Computer Science Review 49 (2023) 100571

Contents lists available at ScienceDirect

Computer Science Review

journal homepage: www.elsevier.com/locate/cosrev

Review article

A survey of the European Open Science Cloud services for expanding the capacity and capabilities of multidisciplinary scientific applications

Amanda Calatrava [a,*], Hernán Asorey [k,l], Jan Astalos [f], Alberto Azevedo [c], Francesco Benincasa [i], Ignacio Blanquer [a], Martin Bobak [f], Francisco Brasileiro [b], Laia Codó [i], Laura del Cano [g], Borja Esteban [e], Meritxell Ferret [i], Josef Handl [h], Tobias Kerzenmacher [d], Valentin Kozlov [e], Aleš Křenek [h], Ricardo Martins [c], Manuel Pavesio [m], Antonio Juan Rubio-Montero [j], Juan Sánchez-Ferrero [m]

[a] *Instituto de Instrumentación para Imagen Molecular (I3M), Universitat Politècnica de València, Valencia, 46022, Spain*
[b] *Federal University of Campina Grande (UFCG), Campina Grande, Brazil*
[c] *Laboratório Nacional de Engenharia Civil (LNEC), Lisbon, Portugal*
[d] *Institute for Meteorology and Climate Research–Atmospheric Trace Gases and Remote Sensing (IMK-ASF), Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany*
[e] *Steinbuch Centre for Computing (SCC), Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany*
[f] *Institute of Informatics, Slovak Academy of Sciences, Bratislava, Slovakia*
[g] *Centro Nacional de Biotecnologia, CSIC, Madrid, Spain*
[h] *MU, Brno, Czech Republic*
[i] *Barcelona Supercomputing Center (BSC), Barcelona, Spain*
[j] *Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas (CIEMAT), Av. Complutense 40, Madrid, 28040, Madrid, Spain*
[k] *Medical Physics Department, Comisión Nacional de Energía Atómica (CNEA), Centro Atómico Bariloche, San Carlos de Bariloche, 8400, Río Negro, Argentina*
[l] *Instituto de Tecnología en Detección y Astropartículas (ITeDA, CNEA/CONICET/UNSAM), Centro Atómico Constituyentes, Villa Maipú, 1450, Buenos Aires, Argentina*
[m] *Control, Observation and tracking systems, Space Management Area (Indra Sistemas SA), Ctra. de Loeches, 9, Torrejón de Ardoz, 28850, Madrid, Spain*

ARTICLE INFO

Article history:
Received 4 March 2022
Received in revised form 11 May 2023
Accepted 18 May 2023
Available online xxxx

Keywords:
Open science
Cloud computing
Federated infrastructure
Multidisciplinary
EOSC

ABSTRACT

Open Science is a paradigm in which scientific data, procedures, tools and results are shared transparently and reused by society. The European Open Science Cloud (EOSC) initiative is an effort in Europe to provide an open, trusted, virtual and federated computing environment to execute scientific applications and store, share and reuse research data across borders and scientific disciplines. Additionally, scientific services are becoming increasingly data-intensive, not only in terms of computationally intensive tasks but also in terms of storage resources. To meet those resource demands, computing paradigms such as High-Performance Computing (HPC) and Cloud Computing are applied to e-science applications. However, adapting applications and services to these paradigms is a challenging task, commonly requiring a deep knowledge of the underlying technologies, which often constitutes a general barrier to its uptake by scientists. In this context, EOSC-Synergy, a collaborative project involving more than 20 institutions from eight European countries pooling their knowledge and experience to enhance EOSC's capabilities and capacities, aims to bring EOSC closer to the scientific communities. This article provides a summary analysis of the adaptations made in the ten thematic services of EOSC-Synergy to embrace this paradigm. These services are grouped into four categories: Earth Observation, Environment, Biomedicine, and Astrophysics. The analysis will lead to the identification of commonalities, best practices and common requirements, regardless of the thematic area of the service. Experience gained from the thematic services can be transferred to new services for the adoption of the EOSC ecosystem framework. The article made several recommendations for the integration of thematic services in the EOSC ecosystem regarding Authentication and Authorization (federated regional or thematic solutions based on EduGAIN mainly), FAIR data and metadata preservation solutions (both at cataloguing and data preservation—such as

* Corresponding author.
E-mail address: amcaar@i3m.upv.es (A. Calatrava).

https://doi.org/10.1016/j.cosrev.2023.100571

EUDAT's B2SHARE), cloud platform-agnostic resource management services (such as Infrastructure Manager) and workload management solutions.

Contents

1. Introduction

The e-Science paradigm studies, enacts, and improves the ongoing process of innovation in computationally-intensive or data-intensive research methods [1]; typically, this is carried out collaboratively, often using distributed infrastructures. Open Science [2] is the practice of science so that others can collaborate and contribute, with research data, lab notes and other research processes freely available, under terms that enable reuse, redistribution and reproduction of the research and its underlying data and methods.

Scientific applications place higher demands on computing power every year. The need for large-scale computing resources, including specific hardware needs such as GPUs, and the increasing demand for storage resources due to the large amount of data generated by these types of applications are challenging for researchers and computer scientists. e-Science uses e-Infrastructures [3], which are collaborative virtual environments that provide digital services and tools to meet this resource demand.

