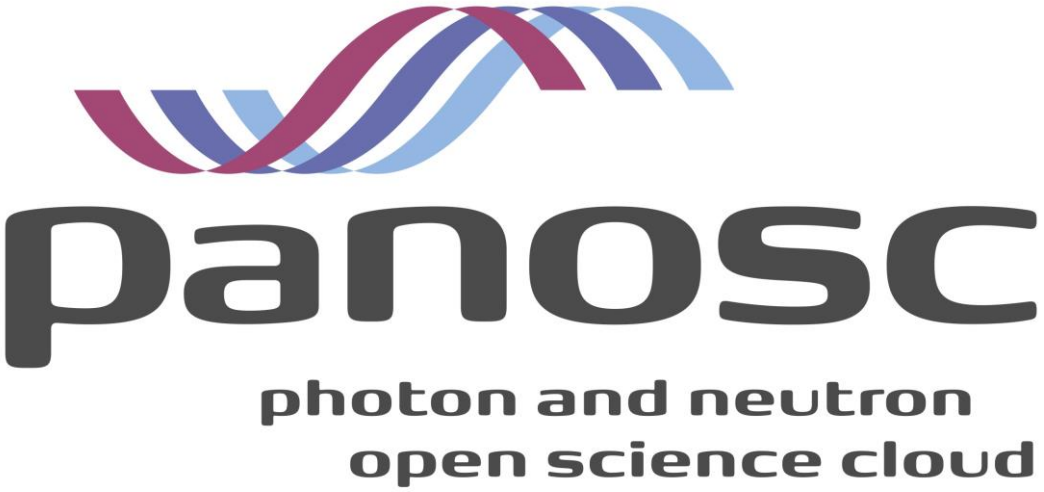
PaNOSC

Photon and Neutron Open Science Cloud

H2020-INFRAEOSC-04-2018

Grant Agreement Number: 823852



Deliverable:

D7.2 Photon and Neutron EOSC metrics and costs model

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

Project Deliverable Information Sheet

Project Reference No.	823852
Project acronym:	PaNOSC
Project full name:	Photon and Neutron Open Science Cloud
H2020 Call:	INFRAEOSC-04-2018
Project Coordinator	Andy Götz (andy.gotz@esrf.fr)
Coordinating Organization:	ESRF
Project Website:	www.panosc.eu
Deliverable No:	D 7.2
Deliverable Type:	R
Dissemination Level	PU
Contractual Delivery Date:	30/11/2021
Actual Delivery Date:	24/01/2022
EC project Officer:	Flavius Pana

Document Control Sheet

Document	Title: D7.2 Photon and Neutron EOSC metrics and costs model
	Version: 1
	Available at: https://github.com/panosc-eu/panosc
	Files: 1
Authorship	Written by: Ornela De Giacomo, Teodor Ivănoaica, Angela Zennaro
	Contributors: F. Dall'Antonia, R. Dimper, F.Gliksohn, G. La Roca, E. Le Gall, J.F. Perrin, R.Pugliese, T. Holm Rod, D. Roccella
	Reviewed by: All authors and contributors
	Approved: J. Bodera Sempere

List of participants

Participant No.	Participant organisation name	Country
1	European Synchrotron Radiation Facility (ESRF)	France
2	Institut Laue-Langevin (ILL)	France
3	European XFEL (XFEL.EU)	Germany
4	The European Spallation Source (ESS)	Sweden
5	Extreme Light Infrastructure ERIC (ELI-ERIC)	Czech Republic
6	Central European Research Infrastructure Consortium (CERIC-ERIC)	Italy
7	EGI Foundation (EGI.eu)	The Netherlands

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

Table of Content

Executive summary

This document reports the result and analysis of the collection of costs reported by partners for the data services provided to the community, including the costs involved in data management, provision of FAIR data and participation to the EOSC.

The collection was performed by five PaNOSC partners that run very different infrastructures in terms of their nature (synchrotrons, neutron sources, free electron lasers, lasers), size, FAIRness level achieved at the moment of the cost collection and services provided, lifecycle and accounting practices. However, although the identification of cost drivers was a challenging exercise, the information collected could help other RIs to estimate the costs involved in the provision of these services.

The costs associated with the EOSC are based on the available budget and not on the actual costs ruled by demand. The future demand of services from a wider community is still to be assessed, since the services are still being setup and made available openly. However, it is clear that with the current budget RIs will provide services, including long term storage and curation, on a best effort basis and that if a wider community will need to benefit, funding will need to be increased proportionally.

1 Introduction

1.1 Description of the deliverable

This is the second deliverable of WP7. The document presents an evaluation of the costs and added value of the services provided to the community, including the costs involved in data management, provision of FAIR data and participation to the EOSC. The cost collection was based on a structured template, which the following PaNOSC partners completed:

- Central European Research Infrastructure Consortium (CERIC-ERIC)
- Extreme Light Infrastructure ERIC (ELI-ERIC)
- European Synchrotron Radiation Facility (ESRF)
- The European Spallation Source (ESS)
- Institut Laue-Langevin (ILL)
- European XFEL (XFEL.EU)

The Partner STICHTING EGI - EGI Foundation also contributed to the cost collection providing standard costs (or prices) to some of the services they are testing with the PaNOSC partnership, that could be crucial for allowing RIs to provide services e.g. to EOSC users.

The costs were collected based on a methodology discussed and agreed upon by all the partners, described in point 1.3. After the definition of the scope of the exercise, a template was prepared to make the cost collection structured and uniform amongst the partners and to guarantee as much as possible the correspondence between the cost categories considered by every partner. However, due to differences inherent to the accounting practices of every facility, these costs are to be considered as approximations. To our best understanding, some cost lines may have been, to some extent, underestimated or overlooked.

The deliverable reports the result of the cost collection and analysis on the main cost drivers. The charts show the cost range, for each cost item, for the five facilities participating in the cost collection plus EGI, that provided the cost for some services deployed by the RIs. The charts are complemented with an explanation of the outliers and the main factors influencing the costs. The data are considered confidential and therefore partners are referred to in an anonymous manner as P1 to P6.

1.2 Scope of the cost collection and analysis

After thorough consideration, we decided to collect the full cost involved in setting up and maintaining the infrastructure required for providing FAIR data. The research infrastructures (RIs) represented in PaNOSC have different infrastructure and readiness levels regarding data management and FAIRness. Some PaNOSC partners have been storing datasets for some years now and are providing access to data for their users as a regular practice. Their setup costs correspond sometimes to actions which date back to the beginning of the Operation Phase, several decades ago. Other PaNOSC partners have started operating only a few years ago, while others

have not started providing access to users yet. All this variability led to completely different scenarios where some of the partners refer to costs and investments already incurred, while others have provided estimated or planned costs for attaining the same maturity level in the future.

PaNOSC project partners have committed to providing FAIR data by the end of the project, so this is what we will consider as the arrival point for the calculation of the costs involved. However, according to the status of the facility at the beginning of the project, some investments will refer to the cost of providing FAIR data by design, while others will instead represent the costs incurred to transform a system that was designed to produce datasets for the exclusive use of the research group generating them, who had all the necessary information to reuse and analyse them. With this legacy, our RIs are adapting to acquire metadata, standardise output formats, define workflows, etc. In addition to the differences in the lifecycle phase and degree of maturity in terms of Data management, FAIRness and services offered to the users of RIs and the research community in general, RIs have adopted an IT strategy based on their specific needs. Finally, the dimension of the infrastructures of the PaNOSC partner RIs is very different, as well as their nature, despite the fact that they are neutron and photon facilities. With these aspects in mind, we performed the cost collection and analysis with the impression that the identification of cost drivers would be very challenging, as we could confirm later.

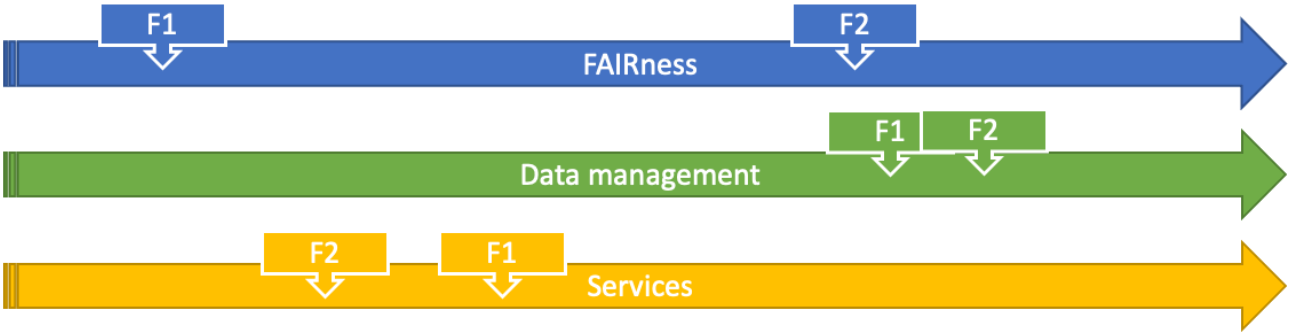


Figure 1. Schematic representation of two research infrastructures, with different readiness levels in data management, FAIRness and other services provided to their users and research community. F1 and F2 are example RIs, the figure does not represent the assessment of two specific PaNOSC partners.

All of the services included in the cost collection and provided by the facilities are considered mandatory for providing FAIR data and linking to the EOSC. For some other partners (ESS, ELI, CERIC), these costs are the costs for setting up the infrastructure that in some cases will refer to future actions, so they will be an estimation based on quotations and a priori assignment of resources.

PaNOSC partners believe that there is a minimum level of services that an ESFRI Infrastructure or an ERIC should offer to their own users (researchers performing experiments at RIs, usually in collaboration with the beamline scientists). The costs of these services are labelled as "Facility operation". The services considered to be additional, necessary for the integration in the EOSC, are labelled as "EOSC". For example, the network, computation and storage costs of datasets generated by users and

elaborated by them within the embargo period are considered as a facility operation cost, while the costs to access, compute or download the same data by other researchers that are not included in the users' group generating the datasets, are considered "costs of the EOSC" by most facilities, with the exception of one that consider these costs as mandatory for providing FAIR and open data and were mandated by their governing bodies to do so.

2 Methodology

The collection of costs of this magnitude and complexity requires an intense dialogue and discussion to ensure that the results obtained are meaningful. For this reason, we adopted an iterative approach by steps, where at the end of every step we analysed the results and revisited some parts of the process. In the following lines, we describe how we approach each iteration:

Step 1:

a. Definition of the goal, objectives and scope of the cost collection;
b. Definition of the cost lines or categories to include;
c. Definition of the approach for the representation of costs;
d. Drafting of a template for the cost collection.

Step 2:

a. First rough cost collection to assess the suitability of the draft template;
b. Re-design of the draft template for the cost collection according to the feedback received by partners;
c. Revision of the cost lines or categories to include, ensuring granularity to allow the analysis but at the same time considering the feasibility according to the accounting practices of each partner;
d. Revision of the scope.

Step 3:

a. Second cost collection, complete;
b. First data analysis on cost drivers, trends and outliers;
c. Collective discussion on the results;
d. Individual interviews with partners to review the cost collection on the light of the data analysis;
e. Final revision of the scope.

Step 4:

a. Third and final complete cost collection;

b. Cost analysis and drafting of the deliverable;

c. Revision of the deliverable by all partners

d. Final version of the deliverable for submission.

In the early stage of the cost collection, partners reported their cost collection but it became immediately evident that costs could have been reported differently according to their RIs' rules and level of granularity for analytic accounting. However, it was necessary for this exercise to be meaningful, to achieve a common ground for the reporting of costs. We strived for partners to take into account similar services that would have allowed for a comparison or at least an analysis. After some iterations and continuous refinements, we identified a subset of services, mandatory for operating a pan-European research infrastructure, that we called "Costs for data processing inherent to the facility operation". The rest of the costs that did not fit into this category were classified as additional services for researchers and for the community in general, under the title "Data processing linked to EOSC Service Catalogue - data reuse"

The template for the data collection was discussed and modified through an iterative process as described above, until the structure was considered compatible with all facilities' accounting practices. In other words, the structure and granularity were such that it allowed all partners to report costs in most of the cost lines defined. The template that we adopted and used for the cost collection is enclosed in Annex 1.

The first table "Costs for data processing inherent to the facility operation" gathers the costs for each facility from the moment the data is generated until the end of the embargo period. It includes all costs related to the download or analysis of the datasets by the research group that created them.

The second table "Data processing linked to EOSC Service Catalogue - data reuse" represents an estimation of the costs for wider use and re-use of the data, including access to datasets no longer under embargo, for download or online computing, by an EOSC generic user.

It is worth mentioning that at the moment this task was performed, there was little experience in the provision of services to a wider community, beyond each facility's user community. Moreover, most of the services enabling facilities to provide open access to their data were under development. We would like to emphasize that these costs are the best approximation facilities could provide. As such, they should be further refined after some years of experience and practice when the datasets that are currently under embargo will become public and the tools developed will allow a larger community to reuse them.

Facilities used different approaches to estimate these costs but in general the starting point was the budget, either available or forecasted. Each partner tried to allocate a reasonable part of the budget to cover costs based on what we anticipate may be the demand for data download, processing, etc. We expect that these costs, or at least the budget allocation between the different cost lines may change in the near future,

when the feedback from experience will allow us to predict the actual demand by the community and after fine tuning with some EOSC aspects that still need to be defined.

Data Management is in most cases a collection of facility-specific tools and services, designed and implemented based on a facility-specific scenario to serve the users of each RI. The following exercise will focus on presenting the cost ranges, foreseen by the PaNOSC partners, for providing FAIR Data and FAIR experiments to their users and other researchers.

Since the costs are impacted by the type of research, by specific modes of operation of the accelerators and by specific types of instruments, it is important to understand that those particularities play an important role in the design of the Data Acquisition, Control Systems and all the specific IT Infrastructure supporting the Data Management Systems.

In this context, for solid documentation of the costs and costs drivers, each partner has provided a high-level description of their Data Infrastructure, emphasizing on the particularities of their specific Computing Model that has a direct or indirect impact to the IT and data management processes.

2.1 Presentation of the PaNOSC partners

2.1.1 CERIC-ERIC

CERIC-ERIC (CERIC) is a European research infrastructure consortium where 8 countries put together their most relevant infrastructures to serve the international research community. The access to these RIs is coordinated centrally by the ERIC, and as a consequence also the data policy. CERIC's data policy was approved in 2020 and is based on the PaNOSC framework data policy, with the intent to be as aligned as possible with the other PaNOSC partners. This policy foresees an embargo period of three years, after which data is made publicly available, unless the research group who created the datasets decides to make them open earlier. For the cost collection, only one RI was taken into account, the Elettra Synchrotron radiation facility in Trieste, Italy. From all the beamlines, Elettra confers to CERIC around 5-10% of the beamtime of half of its beamlines. In particular, 5 beamlines offer a total of about 1650 hours/year and 9 branch lines offer a total of 1685 hours/year. The sum of this beamtime is closely equivalent to one beamline. Two beamlines operated by external institutions are conferring 100% to CERIC, providing almost 10.000 hours per year and two of the Elettra beamlines offer also time to measure with conventional sources, when there is no synchrotron radiation, for about 1200 hours per year. Overall, this corresponds on average to 90 external projects per year plus several in-house experiments, generating 6.5 TB/y of data.

The particularity of CERIC-ERIC is the intrinsic distributed nature of the instruments. A final decision on where data will be stored and analysed for all facilities has yet to be taken depending on the technical tests that are being performed during PaNOSC, the cost and

the specific regulations and requirements of the member countries. Currently, a realistic scenario will see a hybrid infrastructure where part of the storage and calculation will be provided by Elettra Synchrotron, another part will reside in the cloud and finally in the partner facilities that produce huge data (for example the Solaris synchrotron which is providing instruments like a Cryo-EM). This is due to the consideration related to the data gravity. In function of the size of the datasets, their transfer may not be practical and adequate computing resources have to be put in place close to where the data is stored.

The Elettra Synchrotron storage infrastructure is based on a CHEP distributed storage cluster of 4PB capacity. The network infrastructure is currently based on a 10Gbps Internet link currently being upgraded to 100Gbps. Internal connections from the critical components of the infrastructure are 10 to 100GBps. The storage system is complemented by a tape library of 6PB that can grow up to 60PB and acts as a backup and long-term repository of scientific data. HPC clusters and hyper convergent clusters are used to run the services required to support FAIR data management.

