
# ESCAPE: a multi-science data infrastructure for the 2020s

**Simone Campana, Patrick Fuhrmann, Giovanni Lamanna**

**ISGC, Taipei, Taiwan**

**4 April 2019**

ESCAPE - The European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement n° 824064.

# Disclaimer

Project just started. No results yet.


Simone.Campana@cern.ch
Funded by the European Union's
Horizon 2020 - Grant N° 824064

# About the call

H2020-INFRAEOSC-04-2018 call

Clusters to ensure the connection of the ESFRI RIs with EOSC (and the construction of EOSC)

What is

- ESFRI ?
- EOSC ?
- Cluster ?


Simone.Campana@cern.ch
Funded by the European Union's Horizon 2020 - Grant N° 824064

# ESFRI: Strategy Forum for RI's

## ESFRI's mandate

- to support a coherent and strategy-led approach to policy making on research infrastructures in Europe
- to facilitate multilateral initiatives leading to a better use and development of research infrastructures
- to establish a European Roadmap for research infrastructures (new and major upgrades, pan-European interest) for the next 10-20 years, stimulate the implementation of these facilities, and update the roadmap as needed
- to follow-up on implementation of ongoing ESFRI projects after a comprehensive assessment, as well as the prioritisation of infrastructure projects listed in the ESFRI Roadmap

The 2016 roadmap contains details of 21 ESFRI Projects — including six new projects, and 29 ESFRI landmarks. These landmarks are RIs that reached the implementation phase before the end of 2015.

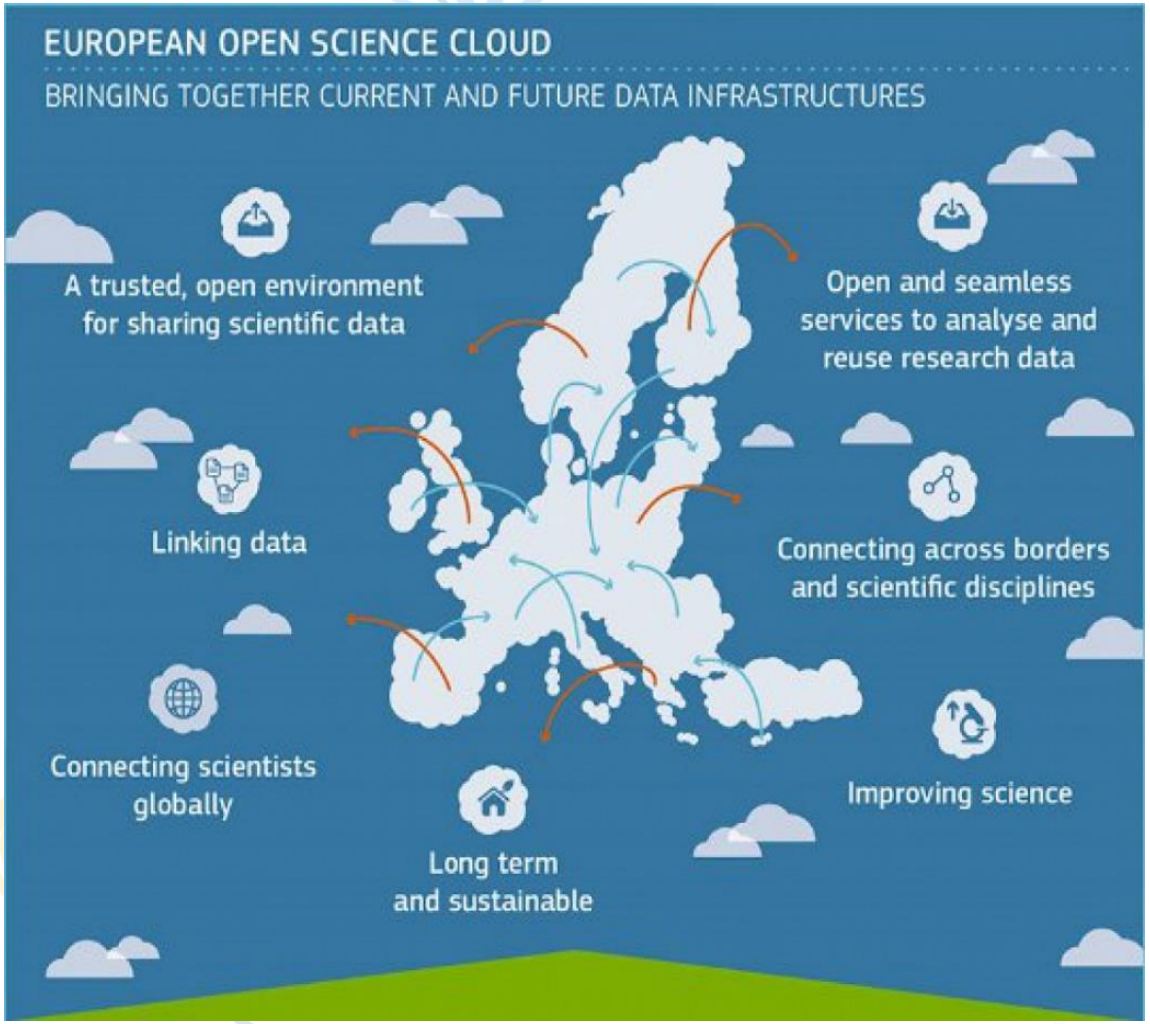

Funded by the European Union's Horizon 2020 - Grant N° 824064

# About EOSC


Funded by the European Union's Horizon 2020 - Grant N° 824064

# About EOSC

- Bridging todays fragmented and ad-hoc solutions, towards a **federation of data infrastructures**
- **FAIR data and services** for data storage, management, analysis and re-use **across borders and disciplines**
- Added value for **data-driven science**, reproducible science, interdisciplinary research, digital innovation