E-Infrastructures are based on distributed backends ranging from High-Performance Computing to Cloud Computing. However, adapting existing software applications to these paradigms is not trivial [4]. This process commonly requires an in-depth knowledge of the underlying technologies to take advantage of their benefits, as it usually requires refactoring the application's architecture. This adaption becomes even more challenging in the Cloud computing paradigm due to the complexity of virtualization and elasticity, the vast range of services available and the variety of resource types. This fact can be a barrier for scientists and researchers, as it requires non-trivial ICT skills.

Another obstacle in adopting the e-Science and Open Science paradigms is the fulfilment of the FAIR [5] (Findability, Accessibility, Interoperability and Reusability) principles, which imply permanent and discoverable identifiers for fully annotated data and metadata. Specifically, in Europe, we can find the European Open Science Cloud (EOSC) [6], a European initiative co-funded by the European Commission that aims to facilitate the deployment and consolidation of an open, trusted, virtual, federated environment in Europe to store, share and re-use research data across borders and scientific disciplines promoting open science practices and providing access to a rich array of related services.

As part of this initiative, the EOSC-Synergy [7] project aims to increase the uptake of EOSC through the capacity and capability building using the experience, efforts and resources of national publicly funded digital infrastructures. The project has identified ten Thematic Services (TSs) in four scientific domains (Earth Observation, Environment, Bio-medicine, and Astrophysics). These thematic services are heterogeneous and cover a wide range of requirements, maturity levels, user targets and usage models. They will increase their functionality by integrating services from the EOSC marketplace [8]. This marketplace comprises a collection of services tackling e-Science application challenges, such as the Discovery of research outputs, the processing and management of scientific data and metadata, the access to research infrastructures and the publication of scientific results.

The ten thematic services also provide helpful best practices for future new services to be developed or adapted to this environment, as they address challenges on federated Authentication and Authorization Infrastructures (AAI), elastic data processing, interoperability with data infrastructures, metadata management and accounting, that apply to many other applications in same or different scientific domains. The analysis of these thematic

services' limitations and the design of the new architecture of the thematic services will be of particular interest to new scientific applications that want to embrace this paradigm.

In this work, we present each thematic service of EOSC-Synergy and then analyse the gaps and bottlenecks in authentication and authorization services, resource provisioning, workload management, and data storage. This section is followed by an analysis of the different tools and services provided in the EOSC Marketplace that can meet the needs of the thematic services. We then present the adoption of services, tools and technologies used by the thematic services to address their needs and then model a generic scientific application that can be the starting point for new scientific thematic services. Next, to better illustrate the work done in adopting tools and services by the thematic services, we discuss the adoption issues in the service instantiation section. Finally, we analyse the state of the art to identify the most critical work related to implementing thematic services in the e-Science paradigm and conclude the paper with the main observations.

2. Thematic services in EOSC synergy

In this section, we briefly describe each of the ten thematic services of EOSC-Synergy, grouped by scientific discipline. Moreover, the expected outcome of the integration with EOSC-Synergy for each one of the thematic services is pointed out in this section.

2.1. Thematic services in Earth Observation

In Earth Observation (EO), three thematic services analyse large satellite imagery, from monitoring coastal changes and inundations to estimating forest masses and crops. They are addressing different types of targets. Specifically, the three services are:

- WORSICA (Water Monitoring Sentinel Cloud Platform) [9,10]: A service for detecting water using satellites, Unmanned Aerial Vehicles & in-situ data. WORSICA can be used for coastline detection, inland water bodies detection and water leak detection on irrigation networks. WORSICA aims to integrate multiple-source remote sensing and in-situ data to determine the presence of water in coastal and inland areas. WORSICA enables the research communities to generate maps of water presence and delimitation lines in coastal and inland regions. These products can be helpful in emergency and planning methodologies in case of inundations or reservoir leaks. In the frame of EOSC, the service will be scaled up to a European level to reach all interested research communities.
- SAPS (Surface Energy Balance Automated Processing Service) [11]: SAPS is used to estimate Evapotranspiration and other environmental data that can be applied, for example, to water management and the analysis of the evolution of forest masses and crops. SAPS allows the integration of Energy Balance algorithms to compute the estimations of particular interest to Agriculture Engineering and Environment researchers. These algorithms can be used to increase the knowledge of the impact of human and environmental actions on vegetation, leading to better forest management and analysis of risks. SAPS is being developed in Brazil, but with the adoption of EOSC services, it is expected to facilitate European scientists to exploit evapotranspiration estimation services from remote sensing imagery.
- G-Core (Acquisition, cataloguing and processing EOS data), [12,13]: G-Core is a production-ready technology used as a service at ESA's and national programs that provides a Data Manager for spatial and non-spatial purposes and a framework for third-party processors. G-Core is a service for acquiring, storing, cataloguing and processing data from several Earth Observing System (EOS) missions. Its two main functionalities are: (i) a Data Manager for spatial and non-spatial purposes; and (ii) a Processing framework to host external processors developed by third parties to generate added-value products based on Satellite imagery. The main goal of its integration in the EOSC ecosystem is to offer the service as a Payload Data Ground Segment (PDGS) in the cloud for future ground segment space missions or as a processing framework to plug in different processors that can make use of the Copernicus resources or private data in order to produce different levels of products to be delivered to the users. Thus, the expected impact of the service adaptation is to democratize the usage of EO data out of the scope of nominal fields. It will help define new products and services, mixing Earth Observation data with other data types for scientific and social environments.