2.1.2 EGI Foundation

The EGI Foundation (also known as Stichting EGI and abbreviated as EGI.eu) is a not-for-profit foundation established under the Dutch law to coordinate the EGI Federation (abbreviated as EGI), an international collaboration that federates the digital capabilities, resources and expertise of national and international research communities in Europe and worldwide. The main goal is to empower researchers from all disciplines to collaborate and to carry out data- and compute-intensive science and innovation.

The EGI Federation is one of the largest distributed computing infrastructures for data-intensive research collaborations. It federates hundreds of major research data centres in Europe and worldwide, making advanced computing services, capacity and research data accessible in a federated manner to members of international scientific collaborations. The EGI Federation provides access to more than 1.2 Exabyte of research data and 1.2 Million CPU Cores for data processing and analysis needs to thousands of researchers. EGI expanded the federation of its facilities with other non-European digital infrastructures in North America, South America, Africa-Arabia and the Asia-Pacific region, as such EGI fully realises the “Open to the World” vision. In order to interoperate at international level, EGI and its partners operate in the context of a lightweight collaboration framework defining rules of participations via a corpus of policies and technical guidelines.

EGI offering includes a federated IaaS cloud to run compute- or data-intensive tasks and host online services in virtual machines or docker containers on IT resources accessible via a uniform interface; high-throughput data analysis to run compute-intensive tasks for producing

and analysing large datasets and store/retrieve research data efficiently across multiple service providers; federated operations to manage service access and operations from heterogeneous distributed infrastructures and integrate resources from multiple independent providers with technologies, processes and expertise offered by EGI; consultancy for user-driven innovation to assess research computing needs and provide tailored solutions for advanced computing.

The EGI Cloud Federation aggregates resources by defining a set of standard open-source interfaces and protocols to access the different cloud functions - such as resource discovery, user authentication, compute and data access services - in a uniform way at all the sites, enabling workloads to span and seamlessly migrate across resource centers. The EGI technical platforms are co-developed with research communities and technology providers. In order to do so, EGI has established processes and technical infrastructures for requirements gathering, software validation, verification and distribution through the Unified Middleware Distribution. Through its solutions for High Throughput Computing, Cloud, Federated Operations and Community-driven innovation and support, EGI is contributing to the Open Science Commons vision (http://go.egi.eu/osc) according to which Researchers from all disciplines have easy, integrated and open access to the advanced digital services, scientific instruments, data, knowledge and expertise they need to collaborate to achieve excellence in science, research and innovation.

2.1.3 ELI ERIC

ELI ERIC was established this year on the 30th of April 2021 when the Extreme Light Infrastructure (ELI) was granted the legal status of European Research Infrastructure Consortium (ERIC) by the European Commission. ELI ERIC will provide access to world-class high power and ultra-fast lasers for science and enable cutting-edge research in physical, chemical, materials, and medical sciences, as well as breakthrough technological innovations. At this moment the ERIC has the Czech Republic and Hungary which are joined by Italy and Lithuania as founding members, while Germany and Bulgaria are joining as founding observers. A third ELI facility is under construction in Romania in the field of nuclear photonics and is expected to complement the current ELI ERIC facilities in the future.

ELI facilities are offering new research possibilities in particle physics, nuclear physics, high energy beam science, nonlinear field theory, and ultrahigh-pressure physics. Besides its fundamental physics mission, a paramount objective of ELI is to provide ultra-short energetic particles (10 to 100 GeV) and radiation (up to a few MeV) beams produced with compact laser plasma accelerators.

At this stage, the two ELI facilities, currently engaged in the last stages of the beams and experiments commissioning process, are now starting to converge, consolidating a common data layer for which ELI ERIC, as a custodian and steward of the Data, is developing and

implementing a data commons across the entire organization.

A geographically distributed Research Infrastructure Consortium, putting together two state-of-the-art Research Infrastructures and preparing them for normal operations, adds multiple challenges due to the particularities of their scientific research instruments and the specifics of the lasers science. With respect to this, one of the major challenges of ELI is the development and implementation of a unified Scientific Data Management System which supports two different geographically distributed laser facilities. Such systems are adding technical and inherent financial challenges that are starting to be addressed by our technical teams, using the experience of the PaN Scientific community.

From the technical point of view, ELI is now focusing on building a fully integrated and functional FAIR Scientific Management System that will be supporting the two facilities following the ELI Data and Access Policies. This challenge also has a financial impact, because ELI will still have a part of the necessary resources geographically distributed and available at the required uptime and performance levels.

Among the technical challenges faced by Large Research Infrastructures, is the transition from the commissioning to the operation phase. This requires the fine-tuning and preparing the ICT infrastructure for supporting the everyday facility operation and the scientists during the entire research lifecycle, from the experiment proposal to the publication of the results.

The associated costs for building the ICT infrastructure for ELI, has two components, one is defined by the costs of maintaining existing infrastructure, referring to already existing IT systems supporting the commissioning process, based on which we have sized the second component, which is the required infrastructure for supporting the future ELI Users' activities and Data operations. The newly sized infrastructure, built to support more than 50 possible beam configurations, as lasers could have different beam paths supporting one or multiple experiments, each of them generating specific data produced by different experimental setups. Additionally to this, the integration of the data generated by each experimental area, by each experiment end-stations, or by different detectors or custom scientific instruments poses new challenges as the facilities are estimated to produce around 2 PBytes of raw data each year which, though it is FAIR by design, it still adds the costs associated to a distributed Scientific Data Management System.

The costs associated with IT Operations and Data Management, in ELI's case, since the facilities are starting the operations using a FAIR by design concept, are mainly driven by the fact that the ELI will develop a central data services aggregation layer harmonizing the operations of the ICT environment by integrating and federating the specific existing ELI site-specific resources. The ELI ICT teams are currently focusing on generating, aggregating, correlating, and curating the data that will be further processed and offered to the users. At this moment the two ELI member facilities, having distributed scientific instruments and accelerators and which are maintained by different specialized teams serving particular scientific challenges,

are now making the first steps to adopt a data commons applying common standards and goals.

2.1.4 ESRF

The European Synchrotron Research Facility (ESRF) was founded in 1988 by eleven European countries to build the world's most performing and bright "third-generation" light source. User operation started in 1994 and by 1998 40 beamlines were in operation. Since 2009 an ambitious upgrade programme has been taking place, with new beamlines and a new source called EBS (Extremely Brilliant Source) making the ESRF the first "fourth generation" hard X-ray light source. With the ESRF upgrade programme, the ESRF is on the ESFRI roadmap. As of 2021, the ESRF counts with 22 partner nations (13 members and 9 scientific associates); has 4 Nobel prize-winners among its users; produces more than 2000 publications per year and receives the visit of thousands of scientists from around the world every year for conducting experiments.

The central ESRF IT infrastructure is located in two, physically separated, data centres. The ESRF data communication network is based on top of the range Extreme Network and CISCO switches. It consists of more than 600 network switches and ~9 000 active network ports out of which 500 ports of 100Gbps. In our two data centres the high-performance storage (10PB) used to process the incoming data from the beamlines is based on the latest release of GPFS. Local communication between GPFS nodes and the compute clusters uses InfiniBand and Ethernet. Data is copied to tape storage for archival and the metadata is stored in a catalogue as soon as the data is produced. The tape libraries are currently StorageTEK SL8500 and the tapes are following the LTO standard. The tape libraries allow using different generations of LTO tape drives and tapes and we are using LTO5, LTO7, and LTO8, with the LTO5 tapes and tape drives being currently phased out meaning that data is migrated from the older generation tapes to the new one. The firmware of the tape libraries is regularly updated in the frame of the maintenance contract. At the time of writing the underlying software for writing data to tapes is ATEMPO from the company ASG. However, we are currently undertaking efforts to use the open-source solution BACULA for archiving our data.

The technical infrastructure used for the data archive is the same as for the operation of the ESRF (network, disk storage, backup). The uptime of this infrastructure is ensured by staff being on standby 24h/24h and 7d/7d to ensure highest availability. During the past two decades, the reliability and hence the uptime of the IT infrastructure was largely better than 99%.

2.1.5 ESS

ESS is an ESFRI landmark designed to be the world's most intense neutron source. It started construction in 2014 and in 2015 it was converted from a Swedish AB company to European Research

Infrastructure Consortium (ERIC) with 13 member countries. While ESS is head-quartered and being constructed in Lund, Sweden, it's Data Management and Software Centre is located in Copenhagen, Denmark, and hence, ESS has two host nations.

ESS has 15 neutron beam instruments planned that can be used to investigate the structure and dynamics of materials from atomic- to macro-scale, and is expected to perform experiments within the materials- and life-sciences. Examples of materials to be investigated are superconductors, batteries, and proteins.

ESS has been designed with FAIR data from the very beginning, although the term FAIR was not established at that point in time. The data acquisition system collects neutron event data, meta-data (e.g. sample environment information, instrument setup, proposal number, information from target and accelerator, etc) and stores them in the NeXUS HDF5 format. It also allows for log books and other auxiliary data to be stored and associated with the experiment in order to make the data reusable and interoperable. The data are findable and accessible through the SciCat metadata catalogue.

The data will be open for people outside the associated proposal team after an embargo period of three years subsequent to completion of the experiment. Moreover, services will be provided so that users can process and analyse their data without transferring the data to somewhere else.

Because ESS is designed as a FAIR data research infrastructure from the outset the cost of facility operations only is not easily discernible from that of EOSC, particularly because the services for accessing, downloading, processing, and analysing data are the same for members and non-members of a proposal team associated with the data. It is only during the embargo period that the permissions (authorization) differ for members and non-members of the proposal team.

2.1.6 European XFEL

The European XFEL is a free-electron laser facility for the generation and utilization of ultra-short X-ray laser pulses of unprecedented brightness, enabling world-class research at the frontiers of science.

The facility consists of a linear accelerator on the DESY site in Hamburg and a research campus with an experimental hall in Schenefeld, the two units being connected by a 3.4 km long tunnel. The facility is run by the non-profit European XFEL GmbH, based on a collaboration of 12 member states, and employing a staff of more than 400 people.

Since the start of operation in September 2017, and during gradual addition of scientific instruments to a total number of six instruments currently, European XFEL has had almost 400 users per year on average. The experiments (user beam-time as well as in-house research and commissioning) have produced more than 50 PB of data (12 PB per year on average), of which about 5 PB are older than three years and hence beyond the embargo period.

At European XFEL the purpose of processing vast amounts of scientific data at quasi real-time – in particular the correction of raw multi-gain megahertz/ megapixel detector data, necessitates the use of high-performance storage solutions.

The facility's storage infrastructure is based on the IBM Elastic Storage System (ESS). Besides the 'standard' IP networking based on Ethernet, the high-performance components of the system are interconnected with an InfiniBand network. Additionally, a dedicated long-haul InfiniBand link (~4km) connects the European XFEL experimental hall in Schenefeld and the DESY data-centre in Hamburg. In order to ensure data integrity and to allow fast data access and efficient data management, many GPFS features are utilized in the system. The storage system is complemented by a dCache instance, which acts as a 'cold' raw data repository, and a tape library for the data archive.

2.1.1 ILL

ILL is an internationally financed scientific facility for research using neutrons generated by a high-flux nuclear reactor. Users can do experiments on more than 50 different instruments in domains like physics, biology, medicine and numerous others.

The ILL was the first international scientific user facility to publish a "Scientific Data Policy" in November 2011, just before the opening of the December 2011 proposal round. The text came into force in October 2012, and prescribed a default non-disclosure period of three years during which access to data is restricted to the experimental team; in cases where no request for data has been made, this period would be extended to five years. With the latest version of this data policy being adopted in 2017, this policy already integrates several aspects of the FAIR principles including the embargo principles and data stewardship. Following the publication of the policy, the ILL created an interdisciplinary working group, the DPP (Data Protection and Processing) group.

One of the main personnel costs was the development of data management tools and services. The first objective was the DOI management defining a workflow for linking datasets to unique identifiers, developing the technical tools and communication and assistance to users. In parallel, ILL invested personnel time and hardware to provide access to the data from outside of the facility via a data portal. Through the data portal, ILL provides access to open data once the embargo period has finished, however users can make the data open at any time. The overall cost to maintain these tools and services is today mainly a hardware support cost.

The next, and current, step is to provide remote access to an analysis environment with direct access to the datasets. Despite the annual influx of international visitors, there has always been interest in options that don't require users to travel to ILL. Remote access to instruments, experiments, and datasets unlocks scientific opportunities for those less able to travel, as well as enabling more flexible working patterns for staff. Supported via PaNOSC, the main

cost of this solution, called VISA, has been the personnel cost for development. The hardware costs depend on the number of users and the necessary compute resources needed to perform their data analyses.

Today, the main cost to deploy VISA at other partner facilities comes from the investment in hardware procurement and associated maintenance. However a non-negligible personnel cost is required to integrate VISA to the existing IT infrastructures and data management strategies.

3 Metrics and costs

The costs considered in the analysis for each category are related to the initial set-up and yearly operation for each Facility. The costs corresponding to each cost category will be represented in the charts below with the following:

- Implementation personnel: costs related to the personnel effort needed for the initial set-up, referred to a time frame of five years;
- Implementation CAPEX: Costs related to the capital (hardware and software) needed for the initial set-up, referred to a time frame of five years, the typical depreciation period for IT hardware;
- Other implementation costs: maintenance (IT, building, cleaning), electricity, heating, cooling, and other set-up or operation costs.
- Operations personnel: yearly costs of the personnel for the operations multiplied by five, to represent a time frame of five years
- Operations CAPEX: yearly costs related the capital (hardware and software) needed for the operations multiplied by five, to represent a time frame of five years

And the totals:

- Implementation total: sum of all costs related to the setup, referred to a standard implementation time frame of five years
- Operations total: total yearly costs (personnel and CAPEX) for operations multiplied by five, to represent a time frame of five years

All the costs in the figures are expressed in thousands of Euros (k€), with exception of the summary charts, where costs are expressed in millions of Euros.

The implementation costs are represented in a time frame of five years because this is the typical depreciation period for IT hardware. Other expenses during the setup period, that cannot be attributed to investments or personnel are grouped in a third category (other implementation costs).

The operation costs reported by partners are referred to the years 2020 or 2021. In full regime (after the implementation), they are expected to change slightly over the years, unless there are upgrades that result in an increase of the operation costs. They are represented in the chart for a period of 5 years just for the sake of comparison with the implementation costs, but we expect an increase over next years if RIs will continue developing and making available new services for their user community and the EOSC.

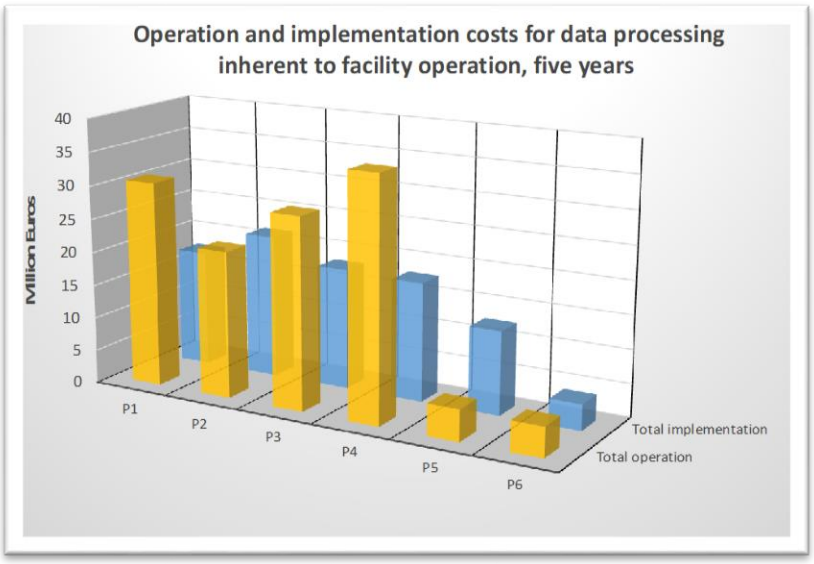


Fig. 1: Operation and implementation (setup) costs for data processing inherent to facility operation. The implementation costs are referred to a time frame of five years. The yearly operation costs were multiplied by five, to represent a time frame operation of five years.

Figure 1 represents the sum of the setup costs for five years, for all cost categories, and the operation costs for one year multiplied for five. Considering the cost lines monitored in this cost collection, the operation costs in full regime are higher than the implementation costs. All the developments of PaNOSC partners during the project and after, are expected to generate an increase in the operation costs to maintain and further develop those outputs. IT is not clear to what extent managers and funders of research infrastructures are aware of this. The difficulties experienced during the cost collection and the uncertainties about the services that EOSC will provide as well as the how the demand will evolve once services will be made available, make it also hard to predict with accuracy the additional funds that will be needed by RIs to meet the demands of the research community.

