*. Archived data will be maintained within the *data hub* by ensuring that at least two copies are replicated across the participating data centres.

[6] The Open Archival Information System (OAIS) is defined by the standard ISO 14721:2012 while ISO 16363 provides a framework for the certification and assessment of digital repositories.

Funding model

In the development of the service delivery cost model, the maintenance of production-quality operational status of each data centre in the *data hub* has a cost that cannot be ignored; it is not just an important factor in the overall service delivery cost calculation, it is politically relevant at the national level.

The funding model will mobilise public and private sector funding with a mix of capital and operational budgets, involvement of commercial service providers as well as public-sector e-infrastructure operators and a broad spectrum of stakeholders. There are, in fact, separate models for capital investment, service development, operational charges and charging structures for scholarly publishing which must be combined to cover the whole lifecycle of the platform. Capital investment can be supported via institutional/national resources, European Structural and Investment Funds (ESIF), European Fund for Strategic Investment (EFSI) as well as by commercial service providers. Development costs can be supported via streams such as institutional/national resources, the Horizon 2020 E-infrastructure work programme, pre-commercial procurement (PCP) and public procurement of innovation (PPI). HNSciCloud[7] has demonstrated the PCP instrument can be used to incite public and commercial service providers to co-design innovative services with research organisations that satisfy the pressing needs of Europe's research communities.

With regard to operational costs, the underlying premise is that service providers will be reimbursed according to metered usage while the services will be accessible to end-users as free at the point of use. This aspect will be developed in collaboration with projects funded via the EINFRA-12-2017 call to support Data and Distributed Computing e-infrastructures for Open Science[8].

Funding schemes vary from country to country and between different communities, for example sometimes computing costs are included in grants, sometimes not. The breadth of these funding schemes needs to be carefully assessed and reflected in the selected models.

It is important that communities beyond the public research sector use the services and so contribute to their funding. Ensuring the *data hub* remains as close as possible to mainstream IT trends will facilitate wider usage and spread the development and operational costs thereby improving sustainability. Innovation needs to be driven by the users with the added-value for end-users clearly demonstrated and viable business models identified. Developing business models does not imply that all services will be operated on a *for-profit* basis but they will ensure their long-term feasibility.

Stakeholders will be provided with regular, accurate and timely information about the status of the platform and individual services including quality levels, market penetration by sector and user community, etc. Information gathered via the operation of the service credit model will be used to produce a map showing which services are being used by each group of users that can feed into the stakeholder investment decisions. The progress with respect to the stated objectives will be reported to the European Open Science Policy Platform (OSPP[9]). The *data hub* must quantify the current needs of the research community and their likely growth in order to prepare a roadmap of investments necessary to serve more Research Infrastructures, users and business sectors.

The platform will support funding models that allow the data market to enter a faster growth trajectory. The supply-demand dynamics will change from technology-push to demand pull. This is a classic virtuous cycle mechanism where network effects multiply the benefits for users in their interactions and makes it easier to consolidate standards and interoperability, reducing further the barriers to adoption.

[7] http://www.hnscicloud.eu/

[8] http://ec.europa.eu/research/participants/portal/desktop/en/opportunities/h2020/topics/einfra-12-2017.html

[9] http://ec.europa.eu/research/openscience/index.cfm?pg=open-science-policy-platform

Governance

In 1627 the French librarian and scholar, Gabriel Naudé, produced a book called *Advice on Establishing a Library,* which laid-out the fundamentals of what has become library science. In order to turn the EOSC vision into a working implementation, we need to embed the practical advice of repository managers, IT service operators and researchers in the *data hub* governance and operation. Getting these aspects right is essential if EOSC is to repeat the impact of public libraries on scientific advancement. The governance model must also take into account the incentives of researchers to share. Important motivations for researchers to share research data are (1) when data sharing is an essential part of the research process; (2) direct career benefits derived from sharing through greater visibility of one's work, reciprocal data exchanges, and the reassurance of having one's data recognised as valuable by others; (3) the norms that researchers are exposed to within their research circle or discipline; and (4) a framework of funder and publisher expectations, policies, infrastructure and data services as external drivers[10]. Cultural change is required to achieve open science and researchers must be convinced that they will not lose control of their precious data. The data centres operated by public organisations and federated via the *data hub* can provide such guarantees. They can rapidly expand the available capacity by making use of commercial cloud services offering commodity compute and data services as part of the hybrid cloud model. By overseeing the data stewardship, the *data hub* can insulate researchers from changes in service provider even if the data custodian is a commercial cloud service provider.