2.2. Thematic services in biomedicine

In this area, the thematic services cover the benchmarking of Genomic data processing tools and the processing of Cryo-electron microscopy imaging. The services are:

- SCIPION (CryoEM data processing for Structural Biology [14]): Cryo-Electron Microscopy Service is an image processing framework used to obtain 3D maps of macromolecular complexes using cryo Electron Microscopy. It has been developed as a plugin-based workflow management system that integrates many necessary software packages in the field. Integrating Scipion with Cloud services allows users from the Instruct Research Infrastructure to deploy a dynamic cluster in the cloud to keep processing the data acquired at an Electron Microscopy facility. This cluster has all cryo-em packages and software needed to obtain a 3D structure and is powered by EOSC compute resources on the backend. It means that scientists with minimal computational background (or compute resources of their own) can access the latest tools and powerful computational resources to obtain a refined 3D structure to be published and shared with the community.
- OpenEBench [15] (ELIXIR [16] OpenEBench is used to evaluate Life Sciences research software. OpenEBench is an observatory for software quality based on the automated monitoring of FAIR for research software metrics and indicators. The OpenEBench platform supports the technical monitoring of scientific software and scientific benchmarking activities carried out by Life Sciences Communities. Its architecture has three engagement levels that allow communities at different maturity stages to use the platform. It also connects with ELIXIR Core Data Resources and Deposition databases to use data needed by the Scientific Communities activities. The expected impact of integrating EOSC services is that Life Science researchers will have semantically annotated, up-to-date collections of benchmarked analytical workflows and tools organized by scientific communities for specific topics, which can be deployed across heterogeneous systems.

2.3. Thematic services in Astrophysics

In Astrophysics, the LAGO thematic service sets up a European service for the Latin American Giant Observatory.

- LAGO, the Latin American Giant Observatory [17,18], is an extended cosmic ray observatory with a vast network of water Cherenkov detectors currently deployed at ten countries in Latin America, from the south of Mexico to the Antarctic Peninsula. The geographic distribution of LAGO allows the realization of diverse astrophysics studies at a regional scale. LAGO is mainly oriented to perform basic research focusing on three main areas: high energy phenomena, atmospheric radiation measurement at ground level, space weather, and climate monitoring. All the LAGO analyses are supported by data-intensive computational frameworks that integrate different simulations tools with their own designed data-analysis codes to determine, exactly, the signals measured or expected at any detector of any type in any particular site around the world and under realistic atmospheric and geomagnetic time-evolving conditions. The final purpose of the LAGO Thematic Service [19] is to enable the universal profit and contribution of this research, within and outside LAGO Collaboration, through a sustainable Virtual Observatory and standardized computational model.

2.4. Thematic services in the Environmental domain

Finally, in the area of Environment, the fourth group of thematic services include sand and dust storm forecasting, untargeted mass-spectrometry (MS) analysis for toxins, water network distribution simulation and the monitoring of stratospheric ozone in climate models. Specifically, these services are as follows:

- SDS-WAS (A Service related to the mineral dust forecast) [20,21]: SDS-WAS is a service that aims at improving capabilities for more reliable sand and dust storm (SDS) forecasts. It supports institutional entities to warn about possible dust events and to foster the study of dust-related phenomena. The framework collects numerical model outputs and observational data from a broad set of worldwide partners plus internally developed. A comprehensive set of post-processed analyses and statistics are generated, and results in plots, tables or numerical (binary) data are disseminated to various users (e.g. public institutions and researchers). Integrating the framework in EOSC will increase the volume of data hosted and processed to reach a wider set of end-users, improve compliance with data FAIR principles and reinforce the robustness of the whole service infrastructure.
- UMSA (Untargeted Mass-Spectrometry Analysis) [22]: UMSA aims at processing mass spectrometry data to correlate the whole spectra (i.e. all the present compounds) with other data (social, medical, other sample analyses, etc.) to work with more complex hypotheses on the impact of the environment in human health. The data are unrecoverable. Therefore long-term data storage is required, together with appropriate data curation. Using the integration in EOSC, uniform access to data and computing resources is provided, scaling the service to the target European-wide user community.
- MSWSS [23] (Water Supply Systems modelling and analysis): MSWSS is a service for modelling and analysing Water Supply Systems which integrates the analysis of toxins in drinking-water supply networks with water distribution network simulation. It allows water infrastructure operators and researchers to analyse hazardous events (e.g. toxin propagation within a pipe system). It may be used to prepare risk management plans for water utilities. The integration with EOSC computing infrastructure and data-sharing services will enable the modelling of more complex water supply systems and increase the number of scenarios for the analysis.
- O3as (Ozone Assessment Service) [24]: The O3as service provides an invaluable tool to extract ozone (O_3) trends from extensive climate prediction model data to produce figures of stratospheric ozone and figures of dates by when depleted O_3 recovers to pre-ozone hole levels. This service is conceived to assist scientists in visualizing ozone data from large climate models by calculating dates for the ozone layer's recovery and providing trends of the ozone abundance in the atmosphere to produce results in the form of figures in publication quality. The integration of the service in EOSC has increased its capacity to process large volumes of data (in terms of TBs) and to facilitate the management of the complex workflow to generate key metrics. Climate model data has many fields of physical quantities. The relevant quantities for the ozone column calculation must be extracted and processed to be visualized efficiently. A preprocessor runs on an HPC system, reducing the large dataset so that a REST API can read it to produce figures without noticeable delays.