The total cost for the initial set-up varies from roughly 12.5M€ to 21.5 M€ for most of the partners, with only one exception due to the dimension of the RI considered for the cost collection, but that would scale accordingly if reported to the dimension of other similar PaNOSC partners. This corresponds to investments over a 5-year period, and considering that this is the typical depreciation time for IT infrastructure, facilities should plan to be able to upgrade part of their infrastructure after that.

Fig. 2: Operation and implementation (setup) costs for data processing inherent to facility operation, by personnel, capital expenditures and other. The implementation costs represented correspond to five years, while the operation costs are yearly.

For some partners, the figures may not include the total costs for long

term storage and curation. Therefore, if RIs will need to adopt a long term data preservation strategy, procuring dedicated infrastructure and providing human resources, or contracting commercial services for this purpose (as the ones that are being tested at a pilot scale in some EOSC projects), additional funding will be necessary. We invite RIs and funders to not underestimate the resources and effort involved. PaNOSC RIs reported the yearly operation costs from roughly 4.4M€ to 7.2M€ for four facilities, and two of them have operation costs around 1M€, in one case this is due to the dimension of the RI.

The variations, both in investments and operation, are due to different factors that will be mentioned later in the document. In addition to the nature, dimension and strategy of the RIs, partners have reported for example different personnel costs between facilities due to their geographical location. Other differences in costs but also in the allocation to different categories arise from the existence of agreements in place with third parties as providers of some services, or the choice to develop in-house solutions for specific needs against purchasing commercial solutions.

An additional source of variability comes from the internal structure of the facilities. Distributed facilities, as opposed to single-sited, cannot always centralise all IT operations and need a certain redundancy of infrastructure and personnel, which naturally increases the costs.

The charts below give an indication of the contribution of every category included in the collection to the total cost of implementation and operation. The bars represent the percentage in which a given category contributes to the total cost. This percentage was obtained as the sum of the costs for each category for all the partners. Although the relative importance of a category can vary from one partner to the other according to their IT strategy or their accounting methodology, the charts can be used as a general guide.

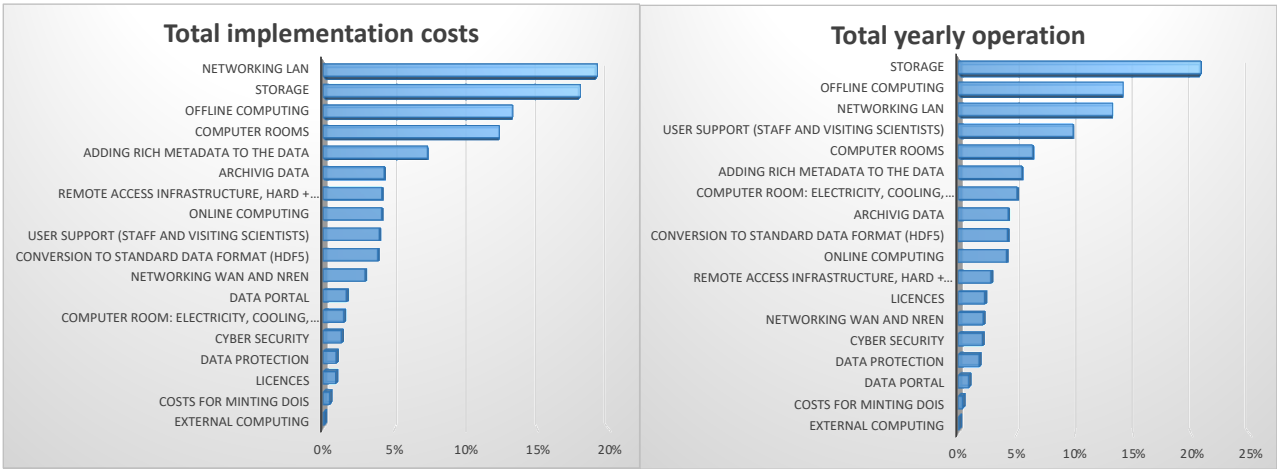


Fig. 3: Indicative percentage of contribution of every category to the total cost of implementation (setup) in five years, and yearly operation.

The costs linked to EOSC were represented in the same way as costs for data processing inherent to the facility operation. However, as will be

discussed in more detail later, facilities currently are developing the services, so the estimation of the demand for these services is challenging.

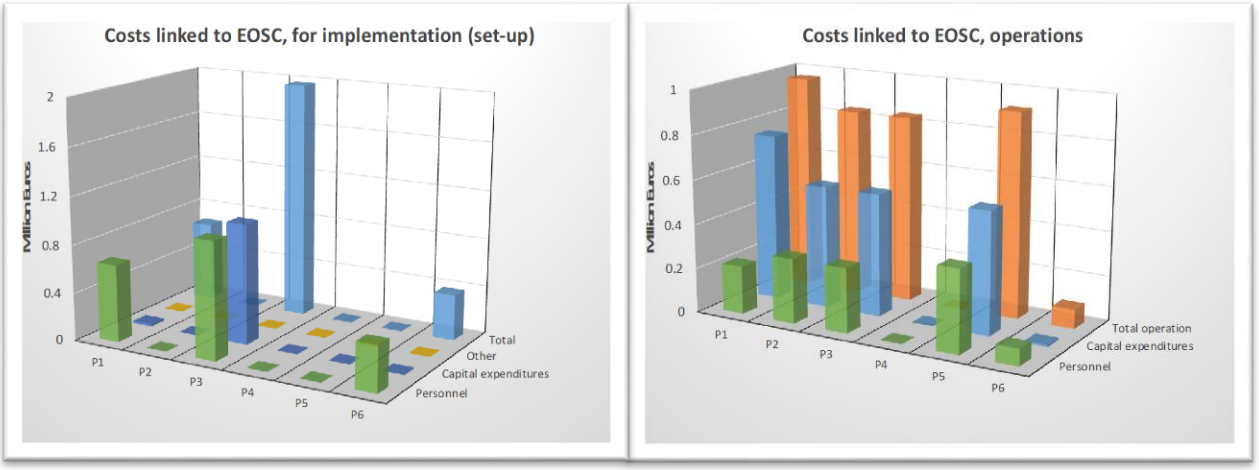


Fig. 4: Operation and implementation (setup) costs linked to EOSC, by personnel, capital expenditures and other. The implementation costs represented correspond to five years, while the operation costs are yearly.

We see that only one of the partners reported capital expenditures, and there are two main reasons. First, some facilities consider open data as part of their mandate, so they don't attribute these costs to linking with EOSC. The second reason

In the following pages we present the detailed distribution of the costs collected to allow an analysis of the main contributions to the implementation and operation costs. Some comments will accompany the charts, since the analysis is not straightforward.

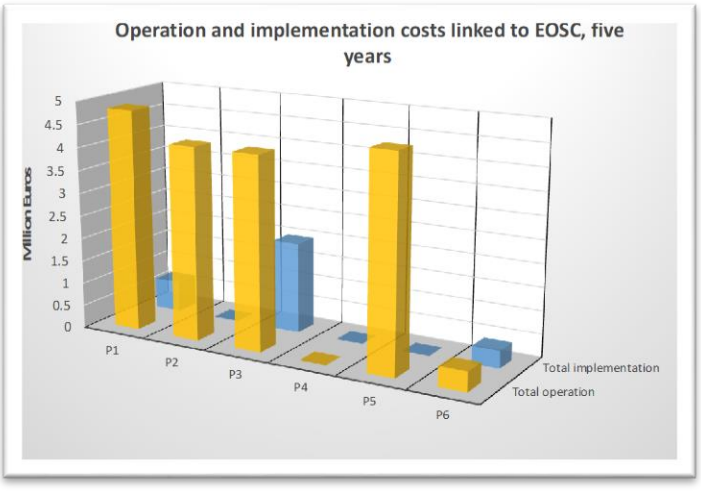


Figure 5 shows that also for the costs associated to the EOSC, it is expected to have high operation costs. It should be however considered that some RIs did not include any implementation cost linked to EOSC not because these do not exist, but because they have committed to provide open data prior to the existence of the EOSC and therefore, they consider these as costs of the operation of the facility.

Fig. 5: Costs linked to EOSC, divided in implementation and operation. The implementation (setup) costs correspond to five years, while the yearly operation costs were multiplied by five, to represent a time frame operation of five years.

3.1 Data processing inherent to Facility Operation

3.1.1 Adding rich metadata to the data

This was identified as the effort of defining and implementing the automatic metadata collection for each beamline/end-station/instrument. The term metadata describes information referring to data collected from instruments, including (but not limited to) the context of the experiment, the experimental team, experimental conditions, electronic logbooks generated during the experiment and other logistical information. It is expected that the metadata definitions will constantly evolve and expand. The design, implementation and further development of tools like the electronic log-book are part of this activity and have been documented by each of the partners based on their specific costs of implementation and team structure.

Right from the beginning we could anticipate that new RIs that are just starting to operate (e.g. CERIC and ELI) have bigger implementation costs than RIs that have been in operation for decades, since there is no pre-existing infrastructure. At the same time, the newly established partners might report lower recurrent and maintenance costs as they have the advantage of starting with some modules and services supporting FAIR and Open Data principles from the very beginning.

In this context, even if all PaNOSC partners are committed to adopting common FAIR policies and standards, the cost of adding rich metadata is variable and, based on collected details, depends on the "FAIR technical readiness" of each partner, on the existing infrastructure and skills, and on the maturity of the control system type and data acquisition strategy of each RI.

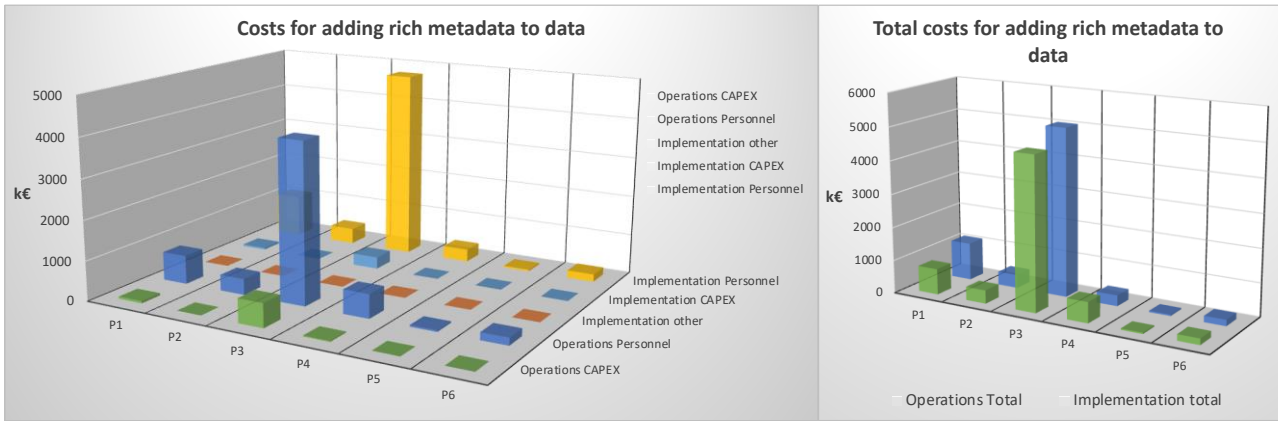


Fig.6. Initial implementation and operation cost for adding rich metadata to data, over a period of 5 years.

The cost of adding rich metadata is reported to be significantly higher for the partner P3 than for the other partners. For P3, the data acquisition system has from the outset collected metadata simultaneously with detector data and stored everything in a common NeXUS (HDF5) file adhering to FAIR principles. Hence metadata are not added to the data as such but treated on equal footing with the detector data. Therefore the cost of adding rich metadata has been set to that of the full data

acquisition system and consequently there is no additional cost for converting data to HDF, minting DOIs, and creating a data portal in the form of a metadata catalogue.

Similar operation costs, as represented in the above chart, are identified only at two of the partners. In the case of the others, this cost is allocated in other categories of operation costs. The allocation to one or another category is highly dependent on the structure of each RI's accounting principles, internal policies and regulations.

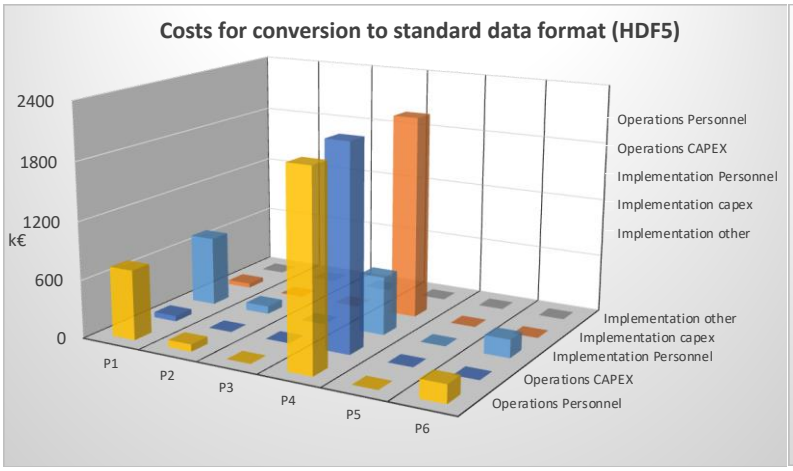
Regarding the yearly operating costs, we can see the impact of FAIR and Open Science on the costs reported by our partners and the great variability in the costs associated with adding rich metadata to data. These costs are impacted mainly by the maturity of the control and data acquisition systems and by the expected effort to update the systems to support FAIR. In some cases, these costs might change after the implementation phase. However, since these are changes that in some cases might be implemented over several years, these costs are relevant for Research Infrastructures that have to perform major upgrades to their Data Acquisition systems. Consequently, for some research infrastructures the operation costs are expected to be similar or even higher than those in the implementation phase.

As the facilities in this costing exercise have reported costs for new hardware or hardware under vendor's support, that adds small costs for operations that might be associated with some minor upgrades (adding extra RAM /new network cards etc.).

The yearly costs are covering the maintenance of the systems generating rich metadata, both for hardware and software.

3.1.2 Conversion to standard data format (HDF5)

This category was defined as the cost associated with the effort of implementing the HDF5 file format for each data source (detector/end station) and adapting the data processing/analysis programs to the HDF5 file format if required. Part of this activity is to convince the beamline scientists to adopt the HDF5 file format which in many cases is very important to ensure that the number of individual files generated is kept low.



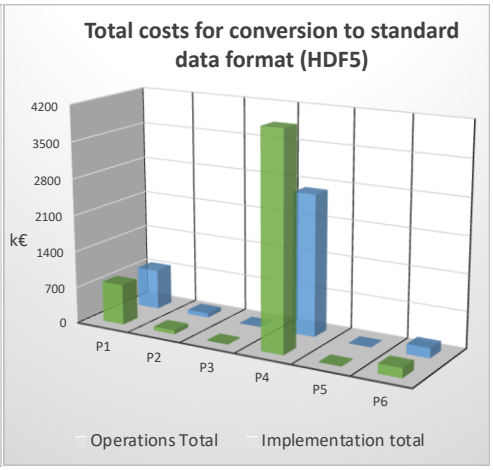


Fig.7. Initial implementation and operation costs for five years, for the conversion to standard data format (HDF5)

For our P4 partner, this cost includes the data acquisition (DAQ) system upgrade since HDF5 is provided directly by DAQ. In particular, it includes multiple detectors, each capable of generating up to 10-15Gbytes/sec, other detectors, slow data - software + hardware streams. In this case, all the technical details, together with the fact that having a distributed data architecture that requires having a big distributed data management team with both scientific and technical team members, is directly reflected in the cost. Two of the partners did not report costs, either because they produce HDF by design or because they aggregate this cost to another, according to their accounting practices.

3.1.3 Minting DOIs

Minting DOIs is defined as the effort to implement the initial DOI minting system and then to maintain it. This is independent of the granularity of the DOI attribution which has to be decided by each RI.