The *data hub* governance model will be composed of 3 elements:

- **Principles of Governance** that express how the 'organisational structure' and 'governance processes' should be set up;
- **Organisational Structure** that define the different levels of responsibility, the roles of the governance bodies, who participates and who influences them;
- **Governance Processes** that formalise the set of activities (including their inputs and outputs) performed by each governance body (i.e. what they do and how they interact).

The following widely accepted governance principles will be adopted:

- Ensure alignment with the Digital Single Market, foster coherence, equitability and inclusiveness
- Enable integration of existing publicly funded e-Infrastructures with commercial cloud services effectively and efficiently
- Publicly governed with participation by all stakeholders to ensure a fair balance of their needs and interests
- Ensure transparency, openness and responsiveness
- Ensure value for money and fair incentives and returns
- Continuously manage legal and ethical compliance and other risks
- Ensure accountability and responsibility of stakeholders and decision makers
- Manage the identity and brand of the *data hub* and ensure sustainable innovation and growth.

The organisational structure will collaborate with and not replace the existing governance structures for research infrastructures, user communities as well as the projects and initiatives that contribute services (see Contributing initiatives). The governance structures will build on the decade of experience gathered with production federated grid service operation at a global scale[11] and will integrate into the model being discussed in the context of INFRADEV-04-2016 pilot EOSC project[12].

[10] Van den Eynden, V. and Bishop, L. (2014). Incentives and motivations for sharing research data, a researcher's perspective. A Knowledge Exchange Report, available from knowledge-exchange.info/Default. aspx?ID=733

[11] Lessons Learnt from WLCG Service Deployment, J.D. Shiers, doi:10.1088/1742-6596/119/5/052030

[12] http://eoscpilot.eu/

Contributing initiatives

Interaction with research infrastructures will be via ESFRI (notably the *Working Group on investment strategies in e-infrastructures* and its successors), the EC funded ESFRI cluster projects listed below as well as the EIROforum. These projects and initiatives will also provide channels to scientific user communities:

- ASTERICS[13] brings together the astronomy, astrophysics and particle astrophysics communities.
- CORBEL[14] is an initiative of eleven new biological and medical research infrastructures (BMS RIs), who together will create a platform for harmonised user access to biological and medical technologies, biological samples and data services required by cutting--edge biomedical research.
- EMBRIC[15] is designed to accelerate the pace of scientific discovery and innovation from marine Bio-Resources.
- ENVRI+[16] brings together Environmental and Earth System Research Infrastructures, projects and networks together with technical specialist partners to create a more coherent, interdisciplinary and interoperable cluster of Environmental Research Infrastructures across Europe.
- The High-Luminosity LHC (HL-LHC) will require a significant increase in data and compute resources compared to that needed for the LHC today and a model for evolving the existing production scientific computing infrastructure to meet these needs has been produced[17].
- PARTHENOS[18] aims to strengthen the cohesion of research in Linguistic Studies, Humanities, Cultural Heritage, History and Archaeology.
- SERISS[19] aims to strengthen and harmonise social science research across Europe.
- SINE2020[20] developing the innovation potential of neutron Large Scale Facilities and preparing for the European Spallation Source (ESS).

Involvement of research libraries will be channelled through LIBER – the Association of European Research Libraries[21]. Many publicly funded associations and projects are implementing specific services and defining relevant polices that can all be brought to fruition in the context of the *data hub*:

- Alliance for Permanent Access (APA[22]) develops information technology and human services that can assist research communities with their digital preservation needs.
- Data Seal of Approval[23] defines criteria for certifying data repositories are in accordance with national and international Guidelines for digital data archiving.
- EDISON[24] project aims to establish the data scientist as a profession.
- eInfraCentral[25] is a recently funded H2020 project that will develop an implementation of a common service catalogue, not just aimed at researchers, but also at industry, government, educators, and citizens; develop access and monitoring tools; and draw policy lessons.