3. Gaps and bottlenecks analysis

Before integrating the thematic services in EOSC, we performed a deep analysis of each one to properly identify the needs and requirements in each specific area and use case. The aim is to increase the services' capacity, performance, reliability and/or functionality according to their needs. Thus, every thematic service has analysed the technical services and solutions they used to manage users, computing and data. This reflection has led to a list of limitations and needs considered as input to design the strategy to implement the improvements in each case. Table 1 shows a summary of the limitations, lacks and needs identified by the Thematic services of the project. We have tagged each identified need in a category to identify and classify them easily. These categories are: AAI, computing resources, workload management, storage and network.

In a preliminary analysis of the results, several technical commonalities and differences have been identified. Firstly, storage is a category that appears in all the thematic services, no matter the domain. All the services share limitations in data management, converting it into the most challenging part. Thematic services have identified important issues in transferring and accessing large volumes of data and require smart caching, advanced data transferring and massive persistent data storage. At the time of this study, there were two main approaches to cope with storage, from deploying their own datastore, e.g. a DATAVERSE [25] instance; to the integration of external Data Infrastructures, like EGI DataHub [26] and EUDAT [27]. Moreover, services must ensure compliance with the FAIR principles to facilitate the access, cataloguing and reuse of the data generated by their services. Thus, a storage service able to manage, together with the data, metadata and unique data identifiers will be required.

Secondly, the need of accessing to a wide backend of computing resources is also a requirement for all the domains. This need is usually related to the workload management category, as the resources need to be properly managed to take advantage of them, both at the level of virtual machines/containers (resource management) and at the level of tasks/jobs to schedule the workflow of the applications (workload management). At the resource management level, some services have expressed an interest in

Table 1
Analysis of the limitations and needs of the Thematic Services (TSs) grouped by area (in green, the TSs in earth observation; in yellow, TSs in biomedicine, in purple the TS in astrophysics; and in blue TSs in the environmental domain).

Thematic service	Limitations and needs	Category
WORSICA	– Improve download speed and num. concurrent downloads of satellite images.	Network
	– Increase storage of the images needed for the algorithm.	Storage
	– Increase computational resources to speedup the image processing.	Comp. resources
	– Seamless authentication and authorization for end users.	AAI
SAPS	– Need for a larger-scale deployment: computing, storage and data access.	Comp. resources/Storage
	– Scalability and standardization of services	Workload Mng.
	– Integrated and widely supported AAI	AAI
GCore	– Overcome limited access to data repository: network bandwidth restrictions.	Network
	– Infrastructure resources for processing and reprocessing large datasets.	Comp. resources/Workload Mng.
	– Data delivery volume. Increasing size of files to be delivered to users.	Storage
SCIPION	– Insufficient Cloud resources for the workflow: GPUs, CPUs and RAM.	Comp. resources
	– Need of a Resource Management able to optimize the use of cloud resources.	Workload Mng.
	– Storage limitations and data transfer performance: 1-3 TB raw data.	Storage
	– Distributed and shared file system.	Storage
OpenEBench	– Need to work on heterogeneous systems to reach Life Sciences Communities	Workload Mng.
	– Need to efficiently store processed data and workflows in a FAIR manner.	Storage
LAGO	– Limitations on data preprocessing.	Comp. resources/Workload Mng.
	– Needs data storage that copes with FAIR, curation and harvesting.	Storage
	– Need for computing power for simulations, together with optimal scheduling.	Comp. resources/Workload Mng.
SDS-WAS	– Lack of services needed for Data storage and curation.	Storage
	– Lack of computing power for data analysis on-demand.	Comp. resources
UMSA	– Long-term data storage is required, together with appropriate data curation.	Storage
	– Execution by means of efficient workloads supporting data provenance.	Workload Mng.
MSWSS	– Access policies to protect data access.	AAI
	– The data has to be stored in a private storage only.	Storage
	– Need to provision computing resources on demand	Comp. Resources
O3as	– Requires larger storage resources, specially improving data availability.	Storage
	– Fast handling of big data.	Workload Mng.
	– Need to provision computing resources on demand.	Comp. Resources

dynamically provisioning processing resources, most of the cases on demand (like MSWSS and O3as). However, each thematic service has different needs: from a dynamic, dedicated cloud backend to an elastic cluster that shrinks or grows according to the workload, and even need to access external High-Performance Computing (HPC) and High Throughput Computing resources for massive Batch jobs execution. Regarding workload management, most thematic services were using batch queues (like SLURM [28] Batch queues or Galaxy [29]), which could be extended to support containerized jobs. The usage of Kubernetes to orchestrate microservices and job queues of containers is envisioned in most cases, as we will present later in Section 5.

