**H2020-EINFRA-2017**

**EINFRA-21-2017 - Platform-driven e-infrastructure innovation**

**DARE [777413] "Delivering Agile Research Excellence on European e-Infrastructures"**

# Platform Infrastructure, Usage & Deployment I

| | |
|---|---|
| **Project Reference No** | 777413 — DARE — H2020-EINFRA-2017 / EINFRA-21-2017 |
| **Deliverable** | D5.1 Platform Infrastructure, Usage & Deployment I |
| **Work package** | WP5: Platform Operation and Maintenance |
| **Tasks involved** | T5.1: Provision of Relevant Cloud Infrastructure<br>T5.2: Provision of Pre-release Testbeds<br>T5.3: Deployment Strategy and Platform Operation |
| **Type** | R: Document, report |
| **Dissemination Level** | PU = Public |
| **Due Date** | 31/12/2018 |
| **Submission Date** | 31/12/2018 |
| **Status** | Draft – v6 |
| **Editor(s)** | Malin Ewering (SCAI), André Gemünd (SCAI) |
| **Contributor(s)** | |
| **Reviewer(s)** | Iraklis Klampanos (NCSRD) |
| **Document description** | This deliverable reports on the computational resources mobilized to provide the necessary development and production infrastructure in |

| | |
|---|---|
| | the scope of the DARE project. Furthermore, it presents WP5s approach to ensure a successful deployment of the DARE platform and provide preliminary usage statistics. |

## Document Revision History

| Version | Date | Modifications Introduced | |
|---|---|---|---|
| | | Modification Reason | Modified by |
| **1** | 30/08/2018 | Initial Structure | M. Ewering (SCAI) |
| **2** | 28/11/2018 | First Version | M. Ewering (SCAI) |
| **3** | 05/12/2018 | Third version | A. Gemünd (SCAI) |
| **4** | 19/12/2018 | Version for internal Review | A. Gemünd (SCAI)<br>M. Ewering (SCAI) |
| **5** | 20/12/2018 | Internal Review | I. Klampanos (NCSRD) |
| **6** | 20/12/2018 | Ready for Submission | M. Ewering (SCAI) |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |

## Executive Summary

This deliverable reports about the computational resources mobilized to serve

- the software development/pre-release infrastructure for DAREs software products,
- the infrastructure required for the final and public deployment of the DARE platform.

Furthermore it defines the approach to transit software, computational and data assets from development to release packages and provides details on the process of deploying the DARE platform and associated components to a publicly available infrastructure.

Finally, preliminary usage statistics will be presented.

## Table of Contents



## List of Figures

## List of Tables

## List of Terms and Abbreviations

| Abbreviation | Definition |
|---|---|
| EPOS | European Plate Observing System |
| IS-ENES | Infrastructure for the European Network for Earth System Modelling |
| C4I | Climate4Impact |
| EGI | European Grid Infrastructure |
| CDI | Collaborative Data Infrastructure (EUDAT) |
| PRACE | Partnership for Advanced Computing in Europe |
| CINECA | A not-for-profit Consortium, made up of Italian universities, Research Institutions, and the Italian Ministry of Education. |
| CSCS | Swiss National Supercomputing Centre |
| GCS | Gauss Centre for Supercomputing |
| GENCI | Grand équipement national de calcul intensif |
| GPGPU-VT | General Purpose Graphics Processing Unit Virtual Team |

# 1 Introduction

In the scope of the DARE project, the consortium will refine, extend, strengthen and integrate diverse software assets like for example dispel4py, S-ProvFlow, Exareme and Semagrow. Independently of DARE, both involved communities EPOS (computational-seismology) and IS-ENES/Climate4Impact (climate), also develop community-specific software and infrastructure. In addition, in the scope of several other European projects and initiatives, relevant components for the DARE-plattform have been and partly still are under development. One of the aims of the project is to demonstrate how these components can be utilized together and brought into the remit of one development platform for each community, to address the requirements of their particular applications and use cases.

In this context, the particular aim of Work Package 5 - Platform Operation and Maintenance, is to ensure that all necessary infrastructure is provided to the relevant actors, to assess and manage the computational resources that are required for the smooth implementation and operation of the platform and carry out the necessary maintenance and update tasks. In this process, WP5 builds upon the results of WP2 dealing with the architecture specification of the DARE platform (D2.1) and taking care of the topic Data Management (D2.5). Interfaces to WP3 and WP4 in the first line emerge during the integration of software assets, tools and services. The cooperation with WP6 and WP7, delivering the high-quality use cases, becomes of key interest in the evaluation process and in particular during the application acceptance testing, which builds an important element during the deployment procedure.

As a first step, the DARE platform was initially set up as a development platform and framework for setting up all services the two use cases need to enable testing and evaluation at challenging scales. In the course of the project, this platform will evolve into a consistent and integrated environment through which application communities can develop and run software, services, data and tools that meet their communities' needs. By the end of the project, DARE will provide an operative instance of the DARE platform integrated into the European e-infrastructure (expected EOSC) and additionally make it available at public GitLab.com repositories to enable interested parties to install and use it locally.

The purpose of D5.1 - Platform Infrastructure, Usage & Deployment I is to report on the current status of the available infrastructure and to present a strategy for the platform deployment. This report only reflects the first project year of DARE. The described deployment strategy/process may evolve during the course of the project and will be adapted if necessary. The final version, D5.2 - Platform Infrastructure, Usage and Deployment II, will take into account developments and priorities throughout the project and will be delivered at M36.

