
**ASTERICS - H2020 - 653477**

# Excerpt of EOSC discussions at the 2nd ASTERICS-OBELICS Workshop, Oct 2017.

**H2020-ASTERICS**

| | |
|---|---|
| Document identifier: | |
| Date: | **24 November 2017** |
| Work Package: | **WP3 OBELICS** |
| Lead Partner: | **LAPP** |
| Document Status: | **Report** |
| Dissemination level: | **Public** |
| Document Link: | |

Abstract

This report gives an overview of the discussions on Astronomy & Astroparticle Physics assets in building the European Open Science Cloud (EOSC) at the 2nd ASTERICS-OBELICS Workshop that took place on 17 October 2017 in Barcelona. The event included participation from Astronomy ESFRI projects, EOSC HLEG, EOSCPilot, EOSChub, EOSC science demonstrator projects, CERN as well as e-infrastructures. The document provides a list of recommendations that emerged during the discussions and can be considered for EOSC implementation.

# I. COPYRIGHT NOTICE



# II. DELIVERY SLIP

| | Name | Partner/WP | Date |
|---|---|---|---|
| From | JAYESH WAGH | CNRS-LAPP | 24-11-2017 |
| Author(s) | | | |
| Reviewed by | | | |
| Approved by | | | |

# III. DOCUMENT LOG

| Issue | Date | Comment | Author/Partner |
|---|---|---|---|
| 1 | 22-11-2017 | First draft | Jayesh Wagh, LAPP |
| 2 | | | |
| 3 | | | |
| 4 | | | |

# IV. Table of contents

# 1. Introduction

The European Open Science Cloud (EOSC) aims to accelerate and support the current transition to more effective Open Science and Open Innovation in the Digital Single Market. It should enable trusted access to services, systems and the re-use of shared scientific data across disciplinary, social and geographical borders. The recent EOSC High Level Expert Group (HLEG) report approaches the EOSC as a federated environment for scientific data sharing and re-use, based on existing and emerging elements in the Member States, with light-weight international guidance and governance and a large degree of freedom regarding practical implementation. The EOSC includes the required human expertise, resources, standards, best practices as well as the underpinning technical infrastructures. Despite the success of the European Strategy Forum on Research Infrastructures (ESFRI), fragmentation across domains still produces repetitive and isolated solutions. The 2nd day of the ASTERICS-second OBELICS workshop 2017 was dedicated to building bridges between ESFRI projects, concerned scientific communities, e-infrastructures, and further consortia, and to call them to action for the implementation of the EOSC for data interoperability in the astrophysics related disciplines.

# 2. Understanding EOSC from the experts

The EOSC High Level Expert Group (EOSC HLEG) has been mandated to support the European Commission in the set-up of data-driven infrastructure that builds on what already exists, that caters for the whole scientific community and brings together the building blocks developed through different categories of stakeholders to provide actionable options for the design of the EOSC governance and services. Some of the key publications and events organized by the EC to define the steps to implement EOSC include:

- EOSC Declaration which is a key input for the EOSC roadmap.
- EOSC Stakeholders Forum November 2017 which aims at endorsing the EOSC declaration and discussing the EOSC roadmap.
- **EOSC Roadmap** which will be announced in December 2017 providing information on the EOSC governance structure, architecture and core services.

To further clarify the concept, EOSC can be compared with Commons Credit Models Pilot which is designed to provide investigators with access to cloud based computing resources as a mean to seed the NIH commons. Emphasis should be given to the fact that finding a right cloud service provider will be of paramount importance considering the General Data Protection Regulations. Workflows appear to be a common denominator for the Science Demonstrators that have currently been selected under EOSC Pilot. EOSC potential services could be seen a broker for Cloud resources, and a marketplace of atomic micro services that can be combined in traceable workflows.

# 3. Insights from other projects and research infrastructures

**H2020 EOSC-hub** mobilises providers from the EGI Federation, EUDAT CDI, INDIGO-DataCloud and 18 major research e-infrastructures offering services, software and data for advanced data-driven research and innovation. The project will create the Hub for the integration and management system of the future EOSC.

EOSCHub will address 2 main challenges:

- How to make large data more easily accessible? We already have some experience with previous projects but other communities have their data managed by third party.
- Policy and financial problems. There's no easy way to access to services that can be used by different communities.

The scientific, technical and cultural challenges for EOSC includes

- Scientific Challenges: Deploying the EOSC to deliver Open Science.
- Technical Challenges: Developing technical solutions to meet scientific needs.
- Cultural Challenges: Adopting new and more open ways of working.

These challenges can be addressed with coordinated efforts to

- Bring the existing research infrastructures together.
- Bring e-infrastructure projects together (GEANT , PRACE..).
- Interoperate between their services and data.
- Allow new resources to be added (HPC providers, Cloud providers, Data providers…).