Thirdly, regarding AAI, only three thematic services (from two different domains) have raised the issue regarding access policies and user management. It is due to the different maturity levels of each thematic service at the moment of this study. Some of them were still under development and had yet to be considered. However, as we will discuss later in Section 5, all thematic services finally have shared the importance of using a robust AAI compatible with the ones used by their target institutions and user communities.

Lastly, there is a category that only arises in a specific domain (earth observation): network-specific needs. This comes from the size of sentinel data images, which are usually huge, and the transmission of this kind of data can represent a bottleneck for the applications in the field.

In summary, only some services have identified needs in all the categories, but most are domain independent. Thus, each thematic service will focus its adaptation to the cloud and the EOSC ecosystem on the more relevant aspects for them, according to their bottlenecks. The next step is to deeply analyse the solutions available in the EOSC marketplace and the open science environment before adapting them into the thematic services.

4. Analysis of the EOSC Portal Catalogue and Marketplace

In this section, we aim to identify the key EOSC tools and services that can address the issues and needs analysed above. Considering the gaps and bottlenecks identified, this analysis also considers potential alternatives to the current technical services used by the thematic services to overcome such issues.

The EOSC Portal Catalog & Marketplace [8] has been developed from users' perspective, identifying the needs to be supported and facilitating all the actors involved in implementing an open approach to science sustainably. The catalogue has more than 420 entries registered by the first quarter of 2023, covering resources from several categories. According to a functional perspective, we can organize them into six categories:

- Access physical & eInfrastructures: offering generalist resources like virtual machines and containers and storage and network transport connectivity. By the beginning of 2023, 113 resources will be listed under this topic. This category includes Compute resource providers, workload managers, Resource orchestrators and data providers. Some of the services were thematic (e.g. discipline-specific). We identified six generic services that could address the requirements of the thematic services: B2SAFE for long-term data preservation, EGI Cloud Compute to provide IaaS cloud resources, EGI DataHub to provide online cloud storage resources, EGI High-Throughput Compute for batch workloads, EGI Workload Manager to orchestrate multi-site batch resources, and Infrastructure Manager (IM) to deploy virtual infrastructures on top of cloud offerings), according to the following criteria: Generic purpose, interoperability and support. A brief description of the services is provided next:
  - EGI Cloud Compute [30], an IaaS from the EGI Federated Cloud that enables users to deploy and scale virtual machines on demand.
  - EGI HTC Compute [31] enables running computational jobs at scale on the EGI infrastructure, which is provided by a distributed network of computing centres and offers more than 1,000,000 cores of installed capacity, supporting over 1.6 million computing jobs per day.

  - EGI Workload Manager [32], a service to manage and distribute your computing tasks efficiently while maximizing the usage of computational resources.
  - EGI DataHub [33], a service that brings data close to the computing to exploit it efficiently and can be used to publish a dataset and make it available to a specific community or worldwide across federated sites.
  - B2SAFE [34] is a service for the long-term preservation of the EUDAT Data Collaborative Infrastructure, one of the largest e-infrastructures in Europe offering permanent storage capacity and integrated management services for research communities. EUDAT also provides other services such as B2SHARE, B2FIND, B2ACCESS.
  - Infrastructure Manager (IM) [35]

- Aggregators & Integrators, where we can find several tools and utilities to facilitate accessing services and resources by indexing and annotation. Out of the 40 resources available in this category, we identified the Dynamic DNS service [36] to easily add a DNS name to an instance deployed in the virtual infrastructure of EOSC, the EGI Fedcloud client [37], that facilitates the access to the federated cloud computing platform, and B2FIND [38], another EUDAT's service, to annotate research objects, considering the requirements of the thematic services and similar criteria as in the first item.
- Processing & Analysis mainly aimed at facilitating the management of computational resources and scheduling the execution of workloads. Despite this category accumulating the highest number of services, most of them are discipline-specific. Moreover, some of the resources were already listed under the first item. Here we can find Elastic Compute Clusters in the Cloud (EC3) [39], a tool to deploy virtual elastic clusters on top of IaaS clouds, and B2Handle (another EUDAT's service) to provide persistent identifiers to resources.
- Security & Operations aims at guaranteeing that the overall system and the services operate securely and according to standard. In this case, thematic instances for authentication and authorization may be preferred as community researchers already have acquired credentials. For this purpose, we identify the EGI Check-In [40] service, the B2ACCESS [41] and EduTeams [42], along with ELIXIR AAI [43] which is not listed in the catalogue.
- Sharing & Discovery relates both to services that produce data relevant to specific disciplines and horizontal services for data deposit and annotation. Only the service B2SHARE to enable sharing and publishing of research data is considered. The catalogue also includes an instance of Dataverse, although we have decided to deploy our own instance. The Dataverse Project, developed by Harvard's Institute for Quantitative Social Science (IQSS) and many collaborators and contributors worldwide, is an open-source web application to share, preserve, cite, explore, and analyse research data.
- Training & Support aims at facilitating the access to high-quality technical information and tailored training materials. Services in this category are not considered.