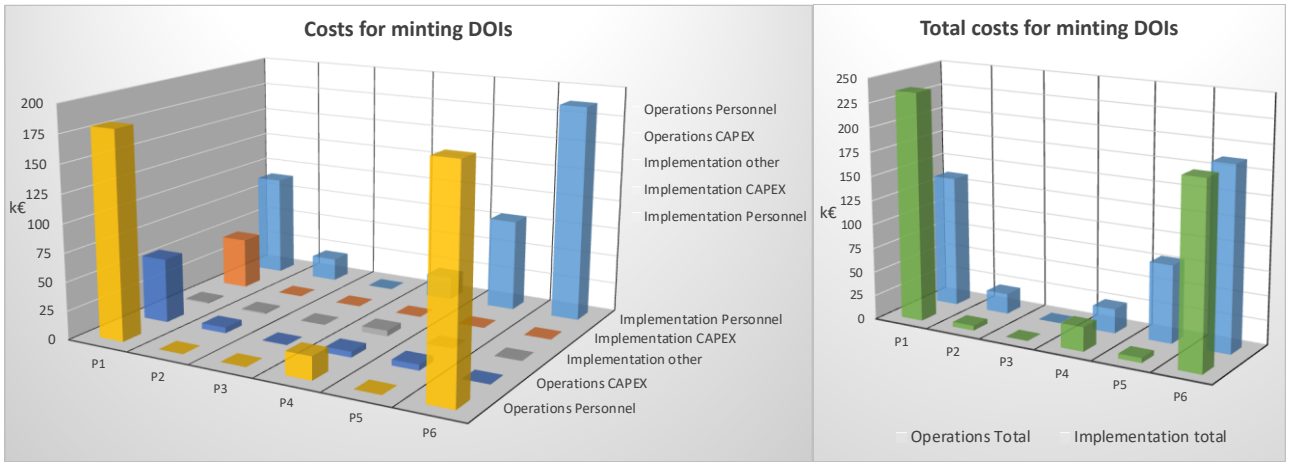


Fig. 8. Initial implementation and operation costs for five years, for minting DOIs,

The cost of minting DOIs is also impacted, in some cases, by the fact that some partners, for example ERICs, have distributed facilities that may have different systems already built for each facility, that require further integration into a single central system.

3.1.4 Cost of standardisation of the data

Based on the input coming from four of the partners, that provided costs in every cost line and considering that the other two partners are either aggregating 3 cost drivers together (adding rich metadata, minting DOIs and/or switching to a standard HDF5 format) or not reporting a conversion cost since they produce standard data by design, we are including below a representation of the reported costs.

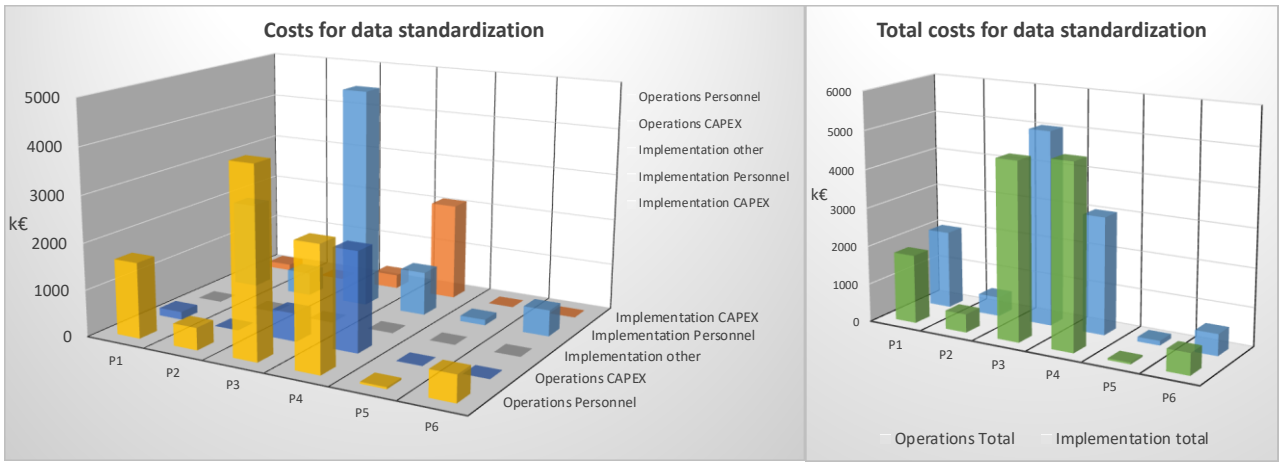


Fig.9. Data standardization costs corresponding to the initial implementation and operation in full regime for five years

3.1.5 Data Portal

In this cost line we report the total estimated costs for building a data portal and adapting it according to the specific requirements and computing model of each RI.

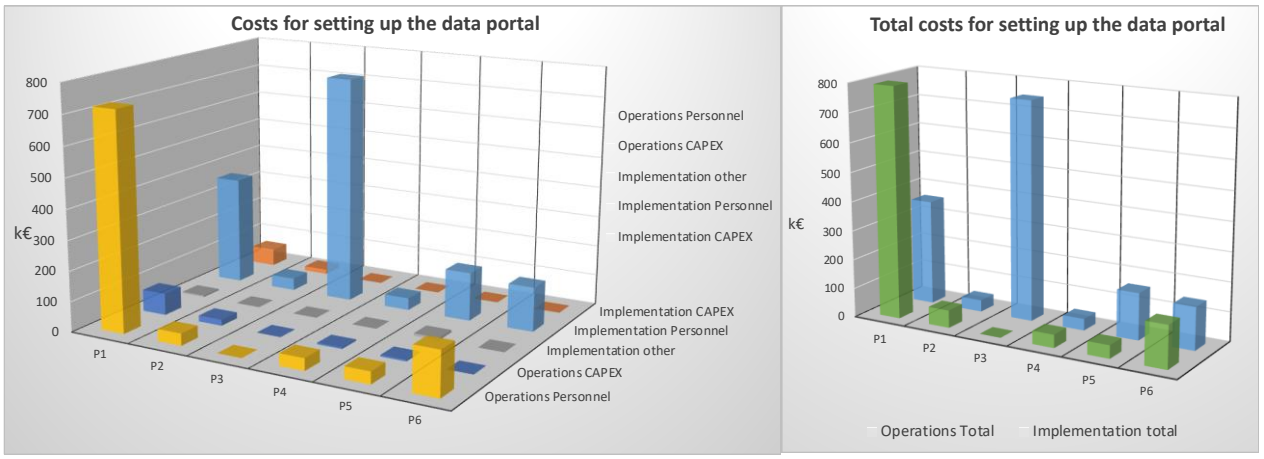


Fig.10. Initial implementation and operation costs for five years, the data portal

Because of the specific requirements of each RI and the intrinsic differences in the cost of personnel, the costs reported are different, but these do not scale with the volume of data, number of datasets, number of proposals or any other of the parameters tested.

3.1.6 Secured Data Access and cybersecurity

In this cost driver we report the implementation and maintenance of cyber security measures to keep the IT safe from intrusion. This topic will vary at each RI in the function of their cyber security culture and, in some cases, they may refer to the provision of a service by a third party as they might have services provided by an NREN/partner.

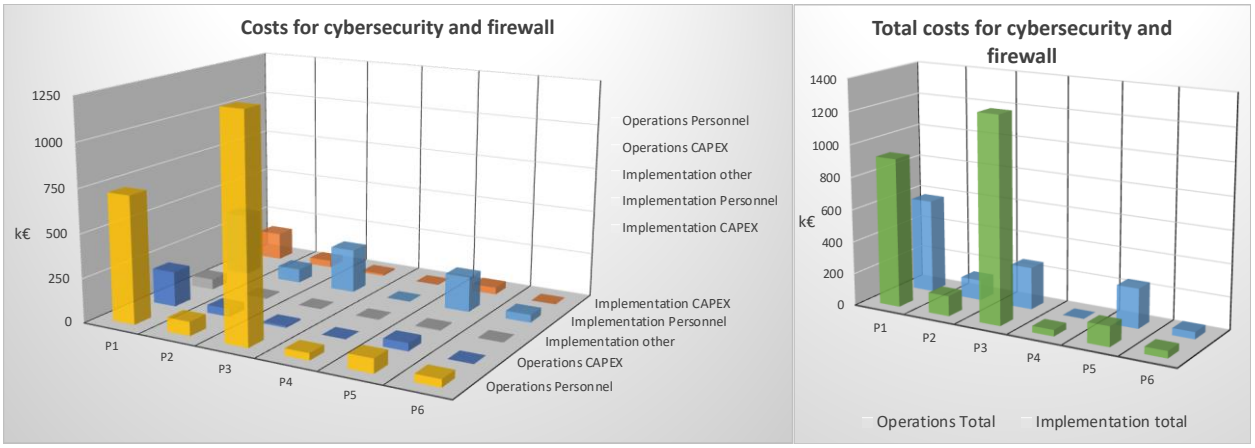


Fig.11. Initial implementation and operation costs for five years for Secured Data Access and cybersecurity

3.1.7 Data Protection - Account management and ACLs

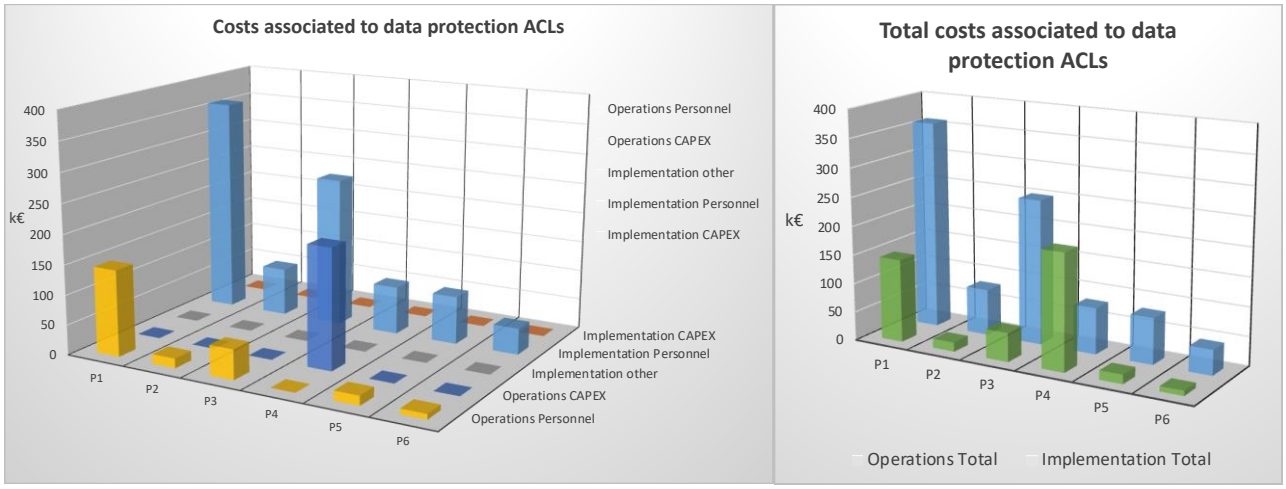


Fig.12. Initial implementation and operation costs for five years for data protection ACLs

Fine-grained access control lists (ACLs) are usually developed and implemented to allow authorized and accounted access to data only during the experiment and until the end of the embargo period. This implies the existence of individual user accounts for visiting scientists at the beamlines, which in turn may complicate access to the beamline control system. The latter point has not yet been implemented at some of the partners and is going to be designed, tested and implemented by each partner Research Infrastructure.

3.1.8 Storage costs

For some of the partners, the cost for data storage is based on high-performance cluster storage systems, systems supporting parallel file systems on ultra-scalable architectures supporting GPFS or CHEPFS (e.g. IBM SpectrumScale). Since the central storage systems are still critical for operation of the beamline detectors, they must have sufficient

bandwidth for the detectors and simultaneously for the data analysis clusters. This adds a high challenge for the IT Infrastructure design and implementation team but also for the computing operations teams.

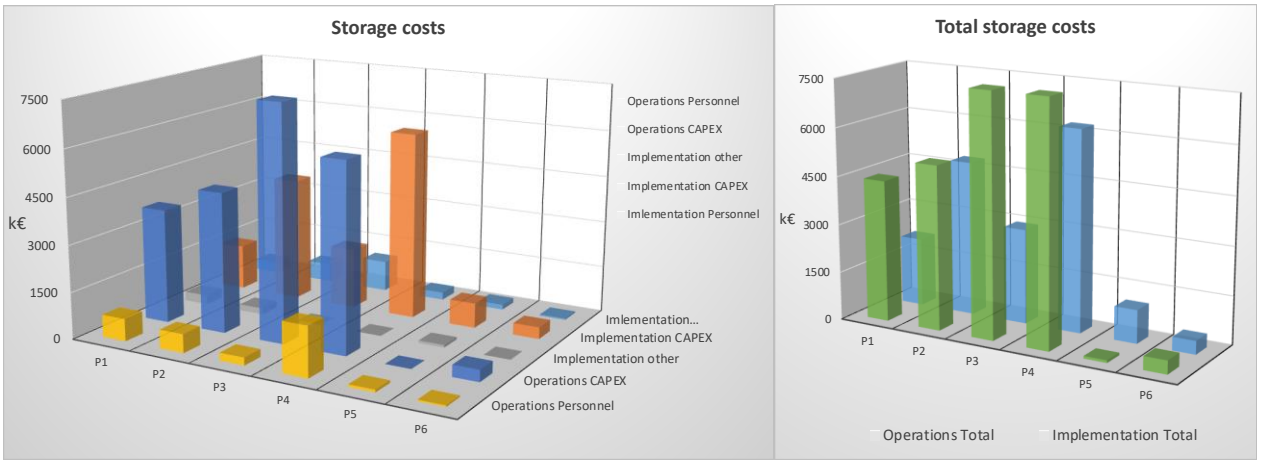


Fig.13. Initial implementation and operation costs for five years for storage.

The costs presented for Partner P2 and P4, are costs associated with new systems, new designs for which the support and all the operations procedures and processes have not been sized yet. It is expected that in the future the storage becomes similar to the mature organization as most of the partners are using the same types of systems and strategies for data storage (the cost per PetaByte is similar, the only difference comes from the complexity of the integration with the beamline control and data acquisition systems, that have a significant impact on the cost of operating such storage systems).

3.1.9 Data Archival

Here we report the cost of tape or long term storage. The initial investment includes the tape libraries with the first set of tape drives and tapes, but also the server infrastructure associated with the tape robots and the software to drive the libraries and manage the tapes and data migration between different storage systems.

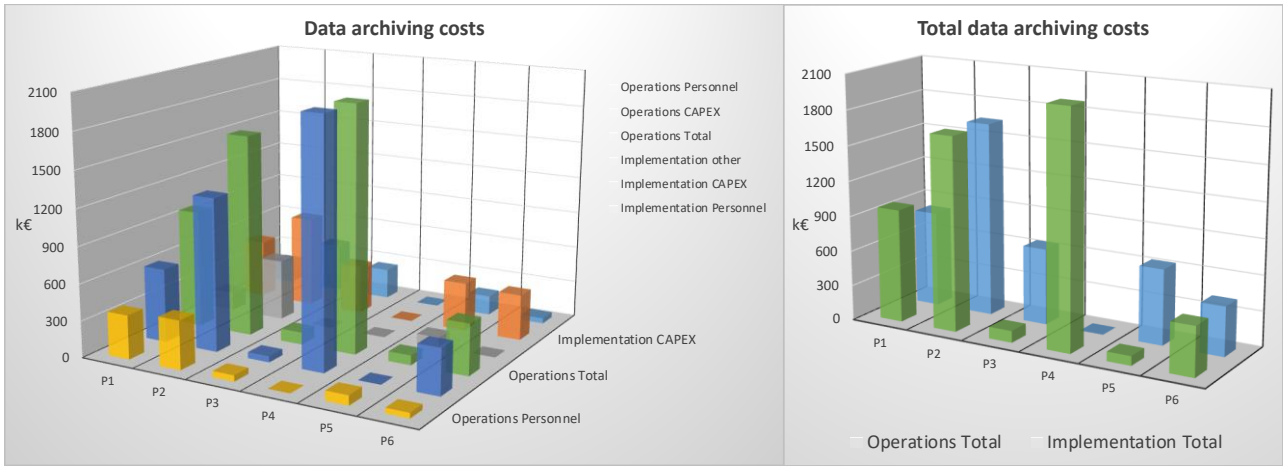


Fig.14. Initial implementation and operation costs for five years for archiving data.

In terms of costs, the above analysis shows that even with a similar cost per PB, RIs might have different costs for operating such systems, determined by the distributed structure (for facilities that are not in the same geographical region) and different volumes of data. Using the same data standards, common formats and common principles governing the data policies (embargo periods) will help to also standardize the costs.

3.1.10 Offline and online computing costs

The following data has been collected about the offline computing, i.e. CPU and GPU compute clusters, their internal network interconnects and the personnel needed to administer the computers. As of today, there is no strict separation between offline and online at most of the facilities, but this will change for many RIs in the future because of increased demands coming from users and staff.