# 2 Infrastructure and Platform

To realize the DARE platform facing the challenges of extreme data, extreme complexity and extreme computing, several components need to be brought together. Low-level services from multiple e-Infrastructures need to come together with the specific DARE services and tools as well as the domain-specific services and applications of the two use cases. The use of external datasets needs to be enabled and the private resources of GRNET and SCAI will need to be bridged to the public e-infrastructures. To get access to the necessary Cloud e-Infrastructure, it will be obligatory to enter into discussions with responsible parties from EGI, EUDAT, INDIGO-DataCloud and EOSC-hub. To

enable communities to make use of HPC resources through the DARE platform, all technical prerequisites will need to be met in order to connect to the HPC-infrastructures.

## 2.1 Private Infrastructure

The basis infrastructure for the DARE platform consists of diverse components. Even though the goal is that, at the end of the project, the DARE platform will be utilizing public e-infrastructures, the private resources of the two project partners, GRNET and Fraunhofer SCAI play an important role, especially in the first project phase.

To set up a first framework for the DARE plattform and enable testing and development, SCAI made available a private, Openstack based cloud and GRNET a Synnefo based cloud. Both clouds are part of the European e-Infrastructure called EGI Federated Cloud which is presented in chapter 2.2.

Openstack as well as Synnefo provide software tools for building and managing cloud computing platforms for private or public clouds.

- OpenStack[1] is a free and open-source cloud platform. The software consists of interrelated components that control diverse, multi-vendor hardware pools of processing, storage, and networking resources throughout a data center, managed through a dashboard or via the OpenStack API. Besides the core services for Compute (virtualization), Image Management, Storage, Networking and Identity Management, the OpenStack landscape is continuously extended with additional services that provide value on top of the core services, such as Orchestration, Workflows, Alarming, Databases-as-a-Service, etc.

- Synnefo[2] is a complete open source cloud stack written in Python that provides Compute, Network, Image, Volume and Storage services. It manages multiple Ganeti clusters at the backend for handling of low-level VM operations and uses Archipelago to unify cloud storage. It is self- developed by GRNET and is powering two of its public cloud services, the ~okeanos service, which is aimed towards the Greek academic community, and the ~okeanos global service, which is open for all members of the GÉANT network.

## 2.2 European Science Clouds

It is envisaged that the final DARE platform will be available through the European Open Science Cloud (EOSC) to the wider research community. The EOSC initiative is expected to unite existing services from e-infrastructures, such as EGI[3] and EUDAT[4]. Beyond that, DARE strives to revise the use of HPC and GPU resources through the cloud. The enormous significance of compute-intensive HPC resources and GPUs becomes apparent while looking at the requirements of DAREs use cases. In this regard, it is important to note that DARE won't be able to provide universal access to such resources due to the funding and access model of European HPC resources. It will rather lay the foundation and provide relevant tools to make use of additional external resources to facilitate the exploitation of the resources that users have at their disposal. To get access to HPC-resources, users can e.g. request resources at PRACE (cf. chapter 2.3) or make use of commercial vendors like e.g. AWS. An alternative to "HPC over Cloud" is represented by the usage of GPUs through the Cloud. This, for example, is relevant for the EPOS use case, whose primary simulation code is able to support a

GPU mode to accelerate computations. The use of GPU resources in this context allows to dramatically reduce the computational burns in terms of economic costs and computing time. However, currently the opportunities to get access to GPU resources in the scope of the European e-Infrastructures seems to be limited. Progress in this direction especially in the scope of the EOSC are followed continuously.

### EGI Federated Cloud

The EGI Cloud compute service[5] (EGI Federated Cloud) is implemented as a hybrid cloud composed by public, community and private cloud providers. It is a public cloud building around open standards and targeted at and accessible by European research communities.

The EGI Cloud enables its users to run and scale virtual machines on demand with complete control over computing resources. It makes it possible to share resources and applications across institutes and national borders, develop portable and standard-based applications and services, operate high-quality services for science, and to establish sustainable e-infrastructures for large-scale, digital science. It offers guaranteed computational resources in a secure and isolated environment without the overhead of managing physical servers and furthermore the possibility to select pre-configured virtual appliances (e.g. with optionally pre-configured CPU, memory, disk, operating system and software) from a catalogue replicated across all EGI cloud providers. The EGI Cloud Service also provides a Training Infrastructure for training events. The Cloud Container Compute (beta phase) offers the ability to deploy and scale Docker containers on-demand. A list of all EGI Services[6] is published at https://www.egi.eu/services/.

In 2012 the GPGPU-VT[7] (General Purpose Graphics Processing Unit Virtual Team) evaluated the impact of GPGPUs and investigated if they provide an added value to EGI[8]. A feasibility study was conducted that demonstrated, the technical realisation of this service. However, at this point in time, only the FedCloud Provider IISAS offers GPGPU resources through the EGI federated Cloud[9].

Since multiple years, SCAI and GRNET provide access to their private clouds within the scope of the EGI Federated Cloud. In this context, GRNET operates a Synnefo technology based cloud providing 70 CPUs with 220GB RAM and 2TB storage. Fraunhofer SCAI operates an Openstack-based cloud comprising 128 physical cores + HT with 244 GB RAM and 20 TB Storage[10]. As mentioned above, these resources build the basis for the DARE infrastructure and will be extended and finally replaced by resources from the publically available e-infrastructures (EGI, EUDAT, EOSC) during the course of the project. Both partners not only introduce their resources but also related know-how in the project.