All these services have two different access modalities:

- Direct access. This model is used by services that are instantiated up-front and do not require intensive access to resources or are provided directly by the user (e.g. EGI Check-in or Infrastructure Manager). Users are automatically forwarded to the service endpoints.
- Access through orders. This model is used in services that require a non-trivial amount of resources (e.g. EGI Cloud Compute or B2SAFE). In this case, the user usually has to choose between different offerings, which may result in costs.

5. Adoption of EOSC services

As shown in Section 3, the ten thematic services have common and complementary requirements, needs and features. In general, all of them are shared between the four scientific domains. Following our preliminary categorization in the gaps and bottlenecks study, we have defined four action areas, which are analysed below.

Authentication and Authorization Infrastructure (AAI). All services require users to be authenticated and authorized. In some cases, there is a need for delegation from the users that access the platform for accessing data or processing resources. In those cases, it is mandatory to have a coherent single-sign-on mechanism. Other cases may require an AAI linked to popular scientific IdPs and implement the authentication via Virtual Organization membership. From the tools and services we identified from the TSs, EGI Check-in has been revealed to be a widely accepted choice. Another option analysed is B2Access, mainly for interacting with the infrastructure. There are also a few cases in which users will use federated credentials to access the services—mainly related to storage.

Workload Management. Most of the cases deal with the execution of a set of batch jobs. In those cases, workload managers should be integrated to take advantage of the computing resources better. This will provide the capability to deal with a larger capacity and workloads. The options here range from using a standard batch queue (SLURM) that can be eventually powered up with automatic elasticity to using Kubernetes to orchestrate a container-oriented approach.

Resource Management. Most thematic services require deploying a virtual infrastructure where the services that provide the functionality and the processing will occur. Most of the thematic services have identified the use of Infrastructure Manager (IM) or Elastic Compute Clusters in the Cloud (EC3) client as a technology capable of filling in this gap. Both tools could provide the capability of defining a virtual infrastructure as code and deploying it on the cloud. IM (for static infrastructures) or EC3 (for dynamic infrastructures), together with recipes for K8s, Slurm or Galaxy clusters on top of a dynamic dedicated cloud backend, is the preferred solution. Moreover, some thematic services require links to external HPC resources (like Marenostrum in BSC) and HTC resources (EGI HTC compute) for the execution of massive Batch jobs.

Data Storage. The services need storage connected to the processing that can be efficiently accessed. In this case, there is a wide range of different solutions proposed or implemented in the thematic services, ranging from external solutions like EGI-DataHub, B2SAFE and B2SHARE to local solutions based on Nextcloud, Dataverse, Elasticsearch and WebDav, where typically the resource manager will also be in charge of deploying and configuring their own Datastore instance (e.g. DATAVERSE instance).

Table 2 summarizes the tools and services performed by each thematic service for the fourth category described above. In summary, three different (although compatible among them) AAI methods have been integrated (EGI Checkin, B2ACCESS, Life-Sciences AAI and eduTEAMS). Job scheduling ranges from solutions based on containers (using Kubernetes) to solutions using batch queues (mainly based on SLURM), supported in some cases by workflow frameworks such as Galaxy and instantiated through EC3. For the interaction with cloud resources, TOSCA [44] and RADL recipes have been developed for Infrastructure Manager. Finally, data access is performed through different solutions such as Dataverse, EGI DataHub OneData, B2SHARE and B2SAFE, which clearly states the complexity of the data management issue and the wide range of solutions.

Table 2
Adoption of technologies for each Thematic Service.

Service	AAI	Workload Mng.	Resource Mng.	Data storage
WORSICA	EGI Check in	ArcCE, Batch (SLURM)	IM (TOSCA)	Nextcloud, Dataverse
SAPS	EGI Check in	K8s	IM/EC3	OpenStack Swift
G-Core	CAS User/pwd & EGI Check in	GCore + K8s	IM/EC3	ElasticSearch
Scipion	EGI Check in	Batch (SLURM)	IM/EC3	Local + EGI DataHub
OpenEBench	Life Sciences AAI	WfExS + NextFlow	OpenNebula	Local + B2SHARE
LAGO	eduTEAMS + EGI Check-in	Batch (SLURM)	Local clusters + IM/EC3	EGI DataHub ONEDATA
SDS-WAS	B2ACCESS	Batch (SLURM)	Local clusters	B2HANDLE/B2SAFE
UMSA	EGI Check in & Life-science AAI	Batch (SLURM) in IM/EC3 (in Galaxy)	IM/EC3	Local + S3
MSWSS	EGI Check in	Batch (SLURM) in EC3 (in Galaxy)	IM/EC3	Local + Dataverse
O3as	EGI Check in	Batch (SLURM) & K8s	Local cluster + IM	Local + WebDAV

Notice that we did not explicitly include "Computing resources" or "Network" because both requirements are solved by using the EGI Cloud Compute and the EGI HTC Compute platforms, which all the thematic services have access to because of the EOSC-Synergy project context. This is not limited to the project context. For other use cases, these computing platforms can be open and freely accessed by requesting access through the EOSC Portal Catalog & Marketplace.