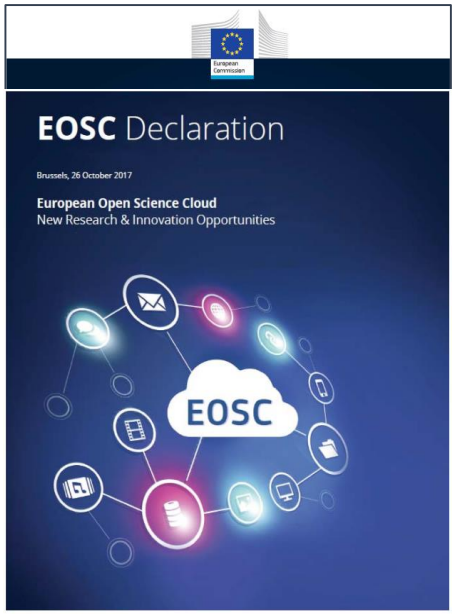


Funded by the European Union's Horizon 2020 - Grant N° 824064

# About EOSC

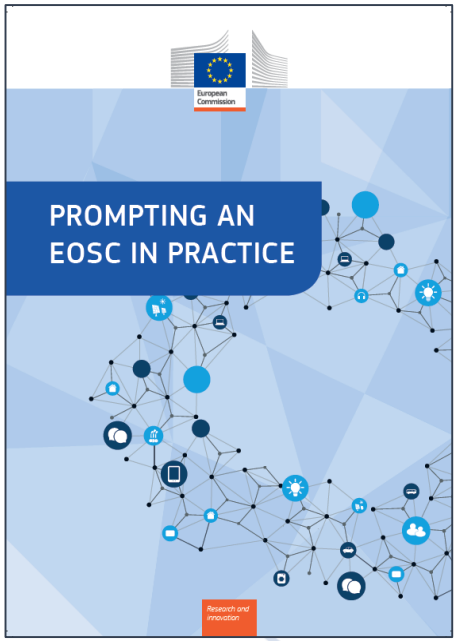
Some readings…

**EOSC Summit of 12 June 2017**
https://ec.europa.eu/research/openscience/pdf/eosc_declaration.pdf



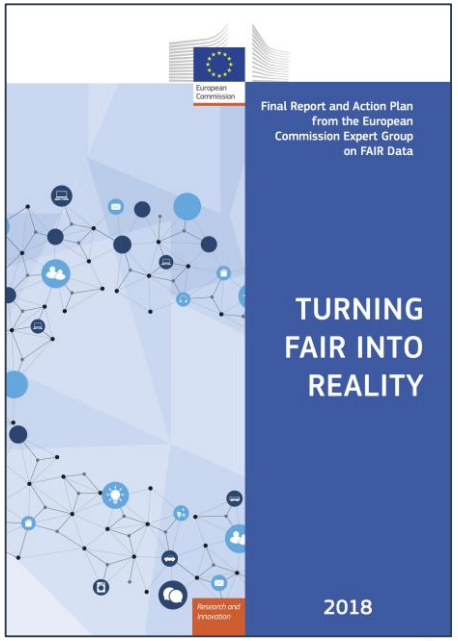

**2nd HLEG on EOSC**

https://ec.europa.eu/info/events/2nd-eosc-summit-2018-jun-11_en




**FAIR Data Expert Group**

https://doi.org/10.2777/1524





Funded by the European Union's Horizon 2020 - Grant N° 824064

# Timeline of EOSC


Vision for the European Open Science Cloud

INFRAEOSC 2018-2019 Calls

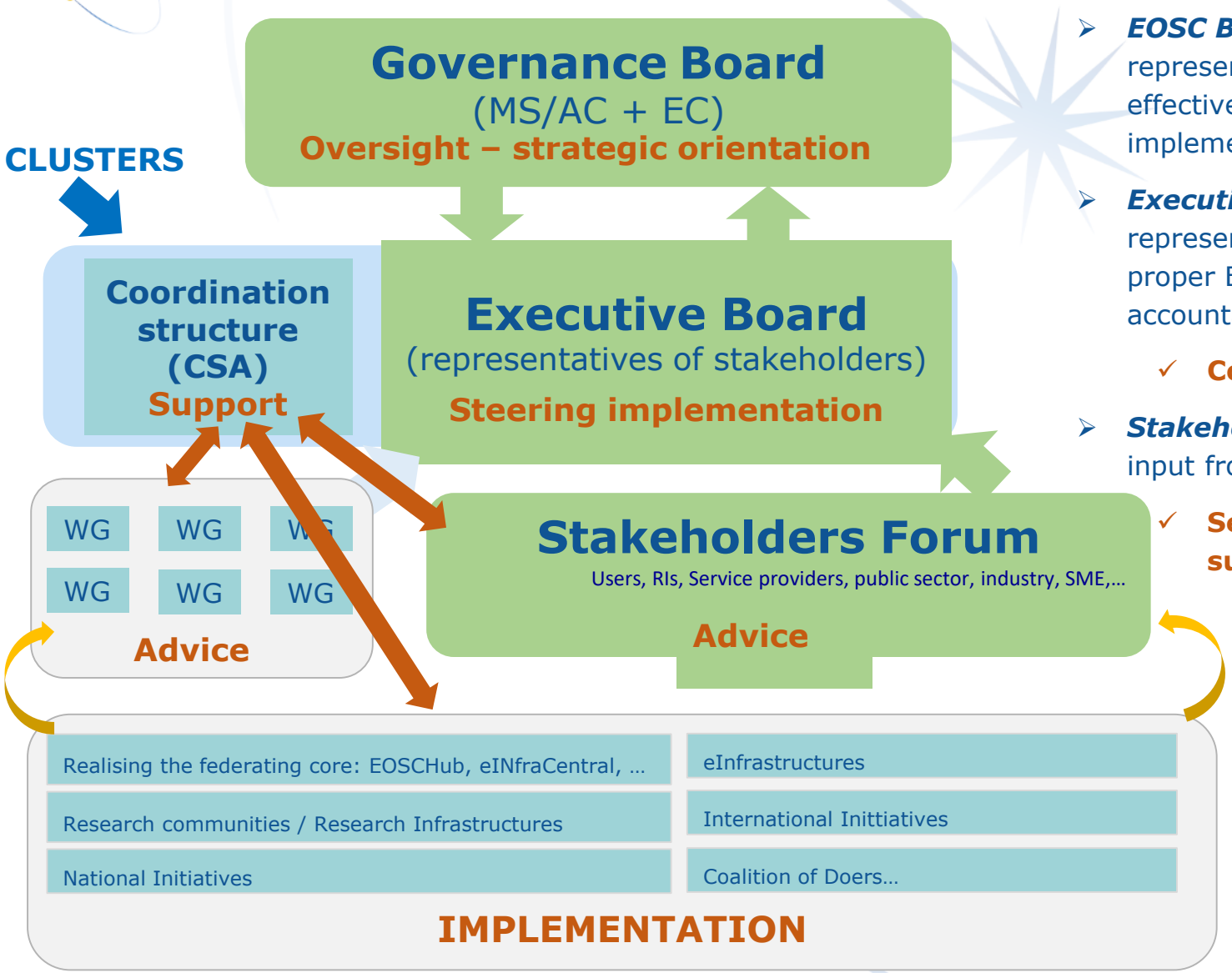
Federated model with 6 action lines

Set up of EOSC governance framework by end 2018

- a. Architecture
- b. Data
- c. Services
- d. Access & Interface
- e. Rules
- f. Governance