[13] https://www.asterics2020.eu/
[14] https://www.elixir-europe.org/about/eu-projects/corbel
[15] http://www.embric.eu/
[16] http://www.envriplus.eu/
[17] Evolution of Scientific Computing; Ian Bird, January 2017; https://doi.org/10.5281/zenodo.291943
[18] http://www.parthenos-project.eu/
[19] http://seriss.eu/
[20] http://www.sine2020.eu/
[21] http://libereurope.eu/
[22] http://www.alliancepermanentaccess.org/
[23] http://www.datasealofapproval.org/
[24] http://edison-project.eu/
[25] http://www.efiscentre.eu/portfolio-item/european_e-infrastructure-services-gateway/

- European Grid Initiative (EGI[26]) is a federated e-Infrastructure set up to provide advanced computing services for research and innovation. The EGI e-infrastructure is publicly-funded and comprises over 300 data centres and cloud providers spread across Europe and worldwide.
- EOSC pilot[27] is a recent project that will support the first phase in the development of the EOSC.
- EUDAT[28] implements data services and is preparing a Common Data Infrastructure (CDI) collaboration agreement intended to support the data services after the current project completes which could be used to formalise the relationship between the data centres participating in the data hub.
- FORTISSIMO[29] enables European SMEs to be more competitive globally through the use of simulation services running on a high performance computing cloud infrastructure. Fortissimo has created a Marketplace to enable users, and prospective purchasers, of high performance computing services to more easily access and purchase such services from suppliers.
- GÉANT[30] is the pan-European data network for the research and education community. It interconnects national research and education networks (NRENs) across Europe, enabling collaboration on projects ranging from biological science to earth observation and arts & culture.
- The GO-FAIR initiative[31] is developing an implementation approach for a global science commons based on three interactive processes: building technical infrastructure (GO-BUILD), which is complemented by a cultural change programme involving relevant stakeholders (GO-CHANGE), and the training of data stewards capable of providing FAIR data services (GO-TRAIN).
- Helix Nebula[32] is a public-private partnership between public research actors and cloud service providers to promote adoption of cloud services for scientific use and has undertaken the first joint Pre Commercial Procurement (PCP). This €5.3 million joint tender, led by CERN, will establish a hybrid cloud platform supporting high-performance, data-intensive scientific use-cases sponsored by 10 of Europe's leading public research organisations and co-funded by the European Commission.
- INDIGO - DataCloud[33] is a Horizon 2020 project that develops an open source data and computing platform targeted at scientific communities, deployable on multiple hardware and provisioned over hybrid, private or public, e-infrastructures.
- OpenAIRE[34] has established a network of data repositories that can benefit from the services offered by the data hub. OpenAIRE also supports national help desks that can encourage the update of the data hub services.
- The Research Data Alliance (RDA[35]) animates a multi-disciplinary international forum through which technical and policy aspects of data services are discussed and promoted. It has recently identified a set of components that it sees as necessary for implementing o virtual layer for the management of complete life cycle of scientific data.
- THOR is a H2020 project that provides a dashboard[36] which monitors the evolution of persistent identifier (DOI) interoperability that can provide a measure for the sharing and re-use of data of research data.
- Up2U[37] project will bridge the gap between education & research by better integrating formal (academic) and informal, self-learning scenarios.

[26] https://www.egi.eu/
[27] http://www.eoscpilot.eu/
[28] https://www.eudat.eu/
[29] https://www.fortissimo-project.eu/
[30] http://www.geant.org/
[31] http://www.dtls.nl/go-fair/
[32] http://www.hnscicloud.eu/
[33] https://www.indigo-datacloud.eu/
[34] https://www.openaire.eu/
[35] https://www.rd-alliance.org/
[36] http://dashboard.project-thor.eu
[37] http://cordis.europa.eu/project/rcn/206177_en.html