To sum up this section, with the adoption of all the services and technologies depicted in 2, thematic services have experienced all these improvements:

- Integration of standardized AAI IdPs to facilitate user management.
- Improvement of processing backends by replacing single computing instances with batch job queues, container management platforms or clients to high-throughput computing backends.
- Publishing the output results in persistent repositories.
- Improving repeatability and platform-agnosticism by describing the application topologies as code using standard TOSCA language.
- Self-management of resources to reduce maintenance costs.
- Persistent Identifiers (PID) annotation of output data and integration in official harvesters.

Thanks to the rich analysis of the experience of these ten thematic services in the adoption of several tools, services and technologies to improve and solve their needs, the path to follow for a new scientific use case is far more accessible. However, to clarify this process and to quickly identify the key services and technologies selected, we present a generic application integrated with the EOSC ecosystem in the next section.

6. Application modelling

This section uses as input the experience of the ten thematic services of EOSC-Synergy to define a canonical generic application architecture leveraging the services identified in the EOSC Marketplace catalogue in Section 4. Thematic services that have similar requirements as those described in Section 3 can use as the basis for this architecture that relies on several tools and services from the EOSC ecosystem, together with well-known frameworks and technologies of the cloud computing paradigm, all of them carefully selected taking into account the selection made by the thematic services.

First of all, after analysing our ten use cases, we identified two different deployment scenarios: (i) a single instance of the service shared by several users or communities, offered as a web portal able to manage users, data access, processing and visualization, supported by a shared or dedicated pool of resources (e.g. WORSICA or OpenEBench); and (ii) an instance of the service deployed on demand, where each user deploys his/her own instance of the service on Cloud resources based on a combination of TOSCA recipes with Docker containers (e.g. SAPS, SCIPION).

Regardless of the approach chosen, Fig. 1 shows the architecture of this generic application that relies on its deployment on top of the EGI Cloud Compute platform. The first layer considered essential by the thematic services is the authentication and authorization infrastructure. For that, EOSC offers the EGI Check-In service, the most popular solution adopted by the thematic services. This service can be easily integrated with a service exposed to users through a web portal. It will be used by both the application manager and the end-users to access the EOSC resources and the service itself. Once the user has been properly authenticated, and depending on the usage model that the service wants to use, he/she will have access to the Application User Interface of the scientific application itself or to the portal of the resource manager that will facilitate the deployment of the scientific application instance. In the second scenario, the user will be redirected to the portal of IM or EC3 to deploy a virtual cluster configured on demand for his own usage. The selection between IM and EC3 has to be taken depending on the needs regarding resource consumption. If elasticity is required, the tool to be used will be EC3. Otherwise, IM is the best tool to provide a static infrastructure configured on demand. Both solutions will require the preparation of a recipe where the application manager specifies the required steps and commands to properly install, configure and deploy the scientific service, together with the credentials to access the cloud provider.

In order to take advantage of the virtual infrastructure where the scientific application is running, we need to rely on a workload manager. From our analysis, we have identified two approaches: (i) a traditional batch job queue managed by the well-known SLURM scheduler, or a solution based on the containerization of jobs, where Kubernetes has proven to be the most popular scheduler. Both options are feasible, depending on the approach that the scientific service wants to follow. However, adopting one of these workload managers might require an effort to adapt the tool's architecture, so this duty must be consciously analysed.

Finally, for the data storage, we recommend using a solution that supports metadata to comply with FAIR principles, i.e. to make data Findable, Accessible, Interoperable, and Reusable (like Dataverse or B2SHARE). No matter if the storage solution will work locally or it will be an external service, one of the most important aspects is the support to metadata to properly index and facilitate reusing the data generated by the services, especially if this will be of interest to other researchers of the area.

7. Service instantiation

In this section, we want to exemplify how adopting the new EOSC tools and services can address the gaps and bottlenecks detected by some of the project's thematic services. Specifically, we present seven examples from seven different scientific services showing the integration in each one of the categories described

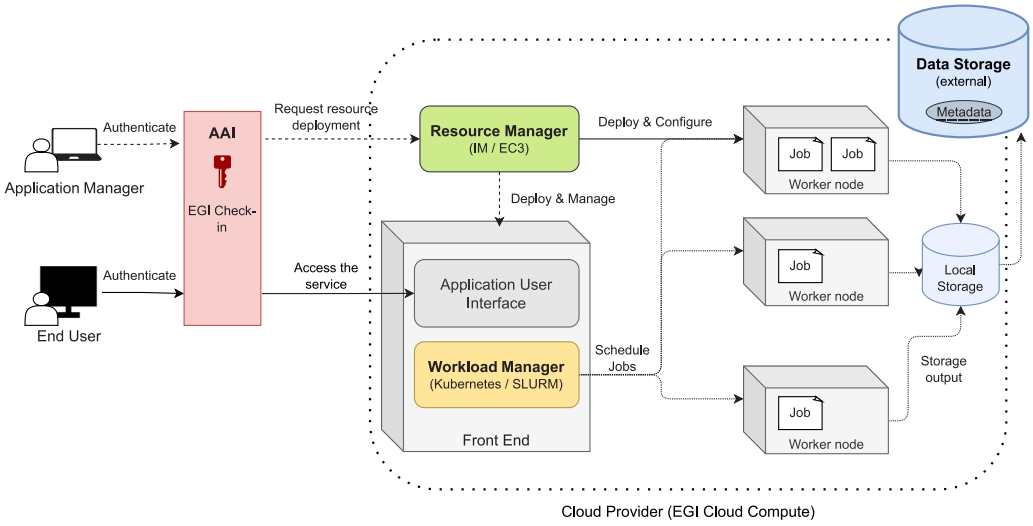

Fig. 1. Architecture of the proposed solution.