### EUDAT- Research Data Services, Expertise & Technology Solutions

EUDAT[11] is a service-oriented, community driven, sustainable and integrated initiative for providing data services for researchers. Its vision is to enable data stewardship within and between European research communities through a Collaborative Data Infrastructure (CDI).

More precisely, EUDAT offers a service-oriented data infrastructure of integrated data services through a geographically distributed, resilient network connecting general-purpose data centres and community-specific data repositories. This enables researchers, policy makers, and members of the public to use the CDI that offers solutions for finding, sharing, storing, replicating, staging and performing computations with primary and secondary research data. The associated services are

summarized in the table below and published at https://eudat.eu/catalogue. DARE intends to make these services easily usable from its technology stack where appropriate and beneficial for the use cases of the User Communities.

**Table 1: EUDAT services[12]**

| | | |
|---|---|---|
| **Data Hosting, Registration & Management & Sharing** | B2NOTE | B2Note allows to easily create, search and manage annotations. An annotation is a keyword or commentary attached to a data object (data collection, file) that explains or classifies it. B2NOTE is a standalone service for annotating data content hosted within the EUDAT CDI. |
| | B2SAFE | B2SAFE is a robust, safe and highly available service which allows community and departmental repositories to implement data management policies on their research data across multiple administrative domains in a trustworthy manner. The service provides an abstraction layer of large scale, heterogeneous data storages and guards against data loss in long-term archiving. It allows to optimize access for users (e.g. from different regions) and brings data closer to facilities for compute-intensive analysis. |
| | B2HANDLE | B2HANDLE is a distributed service for minting, storing, managing and accessing persistent identifiers (PIDs) and essential metadata (PID records) as well as managing PID namespaces. The implementation of the service relies on the DONA/Handle persistent identifier solution. |
| | B2DROP | B2DROP is a secure and trusted data exchange service for researchers and scientists to keep their research data synchronized and up-to-date and to exchange with other researchers. It allows to store and exchange data with colleagues and team members, to synchronise multiple versions of data and to ensure automatic desktop synchronisation of large files. |
| | B2SHARE | B2SHARE is a user-friendly, reliable and trustworthy way for researchers, scientific communities and citizen scientists to store and publish small-scale research data from diverse contexts. B2SHARE is a solution that facilitates research data storage, guarantees long-term persistence of data and allows data, results or ideas to be shared worldwide. The basic production service comprises the following features: self-service registration for any scientists and researchers, free upload and registration of stable research data, data access policy is defined by the data owner, metadata is openly accessible and harvestable, customized metadata handling and customized user interfaces, data integrity is ensured by checksums which are calculated during data ingest, the data is kept online, the storage usage is based on the principle of fair share. |

| | | |
|---|---|---|
| **Data discovery** | B2FIND | B2FIND is the EUDAT metadata service and provides a discovery portal which allows users to find data collections within an international and inter-disciplinary scope. It is based on a comprehensive metadata catalogue of research data collections stored in EUDAT data centres and other repositories. Harmonization of the metadata descriptions collected from heterogeneous sources enables not only the presentation in a consistent form but as well the faceted search across scientific domain boundaries. |

| | | |
|---|---|---|
| **Data Access, Interface & Movement** | B2STAGE | B2STAGE is a service to transfer research data sets between EUDAT data resources and external workspaces mounted on high-performance computing systems. It provides a common API on basis of GridFTP and HTTP and can be used through any standard GridFTP or HTTP client. The service allows users to: transfer large data collections from EUDAT storage facilities to external HPC facilities for processing; ingest computational results onto the EUDAT infrastructure; access stored data sets through associated PIDs; replicate community data sets in conjunction with B2SAFE; ingesting them onto EUDAT storage resources for long-term preservation. The service allows repository managers to use B2SAFE for storing their data that is transferred from their external repository to the CDI. |

| | | |
|---|---|---|
| **Identity and Authorization** | B2ACCESS | B2ACCESS is the EUDAT federated cross-infrastructure authorisation and authentication framework for user identification and community-defined access control enforcement. B2ACCESS allows EUDAT users to authenticate themselves using a variety of credentials. |

### European Open Science Cloud (EOSC)

The European Open Science Cloud, initiated by the European Commission, is a large infrastructure to support and develop open science and open innovation in Europe and beyond. The EOSC was officially launched at 23 November 2018 and is intended to become Europe's virtual environment for all researchers to store, manage, analyse and re-use data for research, innovation and educational purposes[13]. The EOSC implementation roadmap describes six action lines for the implementation of the European Open Science Cloud:

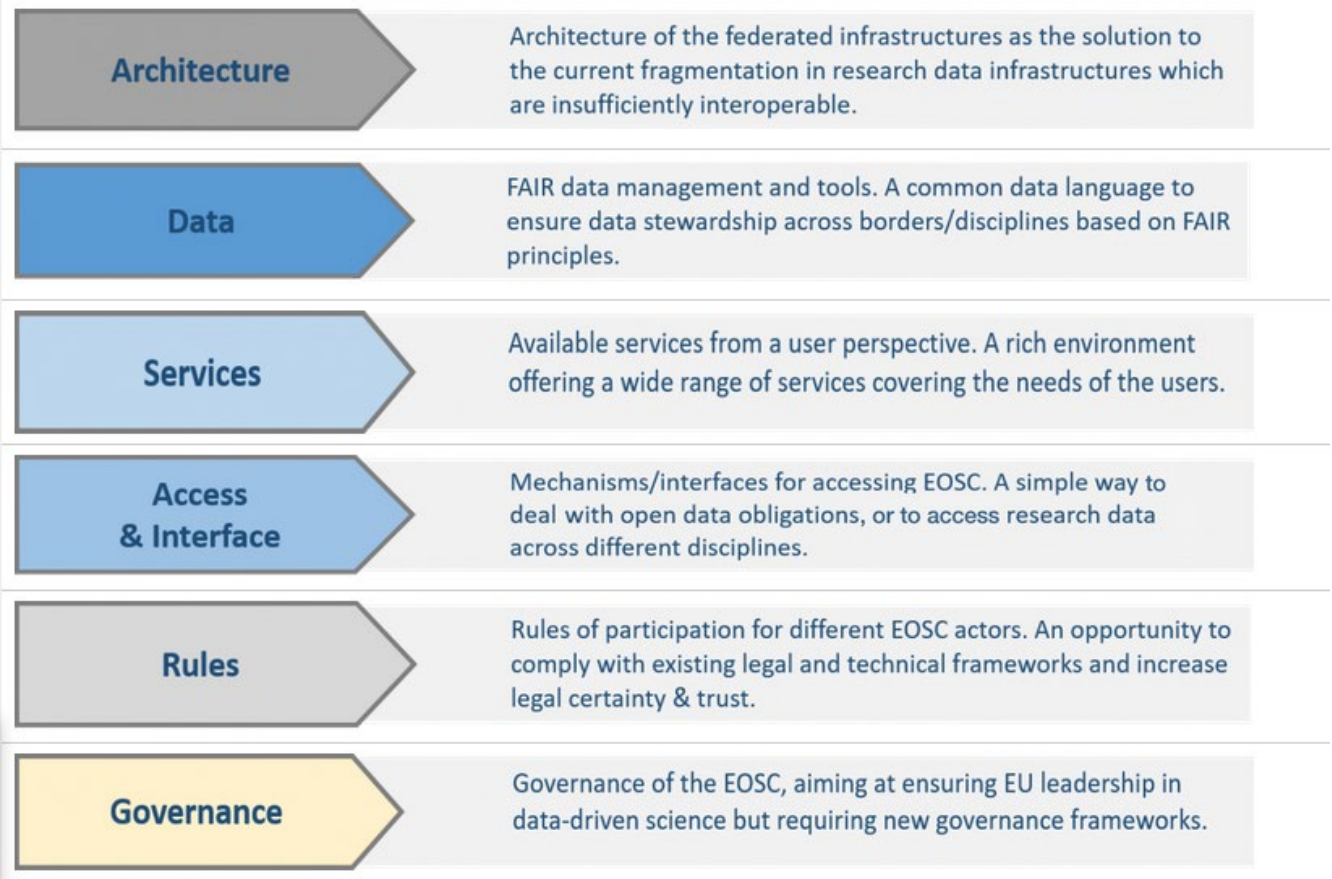


**Figure1: Action lines for EOSC. Source: EOSC Portal [14]**

During the kick-off event hosted by the Austrian Presidency of the European Union the new EOSC Portal (https://eosc-portal.eu/) was introduced. It provides all relevant information and builds the entrance point to resources, services and data[15] Also the DARE platform will be an EOSC-service published at this public e-Infrastructure and accessible for the public.

To implement the vision of this pan-European virtual environment, the European Commission provides financial support to various ongoing and future projects. Figure 2 provides a list of various ongoing and future projects taking part in the evolution of the EOSC by 2022.

To support the first phase of this initiative the EOSCpilot[16] project was launched to bring together stakeholders from research infrastructures and e-Infrastructure providers as well as funders and policy makers to propose and trial EOSC's governance framework. The project furthermore involved several science demonstrators to integrate services and infrastructures to exemplarily show the interoperability and benefits of such cloud in scientific domains like earth sciences, high-energy physics, social sciences, life sciences, physics and astronomy.

In 2018 the EOSC-hub[17] project kicked-off to create the integration and service management structure of the European Open Science Cloud. In general it could be seen as the lead contributor to the development of the EOSC Portal and its components. The project plans to enable an open access to research resources from various scientific disciplines via a digital hub: an integration system of software and services from major European e-infrastructures and research infrastructures. This digital

hub should enable researchers and innovators to discover, access, and use a variety of advanced data-driven resources.