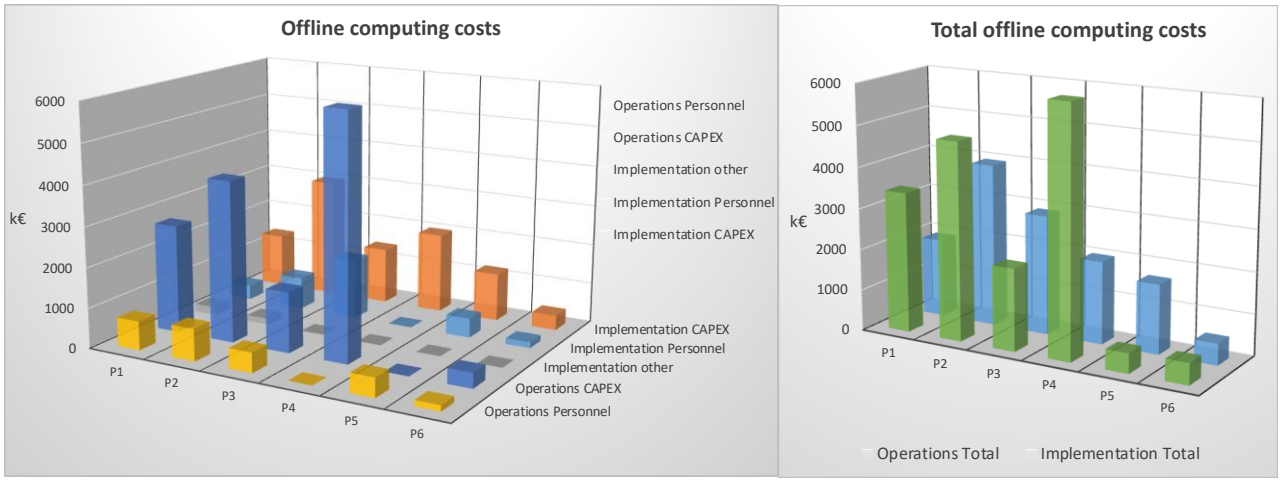


Fig.15. Initial implementation and operation costs for five years for offline computing.

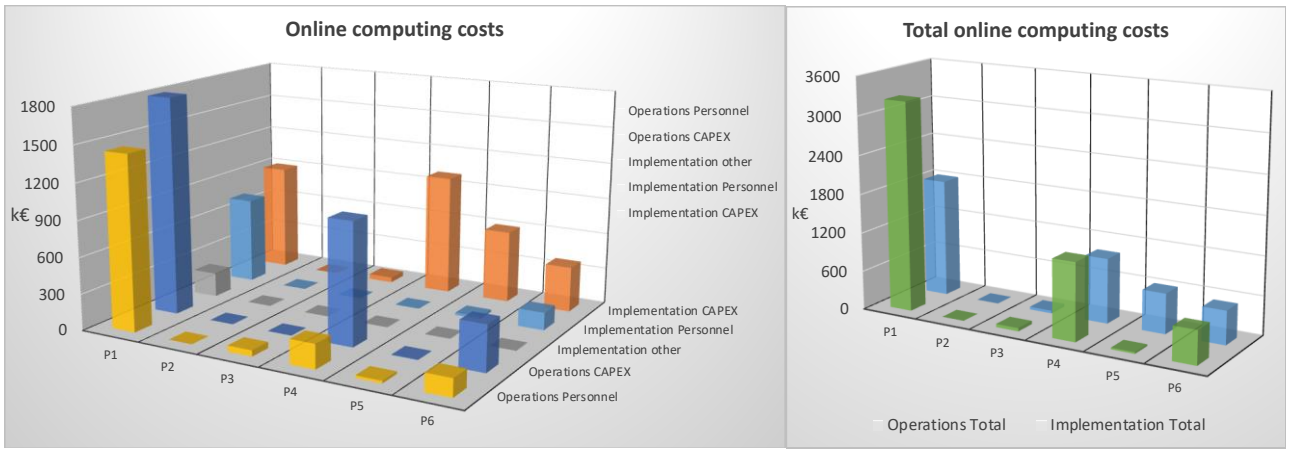


Fig.16. Initial implementation and operation costs for five years for online computing.

As expected, the online and offline computing is very different between the partners because each of them has a custom process developed for sizing this part of the IT infrastructure, as this component is highly dependent on the computing model and specific IT policies and standards. Also, this associated cost driver of the offline/online computing is, in some cases, not completely separated, as the computing resources are usually allocated for online on offline computing tasks based on the specific needs of the RI. This is why, in some cases, our partners might have presented these resources and the associated costs aggregated in one single category which is either online computing or offline computing. This is also the reason why personnel costs are also allocated differently based on the specifics of each partner's computing ecosystem.

Some partners are exploring the possibility to use commercial cloud services to cover peak demands or because it may be more efficient in case of distributed RIs. A possible provider of these services is EGI, one of the PaNOSC partners. We include in the last chapter an estimation provided by EGI on the costs for these services.

3.1.11 External computing

There is only one partner that has adopted the use of external computing provided by a commercial supplier, due to its distributed nature. In this case, on demand computing resources are fundamental for a distributed RI where not all of its nodes have the resources, both personnel and infrastructure, to offer to the users of the ERIC. In this case, the allocation of external resources is designed centrally, by the core IT team of the ERIC.

The costs involved are:

- Initial implementation: 44 k€ for personnel; 25 k€ for capital expenditures
- Yearly operation: 9 k€ for personnel; 5 k€ for capital expenditures (hardware and software).

These costs correspond to the implementation pilot that is being run at the time this document is written with two use cases, but should the pilot be successful, the costs will increase to host data processing pipelines for more instruments.

3.1.12 Networking - LAN

In this cost line we report all costs associated to install and operate the network cabling and network electronics, to connect all beamlines/end stations to a powerful backbone network allowing fast and ultra-low latency data transfers between the beamlines and the data centres of each RI (including transfers to the users). All network backbone infrastructure is based on fibre optics at most of the RIs.

Even if networking is a standard layer of the computing environment for PaNOSC partners, situations differ due to the variety of solutions going from ultra-low latency Ethernet to InfiniBand network connections.

A different and direct impact on cost is at facilities reporting a higher number of beamlines or instruments. This adds extra costs on the networking infrastructure.

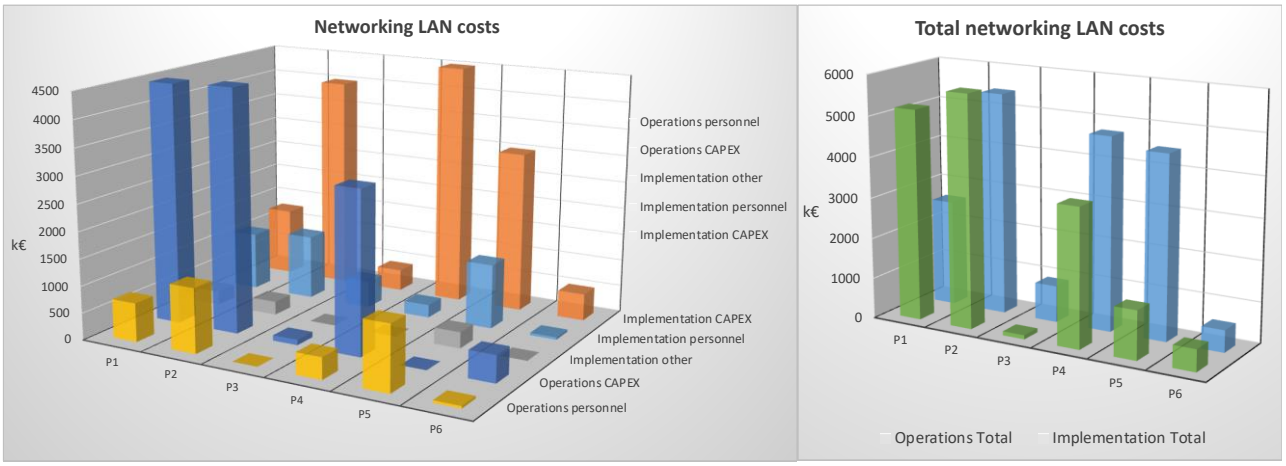


Fig.17. Initial implementation and operation costs for five years for networking-LAN.

3.1.13 Networking - WAN – hardware and NREN

RIs have subscriptions with their NRENs and the NRENs are further interconnected via the GEANT infrastructure.

The subscription cost includes the firewall and the DMZ infrastructure with advanced IPS and IDS. The costs are completely different as each RI might share some of the costs of the backbones. In some cases, e.g. for the ERICs, it is a paid subscription.

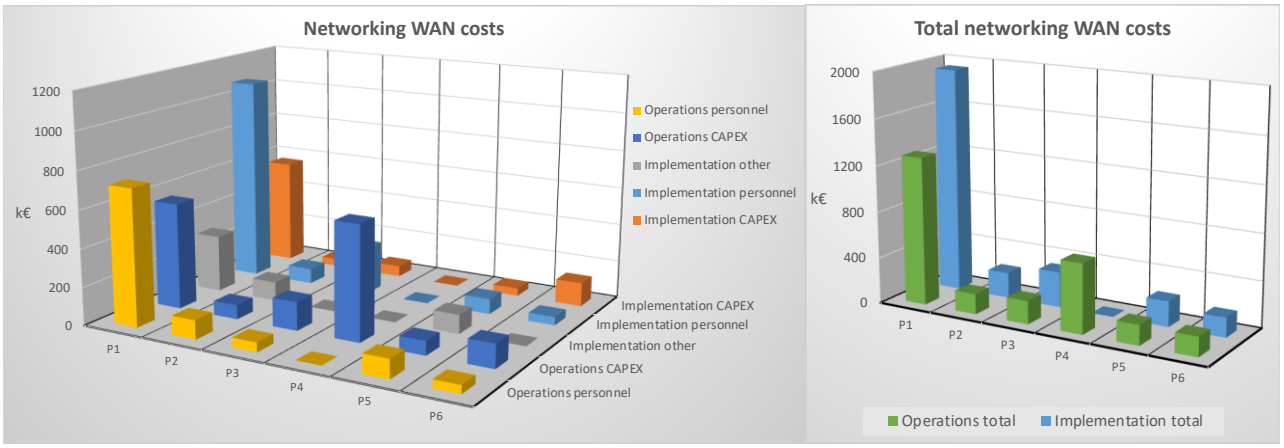


Fig.18. Initial implementation and operation costs for five years for networking-WAN.

Regarding the yearly operations costs, these are doubled in the case of P1 due to the fact that this partner has a distributed architecture and it requires a setup together with two different NRENs.

3.1.14 Remote access IT Infrastructure

In this category we report the cost incurred for the server infrastructure required for remote access including the software licenses. One of the PaNOSC partners has reported that since 2020 they have no licensing costs for the IT Infrastructure supporting remote access.

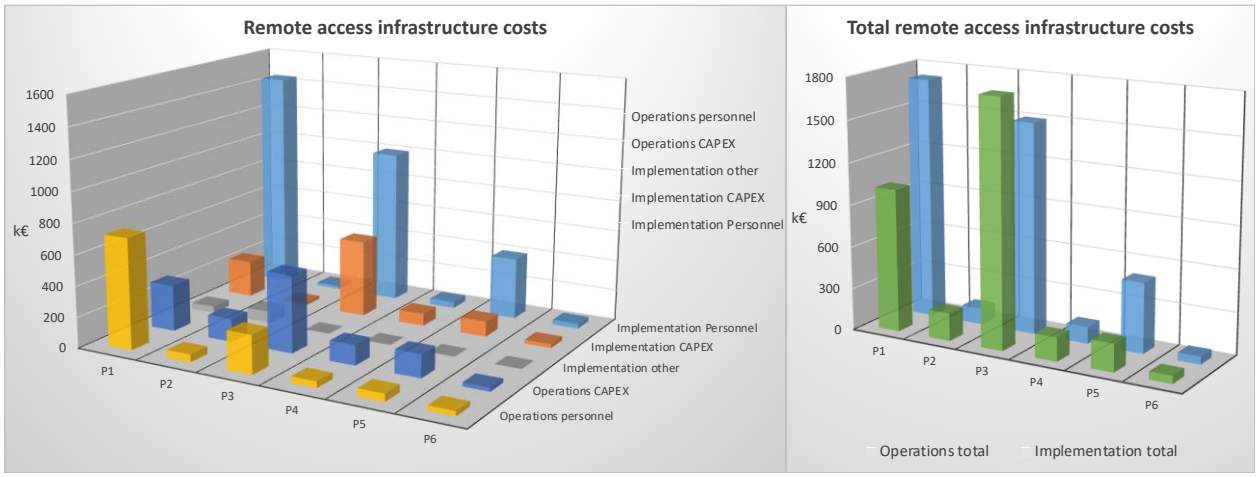


Fig.19. Initial implementation and operation costs for five years for remote access.

Other partners have different costs for this option as they are using different tools and services, especially since for new RIs (ELI, CERIC ESS) some of the solutions are currently under development, a process that is adding extra personnel costs during the implementation phase.

As a commonality for different scientific communities, the costs reported for the implementation of private cloud architectures, like OpenStack, requires qualified IT Engineers dedicated for the development and integration of facility specific requirements. At the same time, maintaining such a solution adds extra personnel costs, as the operations team has to prepare, maintain and scale the cloud architecture supporting every day operations of the facility.

3.1.15 Users support services

Each RI has an operation strategy, which in most of the cases also includes a standby intervention team to deal with issues on the network and server infrastructure, in some cases including 24/24h. In addition to this, there is also support staff that covers the daily operations and support functions, such as help desk tickets that have to be resolved in a timely manner. The support staff also takes care of all requests for assistance falling under the IT infrastructure.

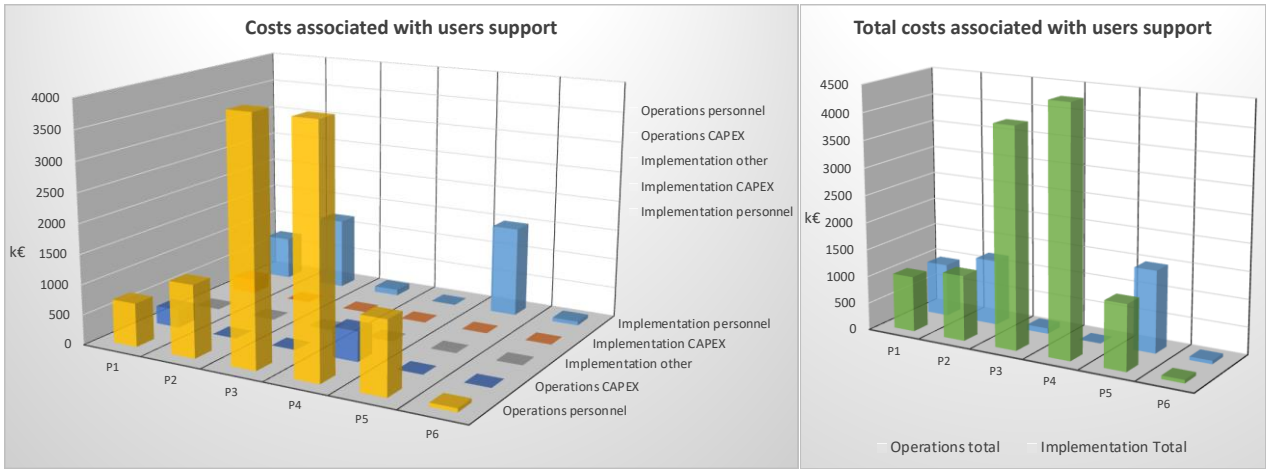


Fig.20. Initial implementation and operation costs for five years for the Users support service.

There is a big difference in the cost model as each facility might allocate the user support differently and in some cases those costs might be included in the IT operations. Although large facilities report higher costs than smaller ones, there is no direct correlation between the user support costs and number of users, datasets generated, or other, for the reasons explained above.

3.1.16 Licenses

The graph below is representing the costs reported by each PaNOSC partner for commercial software (such as Mathworks, Mathematica, ANSYS, etc). These costs vary amongst partners according to the extent to which they develop their software, adopt open software, or use commercial software requiring licenses. This choice is up to each RI and there is no common trend to highlight.