Funded by the European Union's Horizon 2020 - Grant N° 824064

# EOSC Governance Structure



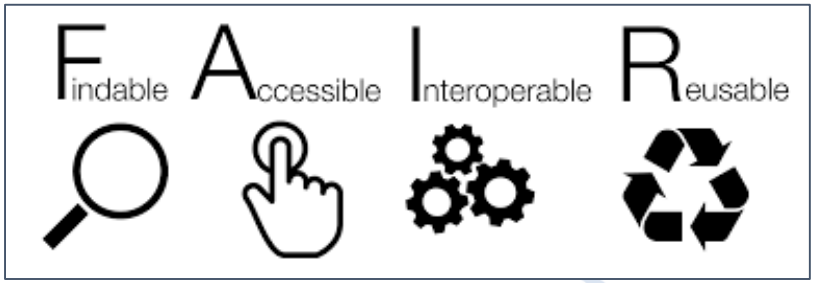

- ***EOSC Board*** of MS/AC and EC representatives to ensure effective supervision of EOSC implementation
- ***Executive Board*** of stakeholder representatives to help ensure proper EOSC implementation and accountability
  - ✓ **Commission expert group**
- ***Stakeholder Forum*** to provide input from a wide range of actors
  - ✓ **Self-organised with EC support**

Funded by the European Union's Horizon 2020 - Grant N° 824064

# Impact of the call

## Expected impact:

- Improve access to data and tools leading to new insights and innovation
- Facilitate access of researchers to data and resources for data driven science.
- Create a cross-border open innovation environment.
- Rise the efficiency and productivity of researchers through open data services and infrastructures for discovering, accessing, and reusing data.
- Foster the establishment of global standards.
- Develop synergies and complementarity between involved research infrastructures.
- Adopt common approaches to the data management for economies of scale.





Funded by the European Union's Horizon 2020 - Grant N° 824064

# About Clusters


Funded by the European Union's Horizon 2020 - Grant N° 824064

# EOSC for big science

A cluster action of Big-Science ESFRI RIs for setting up EOSC, implies technical and policy challenges.

*(As per the European Commission "EOSC Declaration")*

- EOSC as a data infrastructure commons serving the needs of scientists, providing functions delegated to community level, federating resources.
- Researchers should contribute to define the main common functionalities needed by their own community.
- A continuous dialogue to build trust and agreements among funders, scientists and service providers is necessary for sustainability.
- Data Sharing and Data Stewardship are critical issues for the next generation ESCAPE RIs


Funded by the European Union's Horizon 2020 - Grant N° 824064

# Five EOSC Clusters

- EOSC-LIFE: Life science RIs
  - Providing an open collaborative space for digital biology in Europe.
  - EOSC, Biological Medical Research Infrastructures, BMS RI, ESFR, Cloud, Data Resources, GDPR EOSC-Life brings together the 13 Biological and Medical ESFRI research infrastructures (BMS RIs) to create an open collaborative space for digital biology.
- ENVRI-FAIR: Environmental RIs
  - ENVironmental Research Infrastructures building Fair services Accessible for society, Innovation and Research.
- PANOSC: Photon and Neutron sources Ris
  - PaNOSC will contribute to the construction and development of the EOSC, an ecosystem allowing universal and cross disciplinary open access to data through a single access point, for researchers in all scientific fields. The project will work closely with the national photon and neutron sources in Europe in order to develop common policies, strategies, and solutions in the area of FAIR data policy, data management and data services.
- SSHOC: Social Sciences and Humanities
  - The project aims to provide a full-fledged Social Sciences and Humanities Open Cloud (SSHOC) where data, tools, and training are available and accessible for users of SSH data.
- ESCAPE: Hep and Astronomy
  - See below


Funded by the European Union's Horizon 2020 - Grant N° 824064

# ESCAPE

European Science Cluster of Astronomy &
Particle physics ESFRI research Infrastructures



Simone.Campana@cern.ch

Funded by the European Union's
Horizon 2020 - Grant N° 824064

# ESCAPE in a nutshell

**ESCAPE** convenes a large scientific community

- **31** partners (including 2 SMEs)
- **7** ESFRI projects & landmarks: CTA, ELT, EST, FAIR, HL-LHC, KM3NeT, SKA
- **2** pan-European International Organizations: CERN, ESO (with their world-class established infrastructures, experiments and observatories).
- **4** supporting ERA-NET initiatives: HEP (CERN), NuPECC, ASTRONET, APPEC
- **1** involved initiative/infrastructure: EURO-VO (Virtual Observatory)
- **2** European research infrastructures: EGO and JIVE-ERIC
- Budget: **15.98 M€**
- Started: **1/2/2019**
- Duration: **42** months (end date 31/7/2022)
- Coordinator: **CNRS (**Centre national de la recherche scientifique**)**

*Home page: https://escape2020.eu ; Twitter: @ESCAPE_EU*

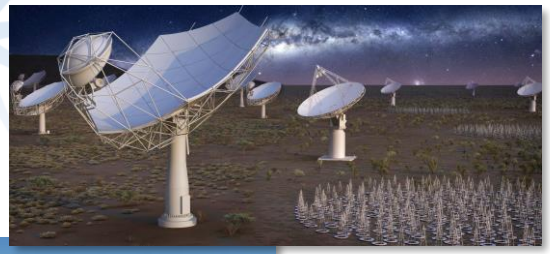

Funded by the European Union's Horizon 2020 - Grant N° 824064

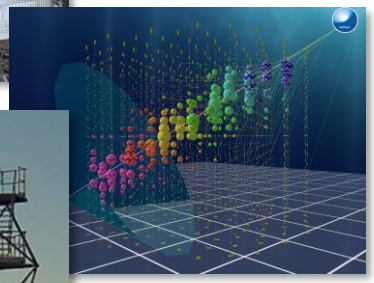
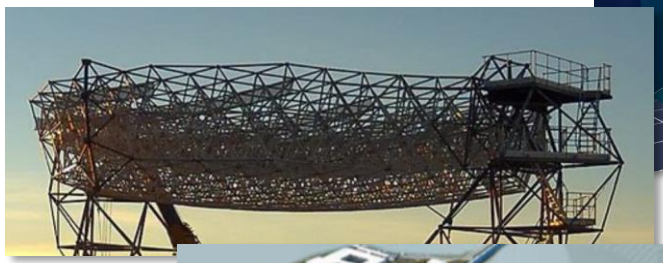