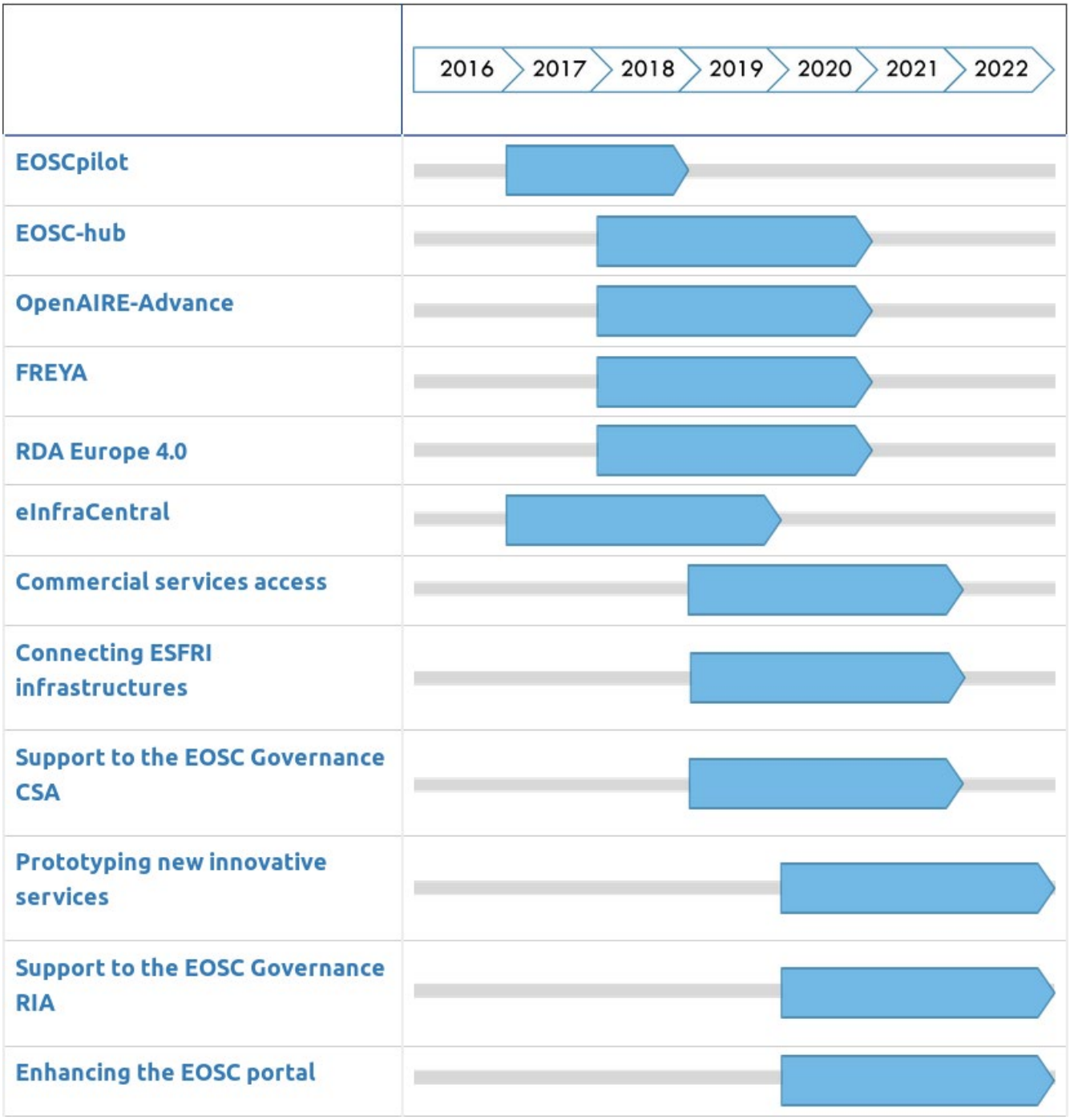


**Figure 2: EOSC timeline, Source: EOSC portal [14]**

The bringing together of e-infrastructure services under the EOSC umbrella is of central importance to DARE. Via EOSC, DARE will be able to provide high-level, intelligent access to relevant 3rd-party services in order to deliver on its objectives.

## 2.3 PRACE Research Infrastructure

One aspect of the DARE project will be to investigate architectural and platform approaches including HPC/MPI computation through the Cloud (e.g. HPC over EGI's federated cloud). It is intended to provide integrated mechanisms to control data-reduction tasks and transfers between HPC and Cloud data-intensive resources at runtime, rapidly freeing expensive HPC storage, decreasing unnecessary idle time and accelerating delivery. For this reason the DARE platform will support the connectivity to HPC resources. One possibility for the future users of the DARE platform is to request computation-intensive HPC resources from PRACE (Partnership for Advanced Computing in Europe).

PRACE[18][19] is an international non-profit association offering a pan-European supercomputing infrastructure. It provides access to computing and data management resources and services for large-scale scientific and engineering applications at the highest performance level. The computer systems and their operations are provided by 5 PRACE members (BSC representing Spain, CINECA representing Italy, ETH Zurich/CSCS representing Switzerland, GCS representing Germany and GENCI representing France). These systems are available to scientists and researchers from academia and industry. Access to the resources however, neets to be officially requested.

- The Call for Proposals for Preparatory Access (http://www.prace-ri.eu/prace-preparatory-access/) is a continuously open call and is intended for short-term access to resources, for code-enabling and porting.
- The Call for Proposals for Project Access (http://www.prace-ri.eu/prace-project-access/) is only open in certain time-frames. It is intended for individual researchers and research groups including multi-national research groups and can be used for 1-3 year production runs.
- For SMEs (Small and Medium Enterprises) the SHAPE programme, offers possibilities to overcome barriers to the adoption of HPC.

However, DARE will not be able to provide access to PRACE resources in a generic manner, because of the resource allocation mechanisms that are in place. Rather it will lay the technical foundation to enable users to exploit HPC resources when they are entitled to use them. In particular, this means that the DARE platform will enable its users to make use of e.g. private, commercial or PRACE resources in a easy and efficient manner.

## 2.4 Docker

Essential element to realize the DARE platform is the ability to isolate and package the DARE software, since it facilitates the deployment and run of developed applications. For this reason and due to its flexibility, reusability, popularity and compatibility with all major cloud provision services, DARE makes use of the Docker container technology.

In general, Docker[20][21] is used to run software packages called "containers". Containers bundle their own tools, libraries and configuration files and can communicate with each other through well-defined channels. All containers are run by a single operating system kernel and are thus more lightweight than virtual machines. They are created from "images" that specify their precise contents. Docker Compose enables the run of multi-container applications by modelling interdependencies and connectivity in a YAML description. Docker Swarm is a clustering and scheduling (orchestration) tool that is delivered together with Docker. It offers features like scalability, interlinking of containers, networking among containers, resource management, load balancing, fault tolerance, failure recovery and log-based monitoring, etc.