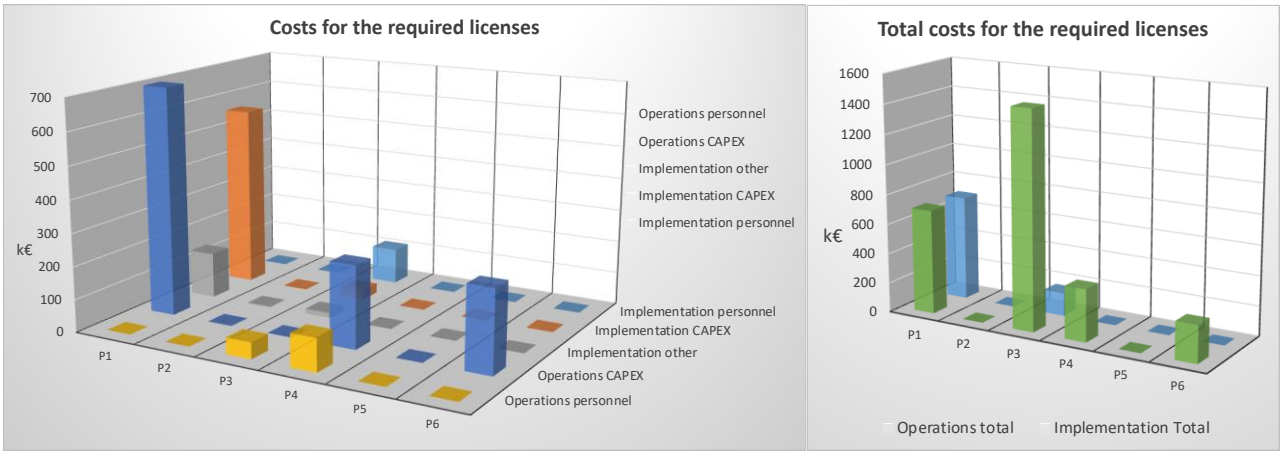


Fig.21. Initial implementation and operation costs for five years for licenses.

3.1.17 Computer Room

Computer rooms are a major cost driver for all partner RIs. Data centre equipment is pretty standard and all RIs are following the industry

standards in terms of redundancy and fault tolerance (Redundant power supplies, UPS, HVAC).

This can be seen in the chart below, the strategy is pretty similar and all partners have comparable costs, both for the implementation and operations.

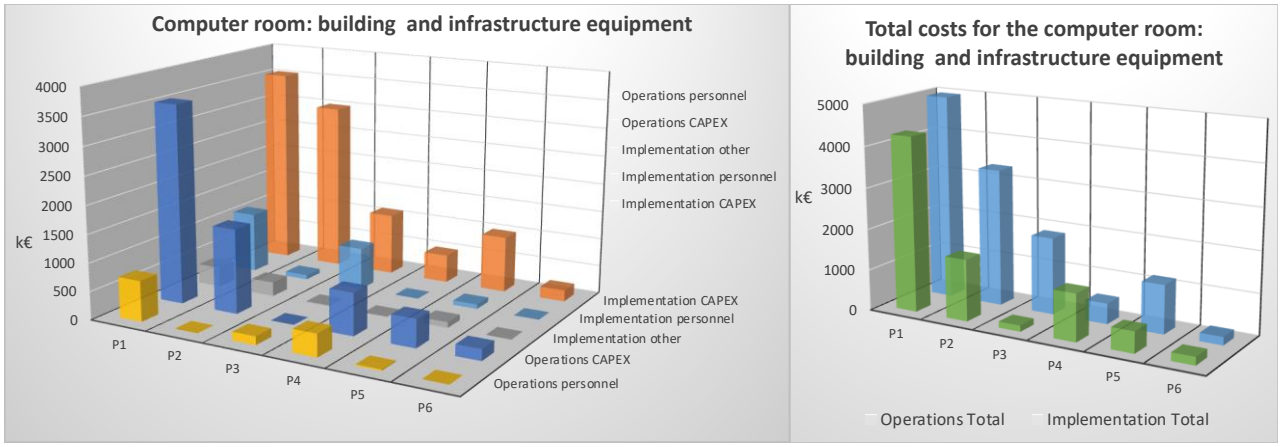


Fig.22. Initial implementation and operation costs for five years for the computer room.

Though the Data centres have similar TCO (total cost of ownership) reports, the implementation and operation of distributed data centres adds to the cost to cover disaster recovery scenarios and higher operation costs. Data centre cost drivers are different and impacted by factors like the CPU density/rack, particular co-processing capacities that might be required (GPU/FPGA), different networking LAN strategies (Infiniband/Ethernet) or the type and amount of storage. All these factors are impacting the auxiliary costs associated to a data centre, mainly focusing on preparing the room for the utilities (industrial cooling and piping system, or hot air extraction tubes etc.).

The same situation appears for newly established RIs that are now starting to define and build their ICT infrastructures. In this case, the partners are reporting the costs based on the current sizing exercises, based on the data volumes expected to be generated and, in some cases, based on the feedback from the commissioning of experiments. In this particular case, each partner is working on preparing the data rooms/data centres that will host the computing infrastructure of the facility. These costs are based on commercial price quotations of the different architectures they are considering.

3.1.18 Computer room utilities

This line is mainly related to the cost of electricity, but also to the general upkeep of computer rooms. In some cases, these are just estimates as computer rooms are currently still being designed. However, also the RIs with computer rooms that are currently in the design process, have clear cost figures as they already have smaller computer rooms used during the experiment commissioning process.

The graph below presents the costs of the utilities needed to support the

computing environment, costs of the UPS, cooling, fire extinguishing and other associated costs included in the auxiliary systems category.

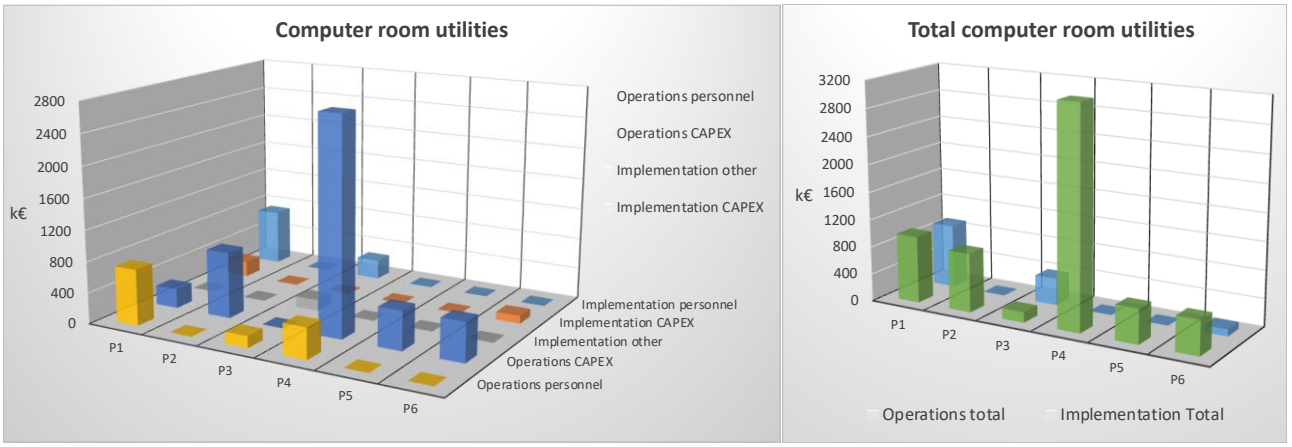


Fig.23. Initial implementation and operation costs for five years for the computer room utilities.

The costs are different from one country to another as there are specific contracts negotiated with utility providers and there are different funding strategies/models for each RI. However, as a cost driver, this is present mainly in operations. There is not an initial cost identified for setting up the infrastructure as most of the mature facilities already have this in place.

3.2 Data processing linked to EOSC

As anticipated in the introduction, only one of the partners had, at the beginning of the project, experience with the reuse of datasets no longer under embargo. As the developments of the project proceed, partners are becoming ready to provide access to datasets and services for their community at large. There are two factors to consider for the costs of data processing linked to EOSC. The first is that generally RIs don't have a budget assigned for these activities. The second is that the costs provided are an estimation according to the expected demand, or to the best effort facilities can offer, but there is no evidence to date to confirm these estimations.

The costs provided reflect different hypotheses formulated by the PaNOSC partners. In one case, the services that will be provided follow a ratio between the PaN community and the EOSC, currently estimated at 3/4 vs 1/4, meaning that precedence will be given to the PaN community. However, this ratio may/will evolve in the future.

Other partners estimated that in the near future, approximately 10% of the data that will no longer be under embargo, will be accessed by the community, without distinction between the PaN community and EOSC. Similarly, it was proposed that part of these data would be analysed on site. The percentage of data that needs to be analysed onsite depends on the facility, as in some cases it is not possible or practical to move the datasets. Some partners are also testing, with small scale pilots, the possibility to rely on commercial providers for data analysis and

storage, or to cover peak load demands with off-line computing. This can be either achieved by purchasing capacity from hyperscalers or by agreement with EU HPC centres such as PRACE or FENIX or national HPC centres.

There was also differences for the allocation of initial investments: some partners decided to allocate the costs for the implementation as part of the facility operation, as they consider that even without the EOSC they would have provided similar services to their users from the PaN community. In this case there would be little additional costs for extending those services to a broader community and those would be mainly operation costs. The budget most RIs can dedicate currently to this activity is very limited. This allocation is somewhat theoretical and based on the current constraints, but it will require a revision if RIs get a proper budget for EOSC activities and are mandated to provide data and services for EOSC (with no limitation on who the final user is).

Some of the facilities are still working on adopting a FAIR Data Policy, thus not really facing the same challenges and costs. We expect to receive a more accurate list of costs once all the PaNOSC and ExPaNDS partners have confirmed the adoption of the tools and services developed in the projects. One of the partners preferred not to provide costs since they considered it was too early to provide a reasonable estimation of costs.

For all these reasons, the analysis we can perform with the EOSC data is necessarily limited in scope. We represent the costs reported for information, but these costs will need to be refined in the future when looking for a sustainable business model for participating to the EOSC.

3.2.1 AAI

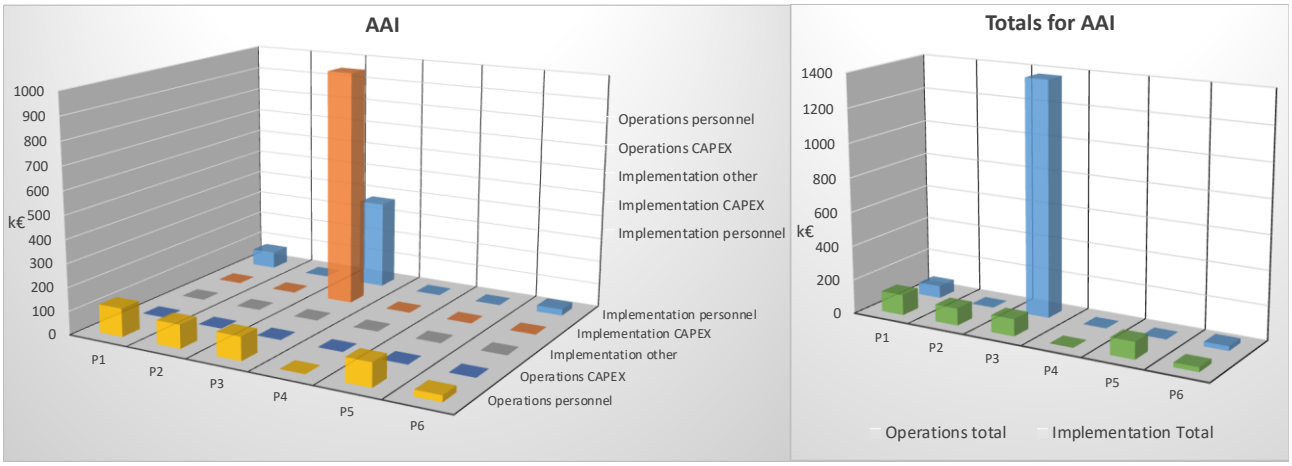


Fig.24. Initial implementation and operation costs for five years for the AAI system associated with EOSC.

In this cost line we report the human resources required to maintain and expand the use of the PaN AAI system at each RI. This includes the time to validate authorisations for the accounts. Some extra resources might be needed for the outsourcing of the Identity Provider (IdP) services.

Most of the partners reported the implementation costs amongst those of operation of the facility. The evident difference for P3 resides in the allocation: P3 had foreseen the reusability of data even after getting

engaged with the EOSC, therefore the costs reported at this point would have occurred in any case to allow reusability independently from their engagement in EOSC.

Some partners are placed in scientific parks or other ecosystems that allow them to use the services provided by other entities, which are those that bear the costs for implementation, maintenance and operation.

3.2.2 Interoperability of the file catalogues

As in the previous case, most of the costs reported correspond to a fraction of the operation costs. P1 and P6 report some costs of implementation since for these partners there is the need to implement interoperable catalogues, while the other partners had chosen solutions in the past that are already interoperable, and do not require any cost for the implementation. Some personnel costs allocated here are due to a foreseen constant need to adapt the catalogues for interoperability, with the evolution of the PaN community and as the implementation of the EOSC moves ahead.

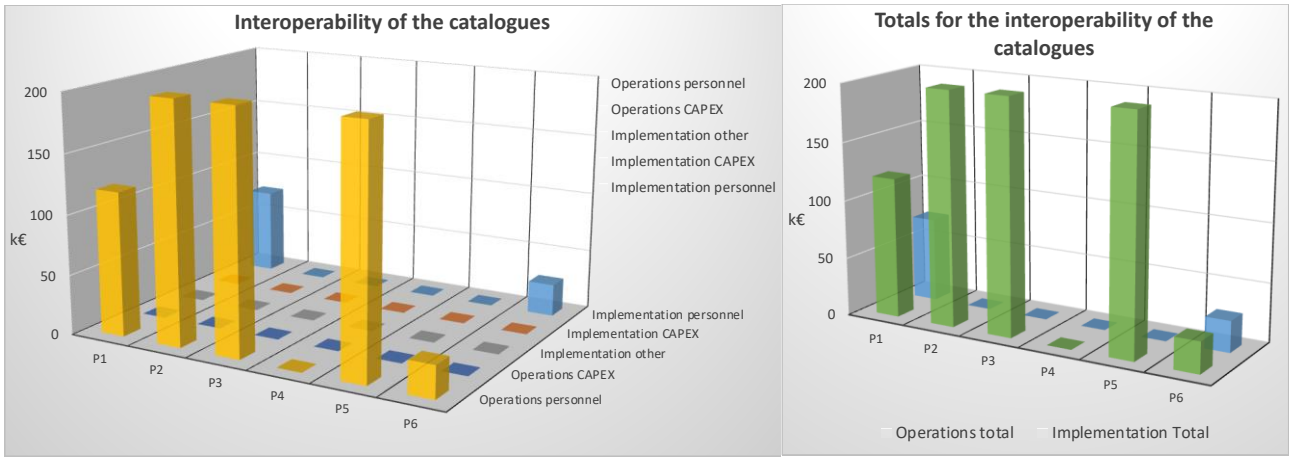


Fig.25. Initial implementation and operation costs for five years for the catalogues' interoperability associated with EOSC.

3.2.3 Data Portal

For the data portal as well, partners have reported mostly operation costs related to the limited fraction of EOSC users they would serve. The PaN data portal under implementation in PaNOSC is considered a standard service to the PaN users in any case, so these costs are not allocated to the EOSC at all or allocated only to a small fraction as personnel costs due to the need to make the portal accessible not only to PaN users but to a larger community.

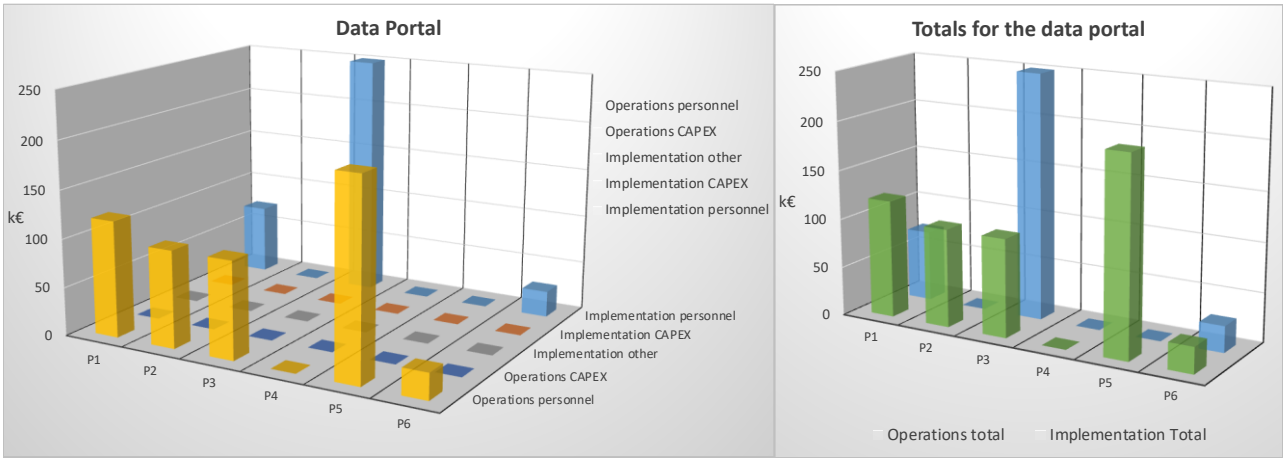


Fig.26. Initial implementation and operation costs for five years for the Data Portal associated with EOSC.