Funded by the European Union's
Horizon 2020 - Grant N° 824064

# ESCAPE Science

- Multi Messenger Astronomy
  - Radio
    - SKA (Scare Kilometre Array)
    - JIVE VLBI (Very large Baseline Instrument)
  - Visible Light
    - European Extreme Large Telescope (ELT)
    - European Solar Telescope (EST)
  - Gamma Rays
    - CTA
  - Cosmic Rays: Neutrinos
    - KM3Net
  - Gravitational Waves
    - EGO-VIRGO
- High Energy Physics
  - HL-LHC
    - High Energy Particle
  - FAIR
    - High density exotic matter physics





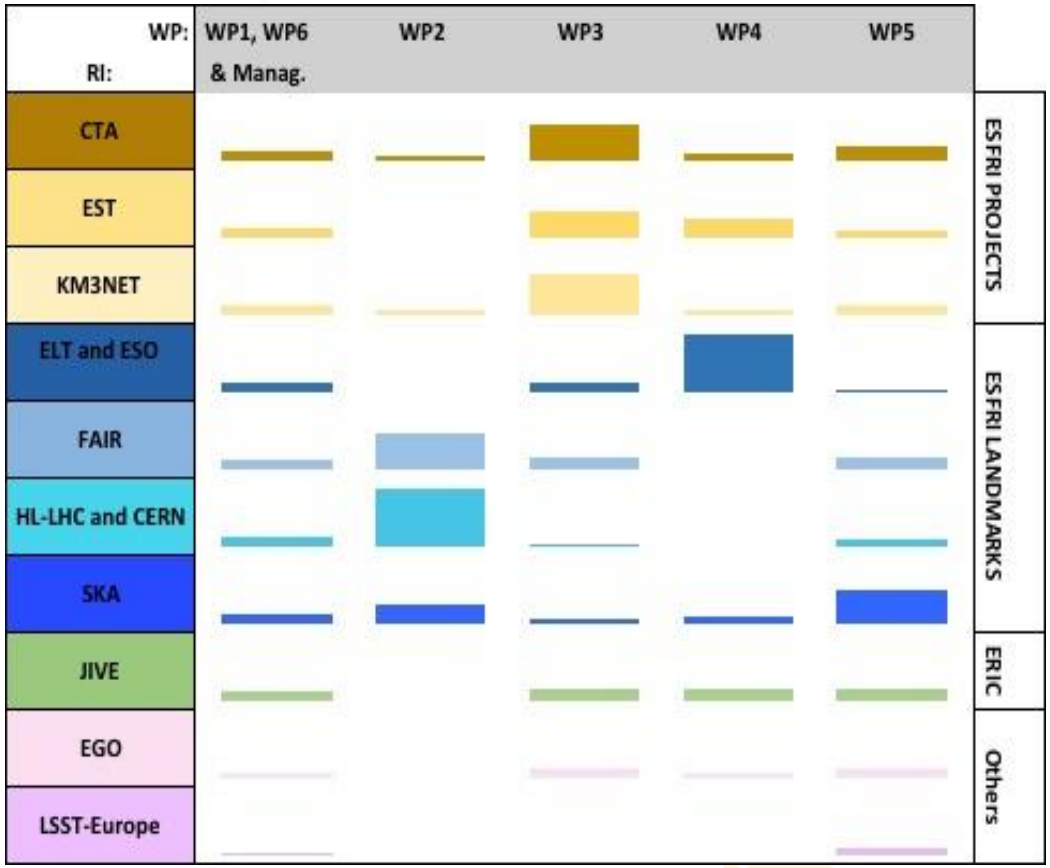
Funded by the European Union's Horizon 2020 - Grant N° 824064

# The work package structure

- WP1 MIND. Leader: Giovanni Lamanna, LAPP-CNRS
  - Management and policy
- WP2 DIOS. Leader: Simone Campana, CERN
  - Contribute to the federation of global EOSC resources through an implementation of the Data-Lake concept (evolution of WLCG and other ESFRI RIs computing models) to manage extremely large volumes of data up to the multi-exabyte scale
- WP3 OSSR. Leader: Kay Graf, FAU
  - Support for "scientific software" as a major component of the ESFR-RI “data” to be stored and displayed in EOSC via dedicated community-based catalogues. Implementation of a community-based approach for the continuous development of shared software and for training of researchers and data scientists.
- WP4 CEVO. Leader: Mark Allen, CDS-CNRS
  - Extend FAIR standards, methods, tools of the Virtual Observatory to a broader scientific context; demonstrate EOSC's ability to include existing platforms.
- WP5 ESAP. Leader: Michiel van Haarlem, ASTRON-NOW (Deputy : Zheng Meyer)
  - Implementation of scientific analysis platforms enabling EOSC researchers to organize data collections, analyse them, access ESFRI's software tools, and provide their own customized workflows.
- WP6 ECO. Leader: Stephen Serjeant, Oxford Open University
  - Citizen Science, Open Science et Communication


Funded by the European Union’s Horizon 2020 - Grant N° 824064

# ESCAPE Work Program




The allocated staff effort is proportional to the respective boxes' surface areas.