The main objective for **EOSCpilot Architecture and the services** is to propose a governance framework for EOSC. There are 10 Science Demonstrator from different domains at present and 5 more will be added at the end of November 2017. EOSCpilot is also defining who the users of EOSC: Data scientists, researchers, managers, service developers, service providers, einfra providers. In the architecture framework we should consider where all these stakeholders fit.

One of the **EOSC Pilot Science Demonstrator for LOFAR** aims to improve the user experience both for power users and non-power users. The project will be built upon existing knowledge and by combining existing tools to show the complete path from facility to user, to

demonstrate it can be done and to demonstrate the capacity and shortcomings, or challenges for the future.

The project plan for this demonstrator includes

- SurfSARA providing connectivity
- Common Workflow Language to standardize existing pipelines
- User Settings to build frontends
- Use of Zenodo platform for persistent storage DOI

The final objective of this demonstrator is to demonstrate a complete system can be built from existing tools.

EVN as cloud version is a pilot from the JIVE institute, to make the **European VLBI Network (EVN)** data accessible through the cloud. The EVN archive has everything that a large archive has, except the size. Providing a cloud version would be a good test case for storage and data reduction in the cloud.

The goal of **H2020-AENEAS** is to design a distributed **European Science Data Centre (ESDC)** to support the pan-European astronomical community in achieving the scientific goals of the SKA. It is very important that the design of the ESDC runs parallel to the emergence of the EOSC and on the way learning from each other. The work carried out in **H2020- HNSciCloud** is a contribution towards the development of the EOSC as well as in view of the next Work programme 2018-2020 call INFRAEOSC.

In order to share data, we need to store the data in a Trustworthy Digital Repository (TDR). Researchers must be certain that data held in archives remain useful and meaningful into the future. In a multidisciplinary environment it is necessary to understand and implement FAIR (Findable, Accessible, Interoperable and Re-usable) Data principles. In case of High Energy Physics **LONG TERM DATA PRESERVATION** refers to documentation, software and the environment in which it runs. As per e-Infrastructure Reflection Group, **Long Term** stands for a period of time long enough for there to be concern about the loss of integrity of digital information held in repositories, including deterioration of storage media, changing technologies, support for old and new media and data formats (including standards), and a changing user community. **From the experience of EOSC Pilot HEP LTDP we understand that even at "modest" scale (100TB), HEP data formats and long-term needs mean that a "generic" TDR is unlikely to work.**

The upgrades of the LHC and its large detectors, planned for the middle of the next decade (the HL-LHC project), will pose significant new challenges for the computing and data infrastructure. HL-LHC will produce several Exabytes of scientific data per year, and will

require some 20 M cores of processing power. The WLCG community is investigating potential changes to the computing models and distributed computing infrastructure that will be needed to meet those challenges over the coming years. In particular, it will address how to create a data infrastructure at the Exabyte scale, that is able to manage and process data in an effective way. High Energy Physics community has produced a community white paper (CWP) through HSF meeting. The main themes of CWP are

- Allow/help countries or regions to flexibly manage compute and storage resources internally
- Investigate the "data-lakes" concept – keep bulk data (down to derived AODs) in a cloud-like realm (data-lake). Plug in processing via traffic-managed networks, bulk processing close to the data.

A prototype of **"data-lakes"** concept would represent a valid asset to build EOSC . This conceot is of interest for other projects such as CTA,SKA.

**Square Kilometre Array** is an international project to build world's largest radio telescope. The project is in its design phase (2013-2018). About 100 organisations across about 20 countries are involved in the design phase. The construction phase 1 is expected to start in 2019. SKA – will generate around 300 PB (petabytes) of data products every year. Hence the computing challenges are simillar to WLCG. Both HL-LHC and SKA will be Exabyte-scale scientific experiments on a 10-year timescale **Recognising this commonality, on 13th July 2017, CERN-SKA signed an agreement for cooperation on Computing and big data management.**

Initial topics for collaboration includes.

- Bi-annual collaboration meetings
- A position paper/roadmap to be produced quickly to focus on Exabyte-scale data infrastructure needs
- Explore collaboration opportunities on common aspects of networking, storage, computing, etc
- Investigate work with industry via CERN Openlab
- Joint projects to demonstrate/prototype concepts for regional centres and computing models

A long tradition of international collaboration has been made to build telescopes and instruments, since 1977 where the Data format FITS which is still used today. Astronomical Interoperability Framework was created in 2012 it defines the development of interoperability standards. In Astronomy there are open interoperable resources that have a centralized access point in the Virtual Research Environment. **Virtual Observatory** can build blocks reused by other data providers. Researchers will expect to find the data services used in everyday in the same EOSC system of system. At a national level ESFRI should be considered as a building block towards the EOSC development. Light Interoperability layer has been one of the key to success as experienced by IVOA which is already included in the Registry of Resources in EUDAT. Astronomy on line services and the VO should be included “in” EOSC as they’re used in everyday by users.