3.2.4 Software Catalogue

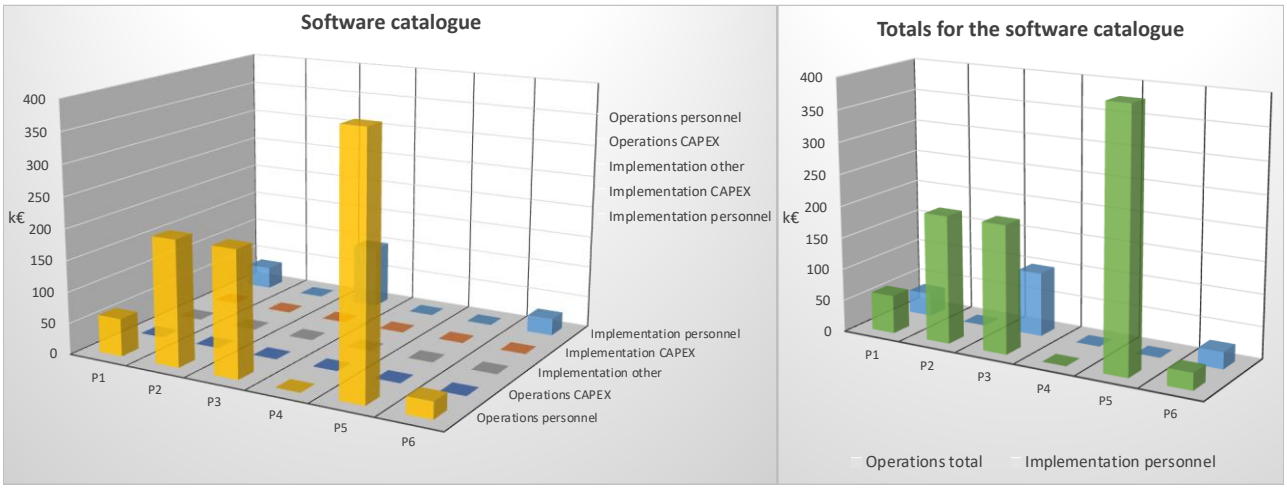


Fig.27. Initial implementation and operation costs for five years for the software catalogue associated with EOSC.

The personnel costs reported here correspond to the human resources required to maintain and extend the PaN community software catalogue. Although this does not include working directly on developing any of the software packages made available through the catalogue, it includes the effort needed to have EOSC infrastructure in place. In this way, partners are assuring the software packages are integrated into the catalogue and meet the minimum quality criteria required (e.g., in particular for documentation, training, periodical updates of the software).

3.2.5 Curation of data archive (data deletion)

Depending on several ownership related aspects and also considering the fact that some of the partners have committed via their data policies to have some data stored forever (as it's the case for the metadata), this cost is driven by the need to curate the data before making it open and deleting the data based on the internal rules and procedures of each RI.

Based on the collected cost information, we can see this process adding a personnel cost for the implementation of the necessary tools/services and setting up the data curation processes, the rest of the implementation costs are not present as this curation process mainly relies on the existing computing and storage capacity.

From the operations perspective, the data curation process adds extra operation costs. In some cases it could also add CAPEX costs as, depending on the volume, the quality and the importance of some data sets, this data curation process could require upgrading some of the existing infrastructure.

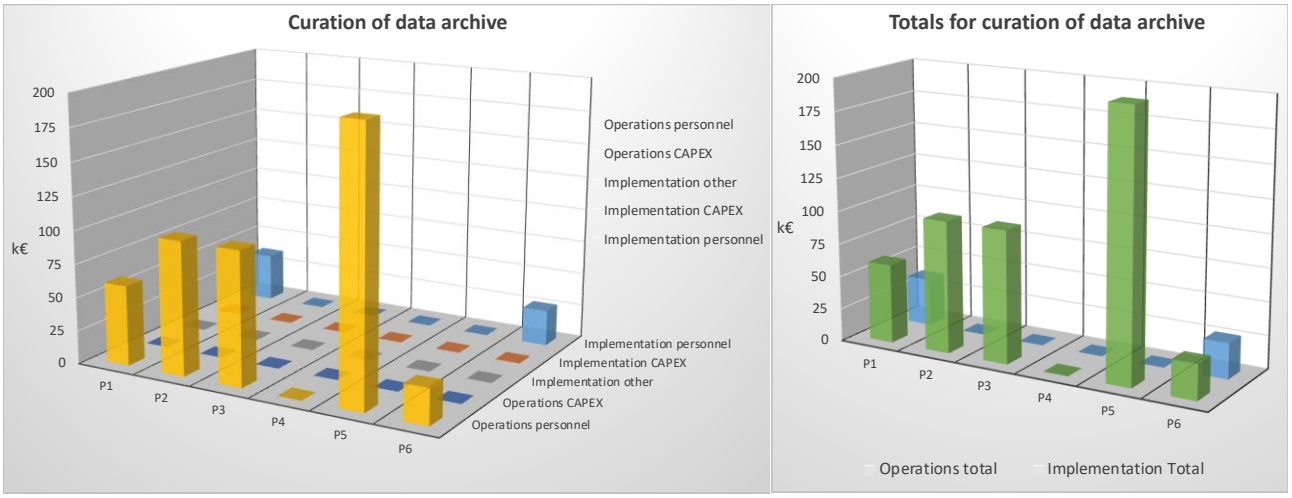


Fig.28. Initial implementation and operation costs for five years for data archive curation associated with EOSC.

3.2.6 User support and training

Remote access to the data portal and the associated EOSC services will require user support and up-to-date training material to ensure a smooth and productive operation. The amount of support and training a facility is willing to offer will depend on the level of the services it wants to provide. However, there are some activities that will be mandatory, such as ticket resolution via the EOSC helpdesk.

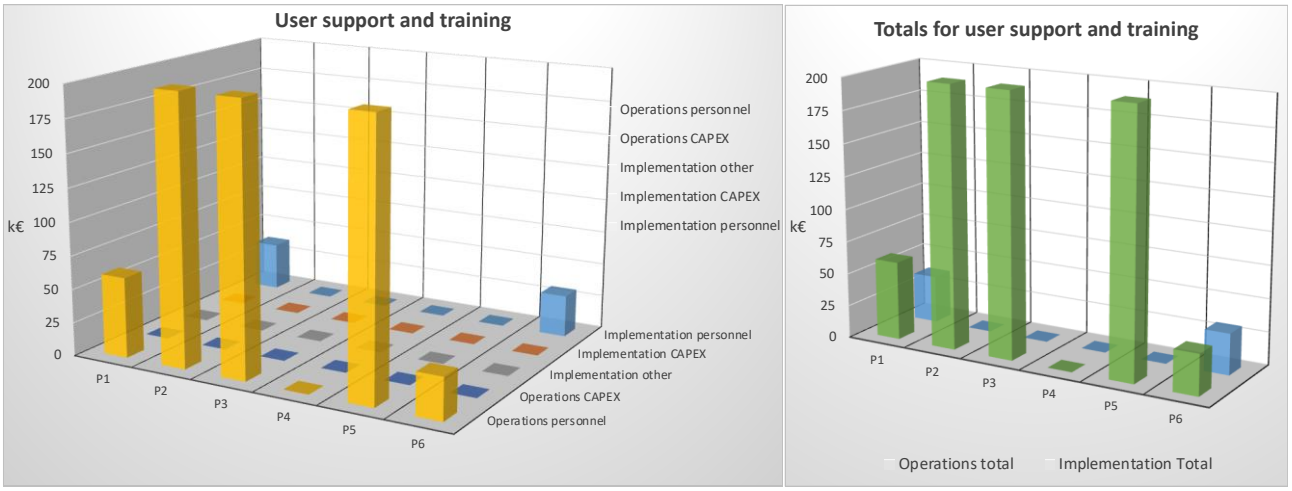


Fig.29. Initial implementation and operation costs for five years for the users support associated with EOSC.

3.2.7 Outreach

Fig.30. Initial implementation and operation costs for five years for the outreach activities associated with EOSC.

In this cost item we report the estimated effort required to make our services and data catalogues known beyond our direct user community. The implementation personnel costs originates from setting up initial processes supporting the communication groups, such as regular data stewardship training and data custodianship, initial implementation and development of such trainings, and dissemination processes to nurture a data stewardship and data custodian culture across RIs. Outreach costs also include the organisation of user workshops, to engage with the user communities for advocating FAIR and Open Science.

Although partners have reported what they would spend individually in outreach according to their company culture, we expect that part of this would be a joint effort, since many of the tools and services will be shared by PaN Facilities.

3.2.8 Network bandwidth for EOSC, data transfer software

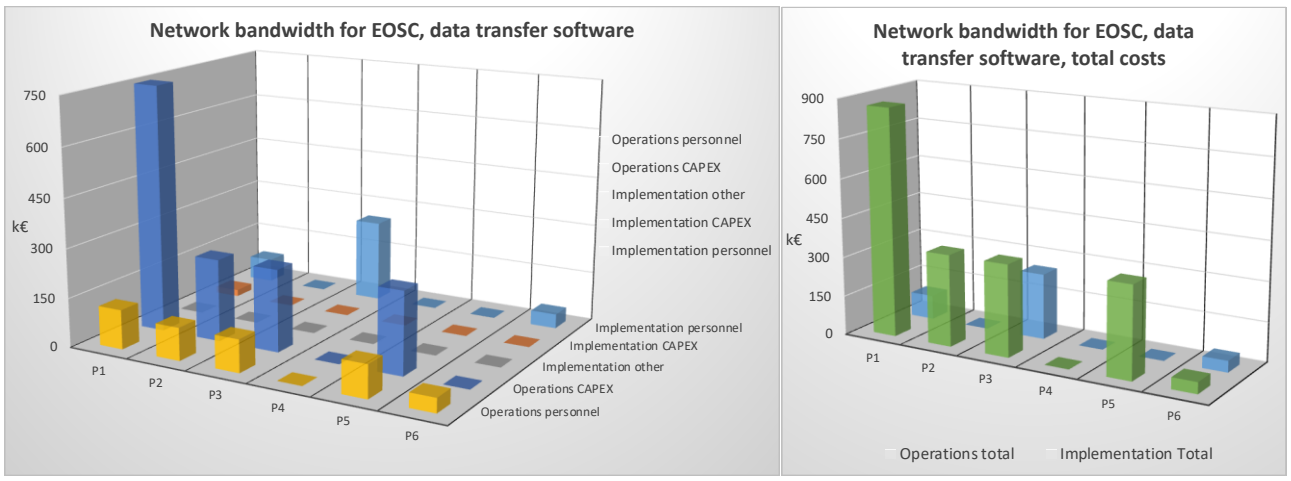


Fig.31. Initial implementation and operation costs for five years for the network bandwidth/data transfer software associated with EOSC.

The bandwidth that will be dedicated to the PaN and EOSC community for accessing RI services and data catalogues was estimated by the facilities based on their forecasts of use and current technology or budget constraints.

Some of the PaNOSC partners are confident that they will be able to dedicate 20-25% of their Internet bandwidth to the PaN and EOSC community. One partner has also included a small provision for the license cost of data transfer software, although at this stage the real cost for this remains unknown. Collecting financial information after a couple of years, once several facilities will have provided open access to a broad community will allow a more accurate estimation of costs.

3.2.9 Storage for EOSC

Linked to the data portal, a dedicated storage area is required for staging archived data from the tape archive. The disk storage capacity required for staging is small compared to the overall storage capacity, however, this may quickly become insufficient if the data catalogues of partners become popular.

Although the storage implies investments, most of the partners reported exclusively operation costs. This is again a matter of approach: RIs considered that they need this kind of infrastructure as a condition to provide FAIR data, independently of the existence of EOSC. For these reasons, the costs (e.g. implementation) are not included here.

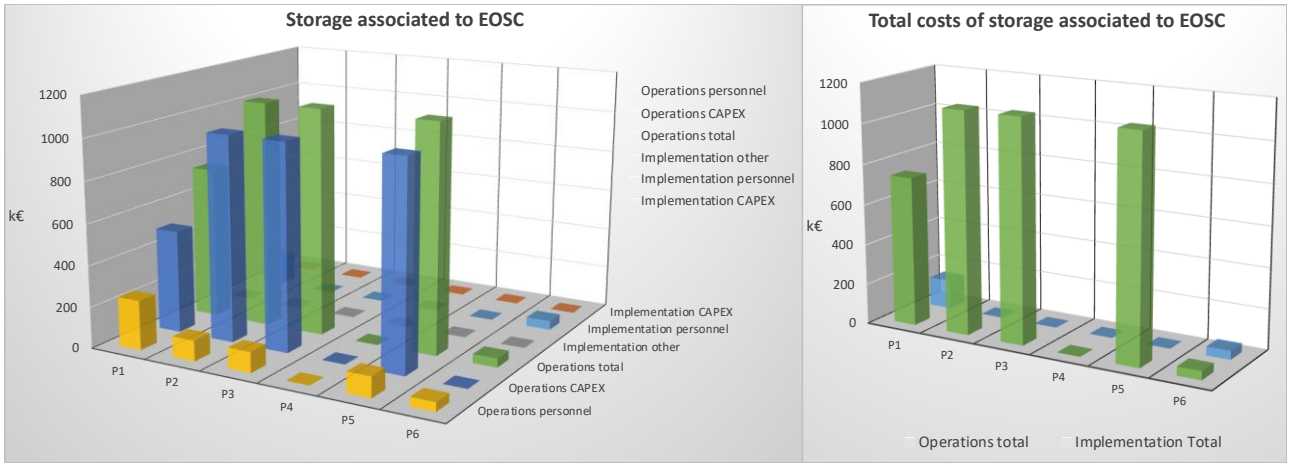


Fig.32. Initial implementation and operation costs for five years for the storage associated with EOSC.

3.2.10 Offline computing for EOSC

A number of server computers is required to allow PaN and EOSC users to interact with the data from RIs catalogues. Computers will allow data visualisations and data processing. The amount of offline computing that can be offered for EOSC is currently limited by the available infrastructure or budget; it can (and probably will need to) be increased

in the near future if there will be a high demand from EOSC users and dedicated funding for EOSC activities. Some partners are considering and testing in a pilot scale, on-demand computing capacities provided as commercial services to cope with the demand from EOSC, but this solution is suitable only for some RIs and not feasible for others. Monitoring and managing the resource allocation has to be carefully done to avoid any abuse or misuse such as orphan jobs using up capacity.

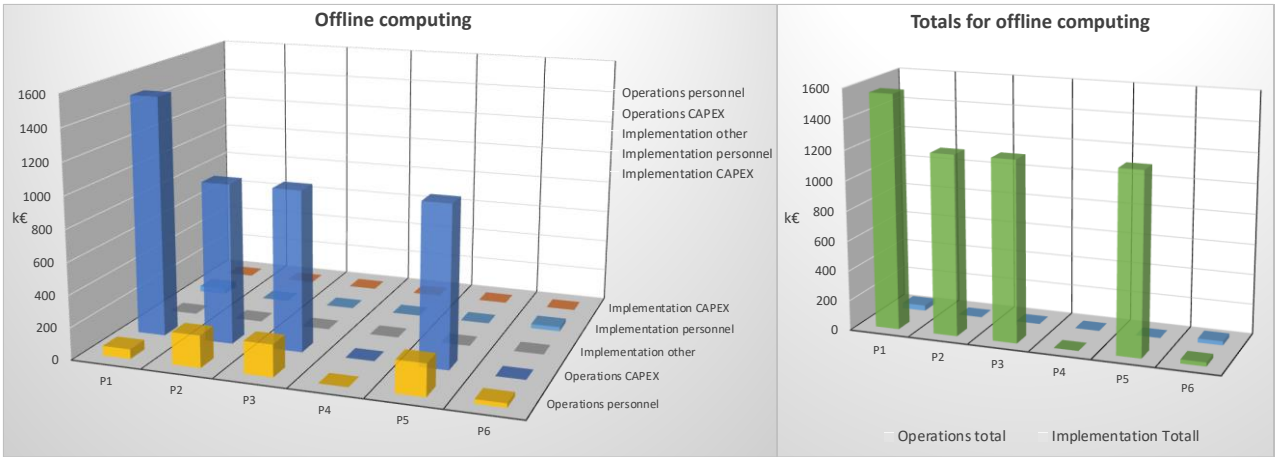


Fig.33. Initial implementation and operation costs for five years for offline computing associated with EOSC.