## An optimal matrix:

- Some clear priorities per each RI
- RIs' use-cases in almost all WPs
- Sub-sets of RIs driving a WP
- All RIs involved in the EOSC support


Funded by the European Union's Horizon 2020 - Grant N° 824064

# Stepping to WP2 :

# Data Infrastructure for Open Science (DIOS)

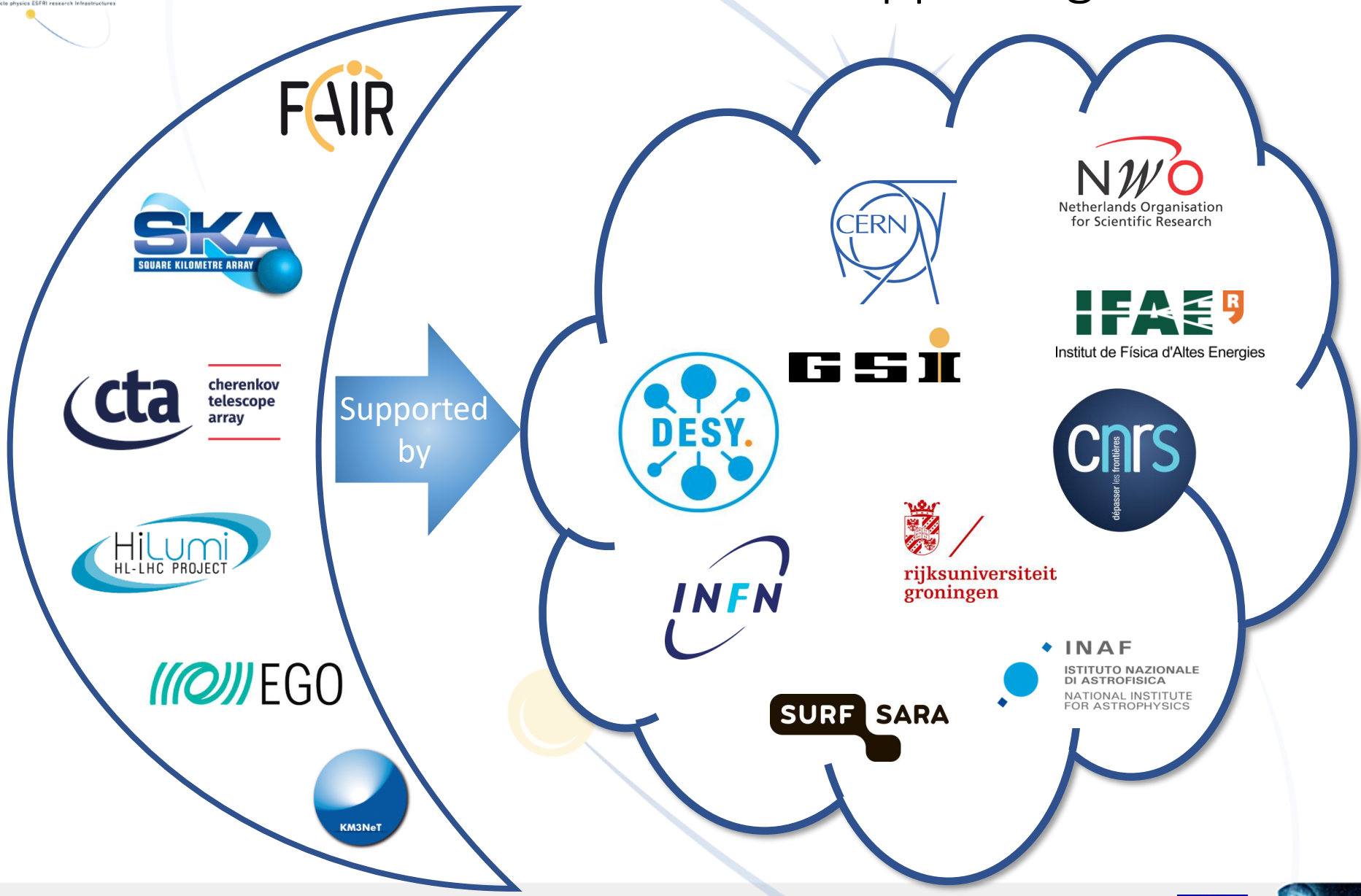

Funded by the European Union's Horizon 2020 - Grant N° 824064

# Data Infrastructure for Open Science (DIOS)

- Goal: design, implement and operate a cloud of data services for open access and open science at the Exabyte scale
- The backbone of the Data Lake are well experienced large national data centers supporting the ESFRIs in ESCAPE
- The data lake will serve as underlying data infrastructure to manage and serve data to the user communities
- This solution will be proposed as key component of the future EOSC framework, supporting FAIR principles


Funded by the European Union's Horizon 2020 - Grant N° 824064

# Involved sciences and supporting RI






Simone.Campana@cern.ch

Funded by the European Union's Horizon 2020 - Grant N° 824064

# WP2- specific objectives

Prototype a reliable and scalable **federated data infrastructure.** Mapping FAI(R)

- Stores and organizes scientific data (**F**indable) and enables the provisioning of data processing (**A**ccessible)
- Enables sciences to build open data repositories (**I**nteroperable)
- In general, supports the world-leading data challenges of the Research Infrastructures in ESCAPE


Funded by the European Union's Horizon 2020 - Grant N° 824064

# WP2- specific objectives

Ensure long term **data preservation** (**R**eproducible) at the infrastructure level

- Archiving of data in certified repositories and capabilities to retrieve the data in the long term
- Complementing other work packages dealing with software, environment and provenance


Funded by the European Union's Horizon 2020 - Grant N° 824064

# WP2- specific objectives

The Data Lake development leverages **collaboration** with and **integrates** the work from previous and ongoing frameworks

- EU projects such as EOSC-hub and XDC
- Initiatives at the Research Infrastructures
- Ongoing work from e.g. GEANT, PRACE


Funded by the European Union's Horizon 2020 - Grant N° 824064

# WP2- specific objectives

**Computing** Interface and Scalability

- Computing capacity available inside and outside the data lake (Grid, Cloud, HPC, volunteer computing)
- Need to integrate compute (and data) resources which are not part of the lake
- Compute-data locality not guaranteed. Need to offer a reliable content-delivery service


Funded by the European Union's Horizon 2020 - Grant N° 824064

# WP2- specific objectives

**Industrial** and **Commercial** involvement

- Commercial storage can be added to the data lake as cache or for resiliency
- Commercial computing can be integrated with the data lake as extra processing capacity
- In the data lake model, for both compute and storage there is no lock-in to the vendor
- WP2 will validate the use of commercial cloud to store and process scientific data


Funded by the European Union's Horizon 2020 - Grant N° 824064

# WP2 objective

Create a cloud of data services, often referred to as a "*Data Lake*" by building on and integrating existing work from a variety of areas:

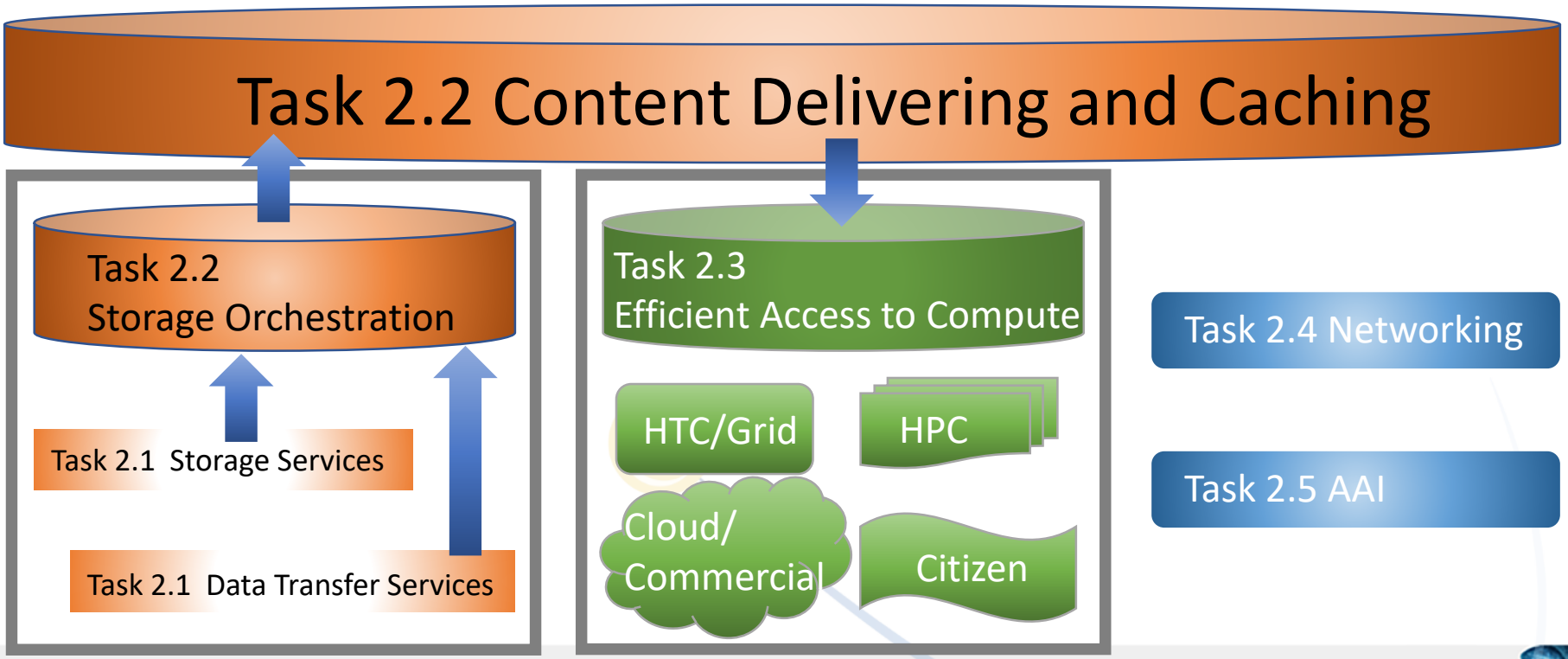
- Research Infrastructures
- previous EU projects, INDIGO, DEEP, XDC,...
- state of the art solutions in the appropriate areas

Collaborates with ongoing work from GEANT, PRACE, and other proposed H2020 projects specifically addressing the European Open Science Cloud


Funded by the European Union's Horizon 2020 - Grant N° 824064

# Structure of ESCAPE WP2

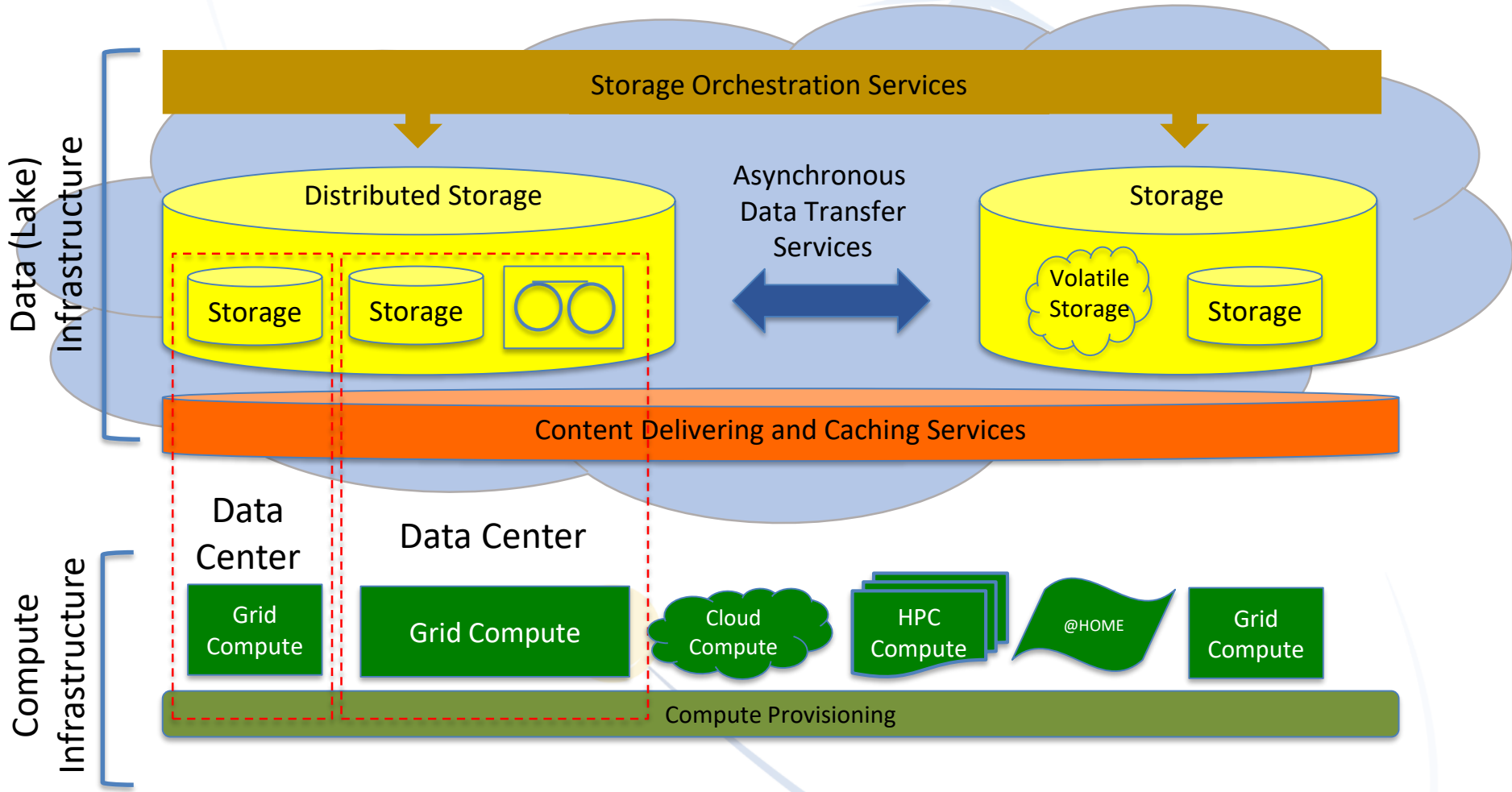
- Task 2.1 Data Lake Infrastructure and Federation Services. CERN (Xavier Espinal)
- Task 2.2 Data Lake orchestration service. DESY (Patrick Fuhrmann)
- Task 2.3 Integration with Compute Services. NOW-I-ASTRON
- Task 2.4 Networking. SKAO (Rosie Bolton)
- Task 2.5 Authentication and Authorization. INFN (Andrea Ceccanti)