While DARE utilized Docker Swarm from the start of the project, it has been decided to replace Docker Swarm with the Kubernetes suite of software. Because of the ease of deployment of Docker Swarm, the fact that it is included in Docker, and because it was already used in some of the software components that comprise the DARE stack, it was well suited for the early stages of development. However, Kubernetes adds a lot of value-added services on top, that DARE plans to make use of in the future.

## 2.5 Kubernetes

In production environments applications usually span multiple containers that need to be deployed across multiple servers. Kubernetes[22] is an open source platform that automates container operations and provides orchestration and management tools allowing its users to schedule application services across clusters of physical or virtual machines. It moreover offers tools to scale those containers in an easy and efficient way and enables to manage their health. It eliminates many of the manual processes involved in deploying and scaling containerized applications.

Kubernetes was first developed by Google, based on years of experience in running containers in production. But also Red Hat was one of the first companies that worked with Google on Kubernetes and has become the 2nd leading contributor to Kubernetes upstream project[23]. Now, Kubernetes counts more than 2,300 contributors and is used by some of the world's most-innovative companies, across a wide range of industries]. In total Kubernetes has a large and rapidly growing ecosystem, that ensures that services, support, and tools are widely available. By now the open source project is hosted by the Cloud Native Computing Foundation (CNCF)[24].

Referring the DARE platform, the planned changeover from Docker Swarm to Kubernetes primary derives for the reason that Kubernetes is supported by a much larger community, offers a more flexible use and additional features. More detailed information is provided in D5.3 Platform Operational and Maintenance.

## 2.6 Big Data Europe Platform

The BDI platform[25] has been developed as part of the Big Data Europe project[26] and has been built to facilitate the installation and development of Big Data tools.

In summary the BDI platform is a customised, cloud-ready and modular integrator platform, bringing together commercial and research, production-ready components for big-data analytics. It offers an easy-to-deploy, easy-to-use and adaptable framework for the execution of big data applications and supports a wide range of common tools for Big Data applications as ready-to-use Docker Compose files. A list of software products included in the BDI is given in table 2.

DARE makes use of the BDE big-data platform as a basis for the integration of pattern matching components and analytics to finally provide a platform for all DARE components to co-exist, interoperate and be readily testable and deployable in Cloud infrastructures. In the scope of WP5 the BDI was installed at the development infrastructure and is accessible by all project partners to serve the development teams of DARE. Also the productive DARE will use the BDI as basis on top of which DARE components will be deployed.

Table 2: Components supported by BDI[27]

| | | |
|---|---|---|
| **Data Processing & Computational Frameworks** | Apache Flink | Open source platform for distributed stream and batch data processing, providing data distribution, communication and fault tolerance for distributed computations over data streams. |
| | Apache Spark | In-memory data processing engine, providing APIs in Java, Python and Scala, with the objective to simplify the programming complexity by introducing the abstraction of Resilient Distributed Datasets (RDDs). |
| | SANSA | SANSA is a collection of open source algorithms for distributed data processing for large-scale RDF Knowledge Graphs. SANSA provides several libraries for RDF Data ingestion, OWL library for RDF/OWL operations, Querying library to support SPARQL, Inference library for rule-based reasoning on RDF/OWL data, and a Machine Learning library for RDF analytics. |

| | | |
|---|---|---|
| **Data storage** | Hadoop | Hadoop is an open-source software framework for storing data and running applications on clusters of commodity hardware. It provides massive storage for any kind of data, enormous processing power and the ability to handle virtually limitless concurrent tasks or jobs. |
| | Hue HDFS File Browser | WebHDFS file browser with Web graphical user interface. |
| | OpenLink Virtuoso | Database management systems for data that is modelled using RDF, backed by an RDBMS. It is available in open-source and commercial editions. RDF data can be queried using SPARQL. Next to RDF, Virtuoso also provides the functionality of a traditional RDBMS. |
| | 4Store | 4store is a database storage and query engine that holds RDF data. It supports both single node and cluster deployment. 4store is available under the GNU General Public Licence, version 3. |
| | Apache Cassandra | Apache Cassandra is a free and open-source distributed database management system designed to handle large amounts of data across many commodity servers, providing high availability with no single point of failure. Cassandra offers robust support for clusters spanning multiple datacenters, with asynchronous masterless replication allowing low latency operations for all clients. |
| | Apache Hive | Data warehouse software facilitating reading, writing and managing large datasets residing in distributed storage using SQL. |

| | | |
|---|---|---|
| **Data acquisition** | Apache Flume | Distributed data acquisition framework used to collect, move or redistribute large amounts of data, based on pipelines that consist of a source, a channel and a sink, setup with simple key value configuration files, either from the filesystem or stored in an Apache Zookeeper node. |
| | Apache Kafka | Distributed publish-subscribe messaging system, scalable without downtime, with messages persisted on disk in a distributed transaction log to prevent data loss and categorized in topics to which is possible to subscribe as a consumer group, where every message is processed once by one member of the consumers' group, and as distinct consumers, where every message will be consumed by every single consumer, distributing thus messages (e.g. data entries) among several databases. |