3.2.11 On-line computing for EOSC

In the context of the PaNOSC project, each partner RI is actively engaged in identifying ways to implement the necessary services, together with the first set of relevant metrics and controls that will allow them to identify what type of support is required by the open science users and what are the costs and risks associated with this kind of support.

In the image below, depending on the level of maturity of each organisation and the particular architecture of their existing computing systems, some partners can already provide some basic limited online computing capacity. In these particular cases, our partner RIs have reported zero implementation costs.

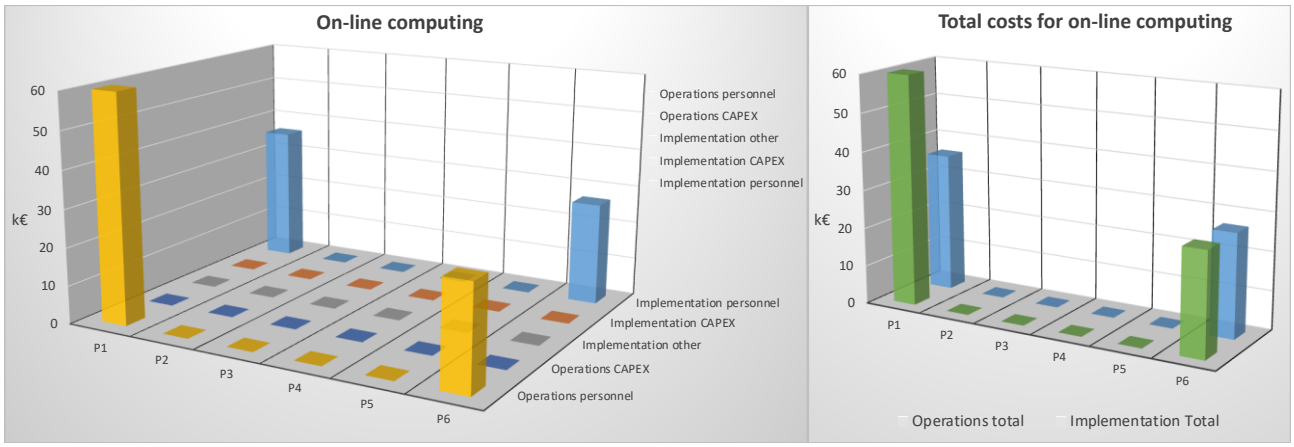


Fig.34. Initial implementation and operation costs for five years for on-line computing associated with EOSC.

3.2.12 Data archive beyond facility data policy

The costs presented in the chart below have been identified by the partners as costs that could be triggered by having data archived beyond the data policy. They vary among partners due to their data policy (how long the data will be preserved in general) and specific performance metrics adopted by each RI (e.g. one can decide that some datasets are extremely valuable and thus, based on the interest of the open science community, could decide to archive such data for longer periods of time).

From the technical perspective, and in some cases even from the implementation perspective, this does not generate a lot of additional cost right now, but once the demand will increase and with new detectors and machine upgrades that will generate increased amounts of data, most of the partners are anticipating an increase of both CAPEX and operations costs not sustainable with the current RI budgets.

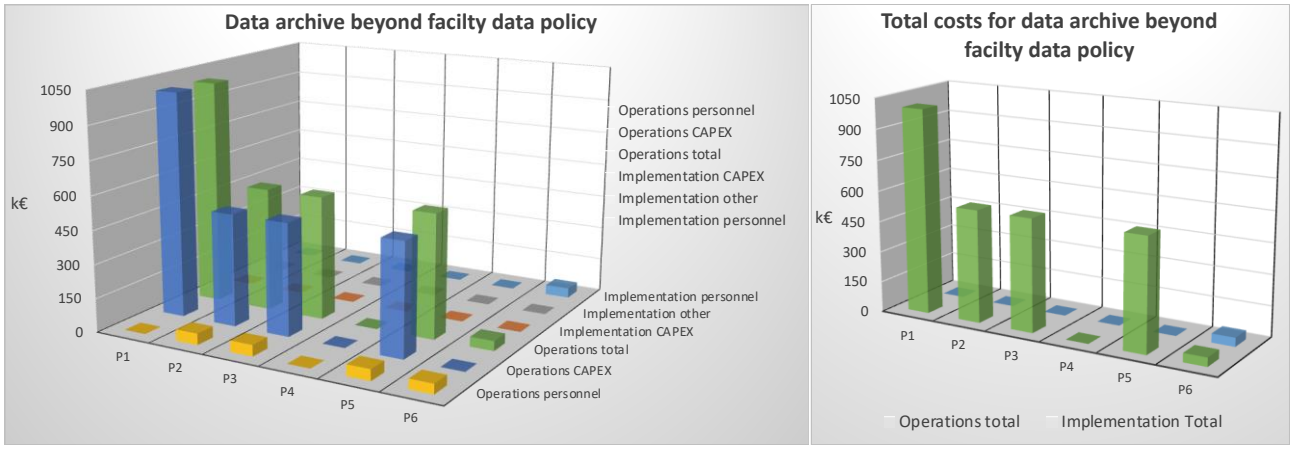


Fig.35. Initial implementation and operation costs for five years for the data archiving beyond the facility policy.

Based on the data collected from all PaNOSC partners, it is clear that there is an impact associated with EOSC. However, it can't be easily defined as it is integrated in the data management system of each PaNOSC partner.

Some of the costs are extracted by the partners from the total data operations and data management. Trying to extract from those costs what is linked to the adoption of FAIR principles and EOSC is extremely complicated as some of the partners have existing tools that can be used, while for new RIs these tools are in the early stages of development.

Though adopting FAIR principles enables PaNOSC partners to support EOSC and open science, the cost of these services is highly dependent on the particular computing ecosystem, computing strategies and computing model of each partner.

3.3Cost of services offered by external providers

PaNOSC partners have the necessary infrastructure to serve their own users, that is, the researchers that perform experiments in their facility and need to access their datasets, analyse them, modify them, assign a DOI, etc. The amount of users facilities host is well known, as well as the operations they perform on their data. However, the future demand from EOSC users, including the long tail of science and citizen scientists, is uncertain and there is no way to estimate it or to understand how it will evolve over time, since the services have been developed and deployed only very recently. For this reason, in the hypothesis that some services could be deployed using services of an external provider, we asked the PaNOSC partner EGI to provide us with a cost estimation for some services that could be useful to PaN RIs where researchers have to get access to the data generated during an experiment at one of the PaNOSC facilities or perform analysis on datasets in a computing infrastructure that is distant from the data infrastructure, and store back results at the same facility.

To support this use case in the PaNOSC community the following EOSC services have been proposed:

- The federated and distributed cloud computing and storage resources of EGI.
- The EGI Notebooks service, based on JupyterHub technology, is used as an agile environment to access and process data sets. Users are enabled to easily share concepts, ideas and working applications, and - combined with Binder - the Jupyter notebooks are allowed to be reproducible and reusable by anyone, anywhere.
- The EGI DataHub solution to access data conforming to required policies (e.g.: unauthenticated open access; access after user registration or access restricted to members of a Virtual Organization), and bring data close to the computing to exploit it efficiently.

This is an alternative infrastructure that can be used for computing open data published by PaNOSC facilities, in other words, data that is beyond the embargo period. In this approach, the user has to install the necessary data analysis software and transfer the data sets to EGI storage.

The costs to operate the EOSC services described in the architecture above are reported in the following table:

Computing (provisioning)	Computing resources allocated to support the use case described above:
	• OneProvider: standard.2core-16ram: (2 vCPU cores, 16GB of RAM, and 80GB of local disk)
	• OneZone: hpc.16core-32ram: 16 vCPU cores, 32GB of RAM and 80GB of local disk)
	• Kubernetes:
	  ○ 1 Master (hpc.8core-16ram): 8 vcpu cores, 16GB of RAM and 80GB of local disk
	  ○ N Workers (hpc.8core-16ram): 8 vcpu cores, 16GB of RAM and 80GB of local disk

	The break-down of the costs for the provisioning of the computing resources in the EGI Infrastructure is the following: • OneProvider: 171.18€ per month • OneZone: 59.56€ per month • Kubernetes: >754.29€ per month (depending on the number of Workers considered in the cluster) The cost of the Infrastructure AAI Proxy service (used to provide access to cloud operators and community managers: 0.5PM/year (4150€ including 25% overhead to the PM rate).
Storage (provisioning)	The costs for the provisioning of Ceph storage is 32.50€ (including the 25% of overhead to the PM rate) for TB/year for size from 100TB to 1-2PB.

Table 1. Costs to operate the services provided by EGI

Operation and maintenance (of the services)	**EGI Notebooks** = The cost to operate the initial set-up of the service (including the Binder support), collecting technical requirements from the community is about 1-2 PMs (7 900€ - 15 800€ including 25% overhead to the PM rate). The costs to maintain the EGI Notebooks (including the Binder set-up) service in production, update the documentation, and provide technical support to users is: 0.6 PMs/month (645€ including 25% overhead to the PM rate). **EGI DataHub** = The Onezone component is hosted at one of EGI's cloud computing sites. System requirements are detailed in the following table: **Requirement** / **Minimum** / **Optimal** No of VMs / 1 / 2 + 1 for every 5000 users CPU / 8 vCPU / 16 vCPU RAM / 32GB / 64GB Local disk / SATA / SSD Local storage space / 20GB / 40GB System requirements for the Oneprovider components are detailed below: **Requirement** / **Minimum** / **Optimal** No of VMs / 1 / 2 + 1 for every 500 concurrent users CPU / 8 vCPU / 16 vCPU RAM / 32GB / 64GB Local disk / SSD / SSD Local storage space / 20GB + 8MB for each 1000 files / 40GB + 8MB for each 1000 files The local storage space is used only in case of the instance installed at the EGI cloud, while for Oneproviders installed at facilities the storage is accessed usually via NFS so it's not local to the service. Both Onezone and Oneproviders are operated via Docker containers, hence the operations and maintenance cost for upgrades are quite limited. The costs to operate the initial set-up of the EGI DataHub service taking into account the requirements from the community is about 1

	PM (7 900€ including 25% overhead to the PM rate). The costs to maintain the EGI DataHub service in production in the EGI Infrastructure (Onezone + Oneprovider), update the documentation, and provide technical support to users is: 0.5 PM (3 950€ including 25% overhead to the PM rate). The cost to operate the Oneprovider at each facility depends on the different PM rate of each institution, but it's in the order of 0.5 PM per year. **EGI Check-in** = The costs to maintain the integration of the community proxy (UmbrellaID) with the e-Infrastructure proxy operated by EGI, and provide technical support requires 0.5PM (4 150€ including 25% overhead to the PM rate).

Table 1 (cont.). Costs to operate the services provided by EGI

3.4 Metrics for data, software, and the EOSC

So far we have presented an overview of the costs associated with data management, provision of FAIR data and linking to EOSC, and these costs are sustained by the facilities because there is an added value that comes with them. As frequently happens for research infrastructures, this value is hardly an economic return but it is mostly socio-economic. We expect that the adoption of FAIR principles and open data will maximise our scientific outputs and impact.

In the following we describe some metrics that can be used to monitor the outputs. The impact and added value will be addressed in more detail when defining the business models for the PaN EOSC.

Use of data and software:

a. Number of FAIR datasets generated
b. Number of times the data generated at the facility is cited in publications
c. Number of downloads of the software packages in the software catalogue.

Reuse of data and software:

a. Number of downloads of open data sets
b. Remote data analysis operations (via the data analysis tools), number of VMs
c. CPU hours used in the case of more intense computing tasks
d. Citations of data sets by research groups not related to the original proposal
e. Number of uses of the software in the software catalogue, by users not related to the group having generated the datasets.

Impact of EOSC:

f. Increase in the number of publications associated to the facility

We propose a single indicator for the impact of EOSC because indicators should be easily measurable and robust. The impact we expect to be much broader than that, but with the current infrastructure we cannot identify other indicators that can be measured easily. The number of publications per se is a performance and not an impact indicator, but we propose the increase as an impact of EOSC, i.e. the increase in the productivity of the facility due to the publications that result from the reuse of data and software. The productivity in terms of publications, and their impact factor, is still today one of the most widely used parameters to assess the scientific performance of RIs during evaluations.

4 Conclusions and future steps

4.1 Conclusions

Task 7.2 allowed us to collect for the first time in a structured manner the costs inherent to the setup and operation, production, conversion (when necessary) and management of FAIR data for the PaNOSC partner RIs. Moreover, we managed to produce a first estimation of the necessary effort to provide data and services linked to the EOSC. Although some this data may not be accurate, PaNOSC partners made their best effort to report all the costs they could collect or extract and therefore we can consider this as a very good approximation. The differences inherent to the nature of our RIs (synchrotron radiation sources, neutron reactors, lasers…), their size, culture, strategic approach to user service, and the technological solutions adopted, introduced a degree of complexity that did not allow to identify parameters to feed into a model allowing for example a new RI to estimate cost for data management according to the number of beamlines, data volume produced, users served or others. We studied all possible correlations but there were no obvious ones because different technologies imply a change in scale, therefore the comparison or parametrisation was not possible.

However, we hope that new facilities find our collection useful in terms of the order of magnitude of the budget involved in data management and in the relative weight each cost category may have. The weight of a cost category, as we have seen many times in the previous chapters, depends a lot on the IT strategy of the RI, on the specific needs and also on the management strategy, since it is up to the management of the facility (and its funders) to decide on the level of services offered to the user community.

The costs collection was challenging due to many differences in the approach of RIs, their structure, their IT strategy and even their accounting systems. The need to keep the identity of the facilities undisclosed made it difficult to make an analysis that allows to the reader to identify the sources of the variations in each category, however, it is still possible to obtain useful information from the data collected and presented in this document.

Every cost category was accompanied by a brief analysis that we expect will make it possible to understand the main sources of the costs and their variations. However, as emphasised several times, some of the costs reported here are estimations, in anticipation of a near future with services fully developed and deployed. There is no experience, statistics or data on actual use to the date that could support some of the estimated costs. Moreover, the current RI budgets imposes strong constraints for the participation to the EOSC, so these constraints were the starting point for most of the partners. With these budget limitations, facilities will serve EOSC users on a best effort basis.

PaNOSC partners considered most of the developments needed to link to EOSC as part of costs inherent to the RI (and not as costs associated with the EOSC) since many of them see these services as necessary for the PaN community, independently of the EOSC. The main contributor to the cost is then represented by investments in software and hardware dedicated to the

operation.

The operations are in many cases higher than the initial setup costs, if we consider a yearly average. This means that facility managers will see an increase in their operation costs linked to the provision of FAIR and open data, and even more if they will make the same services available to EOSC users. Although PaNOSC partners are striving to accommodate the needs EOSC users, the current available budget of most RIs impose strong limitations to the amount of users that can be served, especially when EOSC users would require computing capacities at the facilities. This exercise was useful as a first approach to quantify the costs related to linking with EOSC, but the source of funding or other resources is still not clear for most PaNOSC partners.

An alternative solution, using resources (e.g. computing and storage) offered by an external provider was also considered, and the costs reported. This could be the way out for facilities that do not have enough infrastructure, and could also be a possible model for the EOSC users to deploy the computing demanding services of the RI. An additional advantage would be to have more flexibility to adapt to an uncertain demand, considering that none of the PaNOSC partners has long term experience in providing such services, so the requirements are unknown.

If the EOSC would provide this kind of resources, RIs could make most of their services available with a reasonable effort.

4.2 Future steps

The cost collection was carried out with the objective of defining a business model and a sustainability plan for the PaNOSC. In addition to the costs reported here, there will be additional costs to be taken into account by PaNOSC Facilities, as by the end of the project several services will be operational and adopted by partners. The costs of maintaining PaNOSC outputs (e.g. the federated data portal and search tool, the software catalogue, etc.) will be reviewed for the sustainability closer to the end of the project when there will be a clearer picture on the arrival point or eventual unforeseen additional costs for the adoption of the PaN FAIR and open data approach. These results will be published as well.