Funded by the European Union's Horizon 2020 - Grant N° 824064

# Data Lake strawman



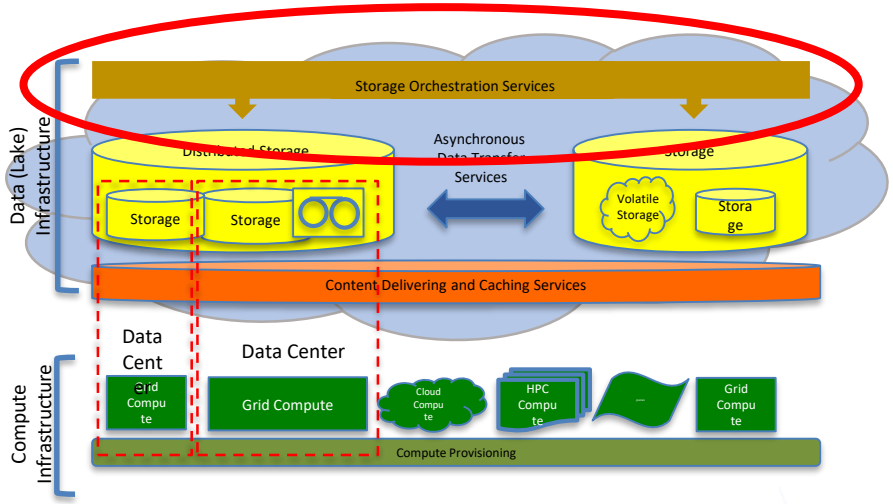


Funded by the European Union's Horizon 2020 - Grant N° 824064

# Task 2.2: Orchestration Service

Implement a system managing scientific user policies while optimizing the service provider costs.

- Replication policies, access policies
- QoS: optimization between redundancy, performance and cost
- Data lifetimes and lifecycles: dynamic replication, deletion, change of QoS




| Partner: | CERN | **DESY** | GSI | SKAO | NWO-I-ASTRON | CNRS-LAPP | CNRS-CCIN2P3 | IFAE | SURFsara |
|---|---|---|---|---|---|---|---|---|---|


Funded by the European Union's Horizon 2020 - Grant N° 824064

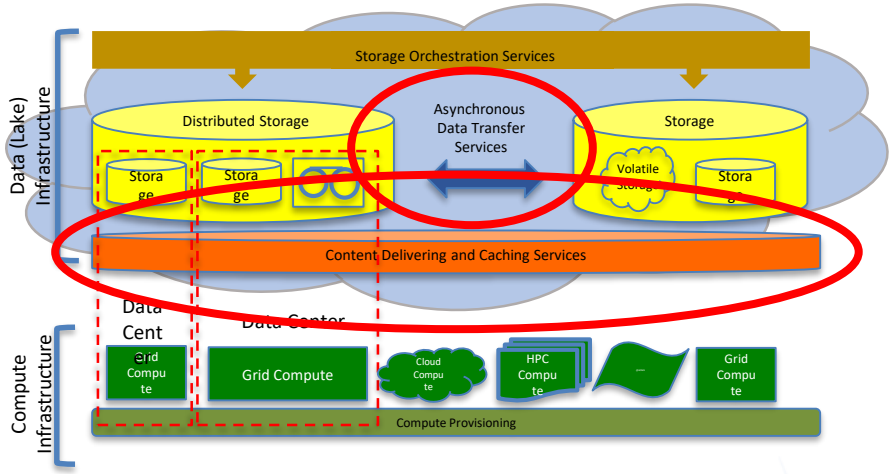
# Task 2.3: Integration with Compute

## Does not focus of provisioning compute resources.

Focuses instead on serving data to large scale processing centers
Processing capacity might be not co-located with data.

Processing capacity might be not co-located with data

- Data Transfer Services
- Caching and latency hiding services (Content Deliver Network)
- Compute services will be heterogeneous: Grid, HPC, Cloud (including commercial)




| Partner: | CERN | INFN | GSI | NWO-I-ASTRON | CNRS-LAPP |
|---|---|---|---|---|---|


Funded by the European Union's Horizon 2020 - Grant N° 824064

# Task 2.4: Networking

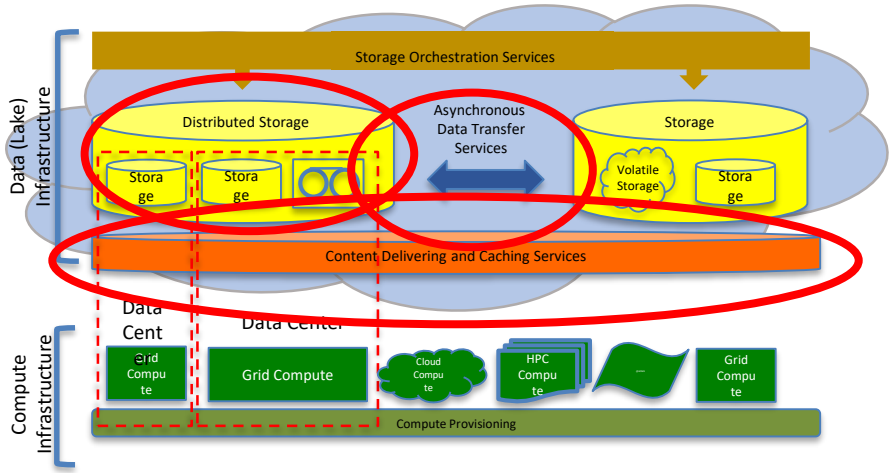
Wide Area Network is a key component in the Data Lake model

Task 2.4 develops the capability to provide high capacity networking between data centers to enable traffic management

Leverages work done in WLCG and GEANT.