# 4. Some recommendations for the way forward

- Today there's already a good network in Astronomy and Astrophysics community based on agreements between different infrastructures. Tomorrow we have to guarantee that these networks are maintained. What we will need is to consolidate the workflows, the access to the archive data and a collaboration between a central EOSC point and the new infrastructures. Astronomy and Astroparticle physics community could contribute to the EOSC design while it's in design phase. Some of these EOSC design aspects can be workflow technologies and pipelines that allow users to change settings and inputs. As an example, easy Access to LOFAR data & knowledge extraction through the Open Science Cloud can be achieved.

- EOSC could be defined as a system of systems made of existing and emerging RIs, e infrastructure, data repositories, registries etc. The system of systems approach should present certain features; the most important one is that EOSC will maintain an Operational and managerial independence. EOSC is not going to take over an existing infrastructure but to add value to it.

- **EOSC should be built upon the existing infrastructure** and by using knowledge and requirements of current large archives and compute facilities, and mapping a scale increase of one to two orders of magnitude. It will stretch the capacity of any cloud or existing infrastructure to the limit.

- **ESFRIs are not represented enough in the EOSC governance model** as for now therefore it is important to make sure that they are well represented. ESFRI projects can participate to the 2018-2020 work programme making use of commercial services and engage with EOSC. The ESFRIs could be used to ensure that there is commitment for Member States in the development of EOSC. It should leverage on the ESFRI contributions with support and inclusion of new models and services such as the **"data lakes"** .

- The next EU H2020 call clearly states "*Research Infrastructures, such as the ones on the ESFRI roadmap and others, are characterised by the very significant data volumes they generate and handle. These data are of interest to thousands of researchers across scientific disciplines and to other potential users via Open Access policies*." Hence **Synergies between ESFRIs** is important to identify and use FAIR data management platforms. ESFRI representatives attending the panel discussion mentioned the need of cluster actions such as H2020-ASTERICS for the EOSC implementation.

- At present HL-LHC and SKA are the research infrastructures facing the Exascale challenges. Other infrastructures such as CTA, LSST, Virgo-LIGO are facing challenges close to the order of Exascale as well. Therefore a common and coordinated approach to access HPC and open science resources is the need of the time. Other ESFRI and world class projects such as CTA, LSST, Virgo-LIGO have similar requirements and constraints. WLCG has significant experience and tools for hybrid clouds with academic institutions, commercial entities as well as EC. **The consortium including ESFRIs and HL-LHC will represent a strong candidature to steer and implement the EOSC.** In order to form such consortium, it is important to understand the commonality and complementarity. It would make sense to work with EOSC as a single grouping of aligned interests.

- If science drives, the cooperative actions with e-infrastructures and service providers to implement open science would be more effective and reliable. **EOSC cannot replace the ESFRI communities** but it can support the ESFRI communities shared interest by helping the Cluster activities. Regarding Long Tail Science, it can support the high bandwidth needed by the ESFRI projects data centres and also interoperability and preservation goals. The EOSC cannot be only about Cloud. EOSC should be a framework to bring together software tools, services, communities. EOSC should be sufficiently open to accept solutions developed or adopted or pre evaluated in the Cluster activities or other communities.

- EOSC should move towards the federation of existing e infrastructure. EOSC should be explored into little steps/issue to make them interoperable with other. Service CESSDA has a huge amount of Catalogue of Services. Being an ERIC, they realized that now they have to change their communication so what they were offering before is now more clear. The fact that EOSC can be promoting Services that have been developed by other infrastructure, could be more interesting for Funding agencies, because in this way, being in EOSC would work as an **incentive for the infrastructures which deployed services used in EOSC.**

- **H2020-ASTERICS** seems to be the only project connecting ESFRIs together to address data interoperability and software reuse and this experience can play a key role for implementing the EOSC even from the point of view of other scientific domains. The ESFRI communities can contribute to data interoperability and software reuse by bringing together people towards common grounds to work together and find solutions/services that can be used across different communities.

- **EOSC for ESFRIs:** Each RI produces data at a continuous rate, and performs basic processing (calibration, geometric registration, etc.): for that purpose it is expected to

have its own dedicated e-infrastructure. The result are archives of persistent data. EOSC should provide data-computing interoperability mechanisms to allow processing on archived data at the archive location (data centre).

- **ESFRIs for EOSC:** Each ASTERICS RI provides an archive of datasets in physical units (i.e. reusable); if the IVOA standards are used, data are FAIR. Additional services could be provided (cut-out service, catalogues of astroobjects, classification, etc.) at the discretion of the data centre. Additionally, a repository of software tools, pipelines, etc., could also be provided.

- **Policy Access:** Datasets are open and public, after a (usually short) proprietary period; metadata are always public except for peculiar cases (e.g. search for extrasolar planets); software is usually public as well –in these cases registering users is not necessary (even undesired). For the use of additional services requiring computing users should be authorised (A&A necessary).