| | | |
|---|---|---|
| **Semantic components** | FOX | Federated knOwledge eXtraction Framework integrating the Linked Data Cloud and using the diversity of NLP algorithms to extract RDF triples of high accuracy out of NL, while integrating and merging the results of Named Entity Recognition tools. |
| | GeoTriples | Semi-automated tool transforming geospatial data into RDF graphs with the use of state-of-the-art vocabularies like GeoSPARQL and stSPARQL, without being tightly coupled to a specific vocabulary, and publishing them as Linked Open Geospatial Data, by extending the R2RML mapping language to the specificities of geospatial data. |
| | Silk | Open source framework, based on the Linked Data paradigm, for integrating heterogeneous data sources, generating links between related data items within different Linked Data sources and setting RDF links from a data source to another one on the Web, while applying data transformations to structured data sources via the declarative Silk – Link Specification Language (Silk-LSL), the RDF path language, the SPARQL protocol for local and remote SPARQL endpoints and the graphical user interface of Silk Workbench. |
| | SEMAGROW engine | Algorithmically sophisticated and well-engineered Query Federation engine, combining and cross-indexing public data, regardless of their size, update rate, and schema, while offering a single SPARQL endpoint and allowing full flexibility in terms of metadata details. |
| | Sextant | Web application for visualizing, exploring and interacting with time-evolving linked geospatial data, with user-friendly interface allowing domain experts and non-experts to use its semantic web technologies and to create thematic maps by combining geospatial and temporal information from various heterogeneous data sources ranging from standard SPARQL endpoints, to SPARQL endpoints |

| | | |
|---|---|---|
| | | following the standard GeoSPARQL defined by the Open Geospatial Consortium (OGC), or well-adopted geospatial file formats, like KML, GML and GeoTIFF. |
| | Strabon | Semantic spatiotemporal RDF store, used to store linked geospatial data that changes over time and to pose queries using two popular extensions of SPARQL, enabling thus the serialization of geometric objects in OGC standards WKT and GML and offering spatial and temporal selections and joins, a rich set of spatial functions similar to those offered by geospatial relational database systems and support for multiple Coordinate Reference Systems. |
| | UnifiedViews | Open source platform for distributed stream and batch data processing, providing data distribution, communication and fault tolerance for distributed computations over data streams. |
| | Ontario | Query processor for Data Lakes, it allows to query heterogeneous data (e.g., csv, json, rdf) while they are in their original formats. Ontario is a realization of the so-called Semantic Data Lake, where Semantic Web techniques, e.g., SPARQL, RDF mapping languages, etc. are used underneath the hoods to build the “virtual” data integration process. |

## 2.7 Community Infrastructures

Since DARE will make use of external data, data management and the transfer of data between internal and external infrastructures play an important role in the DARE project. The main source of external datasets are the data repositories from the communities that may use the DARE platform in the future.

In the scope of the DARE project, primary the two use-case domains solid earth science (EPOS) and climate (IS-ENES/Climate4Impact) are considered. How the individual data repositories are made available through the DARE platform differs. Some could be accessed through the VERCE platform[28] or rather C4I platform[29]. Others could be made available through specific webservices of the databases or by direct queries. More details and a list of all main datasets, both internal and external, are offered in D-SA1.2-Data Management Plan.

# 3 Deployment Approach

By the end of the project the DARE platform will build on top of the current and next generation e-Infrastructures deployed in Europe. However, to enable a quickly available development environment and at the same time ensure a smooth deployment of the DARE platform WP5 will realize infrastructures on public and private resources.

For the development and testing of relevant components, WP5 set up a software development platform/pre-release testbed for DARE’s software products which is continuously adjusted to meet

the requirements of the development teams. Currently this environment is available on GRNETs infrastructure okeanos and reachable by the DARE developers for direct use.

## 3.1 Software Development Cycle

The DARE consortium will bring together several software assets like dispel4py, S-ProvFlow, Exareme and Semagrow whose integration, along with registries and other technologies, will contribute to the DARE platform. These assets are in the process of being integrated, extended and strengthened. Concerning the development of these assets, a modern agile development and integration methodology is applied.

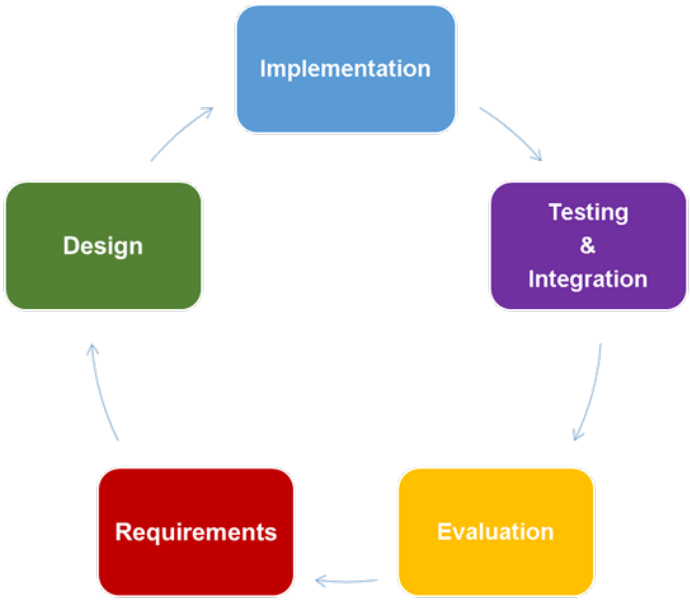


Figure 3: Software Development Cycle

In this context, partners decided upon GitLab to be used as code repository and development platform (https://gitlab.com/project-dare). GitLab enables a flexible and joint development taking into account the whole development cycle.

## 3.2 Deployment Pipeline

As soon as any software components have passed the Software Development Cycle, they enter the Deployment Pipeline.