Applies to all scenarios in Task 2.3:

- Asynchronous Data Transfer of large data volumes
- Content Deliver Network for processing
- Integration of commercial compute resources




| Partner: | CERN | GSI | SKAO |
|---|---|---|---|



Funded by the European Union's Horizon 2020 - Grant N° 824064

# Task 2.5: Authentication and Authorization

Integrates solutions from different projects/activities to build a federated storage infrastructure

- provide the appropriate level of granularity of authentication and access control to manage and protect data
- provide the means by which to enable open access once data is released to the broader community

Heterogeneous authentication mechanisms, management of memberships and policies, controlled delegation, leverage off-the-shelf libraries and components

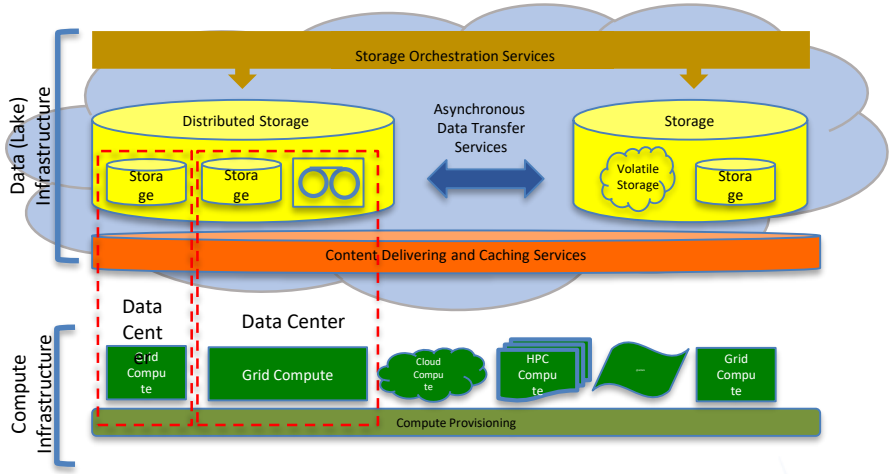
| Partner: | INFN | SKAO |
|---|---|---|


Funded by the European Union's Horizon 2020 - Grant N° 824064

# Task 2.1 puts all this together

Builds the prototype which federates the storage of several of the data centers that support the main science communities in the project

- Store scientific data of the Research Infrastructures with the policies by them defined
- Provide the needed monitoring and analytics tools
- Certifies the data centers for bit preservation




| Partner: | CERN | DESY | FAIR GMBH | GSI | INFN | NWO-I-Nikhef | RUG |
|---|---|---|---|---|---|---|---|


Funded by the European Union's Horizon 2020 - Grant N° 824064

# Milestones and Deliverables

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1 Aug 2019 | M2.1 | First WP2 workshop on the initial design and goals of the first pilot data lake, prepare D2.1 | WP2 | M6 | Workshop summary report | | |
| 1 Oct 2019 | D2.1 | Implementation plan and design of pilot; R&D questions/metrics that will be addressed in the pilot and prototype. (R) | | | 2.1, 2.2, 2.3, 2.4, 2.5, 2.6 | CERN | 8 |
| 1 Aug 2020 | M2.2 | Initial pilot data lake with at least 3 core data centres | WP2 | M18 | Progress report; Active monitoring of activity (web site) | | |
| 1 Dec 2020 | M2.3 | Second WP2 workshop to analyse the performance of the pilot, prepare D2.2 | WP2 | M22 | Workshop summary report | | |
| 1 Feb 2021 | D2.2 | Assessment and analysis of the performance of the first pilot data lake (R) | | | 2.1, 2.2, 2.6 | SKAO | 24 |
| 1 Feb 2021 | M2.4 | Expanded prototype – more data centres including 3rd party centres, demonstrate integrated data management tools, verify RI data accessibility from compute platforms including commercial clouds | WP2 | M24 | Review of D2.2; Monitoring web site | | |
| 1 Aug 2021 | M2.5 | Extension of the data lake to efficiently serve data to external compute resources providers | WP2 | M30 | Progress report; Monitoring web site | | |
| 1 Oct 2021 | M2.6 | ISO 16363 certification process underway in core data centres | WP2 | M32 | Progress report; core data centres finished self-certification audit and ready to submit to external audit. | | |
| 1 Apr 2022 | M2.7 | Third WP2 workshop to review performance of the full prototypes, and to explore future directions, prepare D2.3 | WP2 | M38 | Workshop summary report | | |
| 1 Jun 2022 | D2.3 | Final assessment and analysis of the full prototype, outlook for further development and deployment towards full production services within EOSC (R) | | | 2.1, 2.2, 2.4, 2.6 | CERN | 40 |



Funded by the European Union's Horizon 2020 - Grant N° 824064

# Deliverables

- Oct 2019:
  - Implementation plan and the design of the of the pilot
- Aug 2020 (Milestone)
  - Initial Data Lake with at least 3 centers.
- Feb 2021:
  - Analysis and assessment of the first version of the pilot data lake.
- Feb 2021:
  - Extended prototype with more centers and tools
- 1 Jun 2022:
  - Final analysis and assessment of the full prototype.


Funded by the European Union's Horizon 2020 - Grant N° 824064

# First practical steps

Focus on our first milestone (and deliverable): discuss and design the first implementation of the data lake

- Evolve the strawman into an architecture
- Which components to focus on in the initial phase
- Sciences drive the needs (and needs drive the design)
- This will be an initial design. Flexibility (both in design and components) is one key aspect


Funded by the European Union's Horizon 2020 - Grant N° 824064

# END


Funded by the European Union's Horizon 2020 - Grant N° 824064