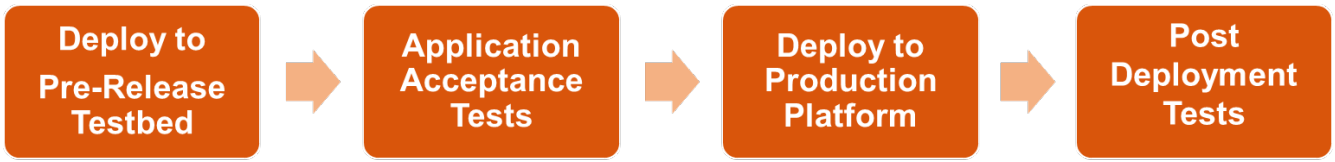


Figure 4: Deployment Pipeline

As shown in figure 4, this pipeline comprises four steps. In the first step all needed Docker and Kubernetes files will be created in order to afterword's integrate and deploy the assets to the internal Pre-Release Testbed that is currently located at GRNETs cloud service okeanos. In the second step the

Application Acceptance Tests will be performed to ensure that the applications of the involved use-cases work and meet the expectations of the domain specific communities. In the development of these Application Tests, software developers as well as representatives from the use cases need to be involved. After the successful conclusion of this phase in the third step the components will be published to the final production platform (expected EOSC). In the fourth and last step the functionality, performance and safety of the Dare platform will be tested and proved.

In case of error messages or if certain problems become apparent during the deployment process, issues will be reported via GitLab in order to inform the developers. As soon as occurred problems have been sold the software assets again enter the deployment pipeline.

## 3.3 Deployment Schedule

Regarding the deployment of the DARE platform, we distinguish between the internal and the public deployment. The internal deployment comprises the integration of all components to the Pre-Release Testbed and the application acceptance tests. This internal deployment happens continuously to ensure a rapid progress. The public deployment furthermore includes the deployment to a European e-infrastructure (expected EOSC) and corresponding tests. The public deployment for the first version of the DARE platform is planned for 12/2019, at the end of the second project year. The deployment of the final platform will take place at the end of the DARE project in 12/2020.

**Table 3: Deployment Schedule**

| Date | 12/2018 | 12/2019 | 12/2020 |
|---|---|---|---|
| Deployment to Pre-Release Testbed (internal) | continuously | | |
| Deployment to Production Platform (public) | | x | x |

# 4 Usage Statistics of Dare Platform

Usage statistics are an important tool for monitoring. On the one hand they present a history of events that occur over a specific time frame. On the other hand they help to analyse how future users will use the DARE platform. Since the release of the public platform is scheduled after the second project year this deliverable could only provide basic information referring the usage of GitLab.

**Table 4: Gitlab-Statistics [30]**

| Members | 27 |
|---|---|
| Subgroups and Projects | 11 |
| all Issues/ open/ closed | 98/ 48/ 50 |
| all Milestones/ open/ closed | 2/ 1/ 2 |
| all Merged Requests /open/closed | 2/ 0/ 2 |

D5.2-Platform Infrastructure, Usage & Deployment II (M36) will provide the relevant usage statistics of the publicly deployed DARE platform. These are expected to be derived from selected Monitoring tools (D5.3).

# 5 Conclusions

The present deliverable described the underlying infrastructure of the DARE platform, including private computational resources, resources provided by the European Open Science Clouds and further components contributing to the platform. It has defined the process for transferring assessed development versions of platform components to the publicly deployed DARE platform, integrated to the European e-infrastructure, and finally presented preliminary usage statistics. These data is not particularly representative, since it only reflects a short time period and moreover not yet considers the usage statistics of the publicly deployed DARE platform.

This deliverable can be considered as living document. Described platform components as well as the deployment strategy may evolve taking into account new information.

In addition to this document, D5.3-Operational Requirements and Guidelines I will present best practices for managing and maintaining the deployment and provide information about operating the DARE platform. It will e.g. cover the topics monitoring, security policies, accounting and user management, etc.

# 6 References

[1] https://www.openstack.org/software/
[2] https://www.synnefo.org/about/
[3] https://www.egi.eu/
[4] https://eudat.eu/
[5] https://www.egi.eu/federation/egi-federated-cloud/
[6] https://www.egi.eu/services/
[7] https://wiki.egi.eu/wiki/VT_GPGPU
[8] https://www.egi.eu/wp-content/uploads/2016/08/Inspired-issue-10.pdf
[9] https://wiki.egi.eu/wiki/Federated_Cloud_GPGPU
[10] https://wiki.egi.eu/wiki/Federated_Cloud_infrastructure_status
[11] https://eudat.eu/what-eudat
[12] https://eudat.eu/catalogue
[13] https://eosc-launch.eu/home/
[14] https://eosc-portal.eu/about/eosc
[15] https://www.eosc-portal.eu/
[16] https://www.eoscpilot.eu/about-eoscpilot
[17] https://www.eosc-hub.eu/about-us
[18] http://www.prace-ri.eu/prace-in-a-few-words/
[19] http://www.prace-ri.eu/
[20] https://www.docker.com/why-docker
[21] https://opensource.com/resources/what-docker
[22] https://kubernetes.io/docs/concepts/overview/what-is-kubernetes/
[23] http://stackalytics.com/?project_type=kubernetes-group&metric=commits
[24] https://www.cncf.io/
[25] https://www.big-data-europe.eu/platform/
[26] https://www.big-data-europe.eu/about/
[27] https://www.big-data-europe.eu/bdi-components/
[28] https://portal.verce.eu/home
[29]https://climate4impact.eu/impactportal/general/index.jsp
[30] https://gitlab.com/project-dare