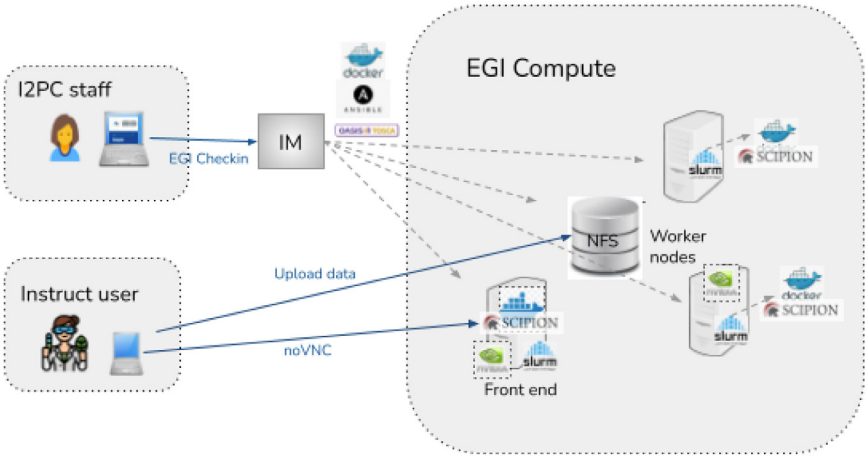

Fig. 2. Scipion service usage and architecture.

in Section 5. We have omitted the integration with the authentication and authorization service of EOSC (EGI Check-in) because this is a well-known process that is properly documented [45]. The following subsections cover the cases for the rest of the categories.

7.1. Workload management

During the analysis of the thematic services, we have detected two main needs for workload management: (i) services that use a more traditional approach relying on local resource management systems based on queues of jobs; and (ii) services that encapsulate the tasks in a container and rely on a container-based management system. To exemplify these two models, we analyse both Scipion and O3As.

7.1.1. Batch job oriented

The Scipion service aims to facilitate the life of users from the Instruct-ERIC [46] Research community for processing their electron microscope data. Users who obtained her data through an Instruct granted project in an Electron Microscopy facility can request the Scipion service by contacting the Instruct Image Processing Center (I2PC). Then, the request is reviewed, and if the available quota permits it, the I2PC administrator will deploy a cluster using the Infrastructure Manager (IM). The user will then receive an email with instructions on accessing the front-end node and copying the data to start processing. She can use the service for a maximum of one month, although if no other service requests are pending, this time might be extended. The cluster will be destroyed by the I2PC team after the granted period finishes, giving the user enough time to download her results.

As shown in Fig. 2 the cluster is defined in a TOSCA description that includes Ansible recipes and Docker images to launch and configure the different nodes in the cluster. This TOSCA description is deployed through the IM. Once the cluster is running, the user can transfer the data to the front-end node and access the service using either a VNC client or NOVNC through a web browser.

The front-end host node runs a SLURM master and a Docker container that includes Scipion and related Cryoem packages configured to launch their jobs through SLURM. Once a job is sent to the queue, a Docker container is run in one of the worker nodes to run the Scipion command.

Hardware resources deployed in the cluster are part of the EOSC EGI Cloud Compute service that controls access through Virtual Organizations (VO). In the case of the Scipion service, deployment is only granted to members of the cryoem.instruct-eric.eu VO. The cluster-shared storage is currently based on a local Ceph disk.

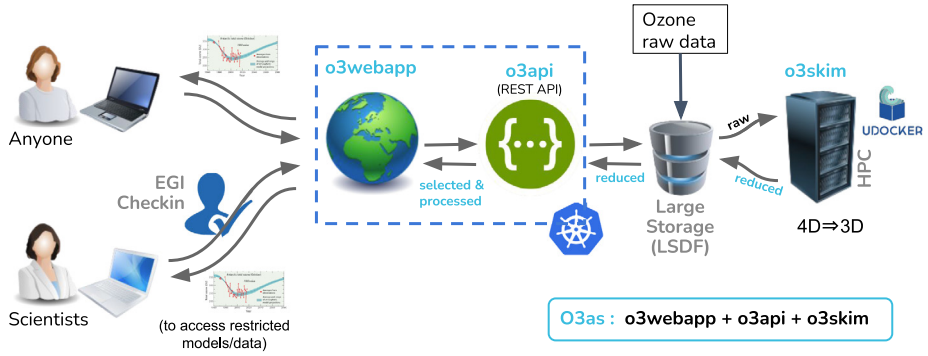

Fig. 3. O3as service overview: it consists of three main components: o3webapp (*to come*), o3api, and o3skim.

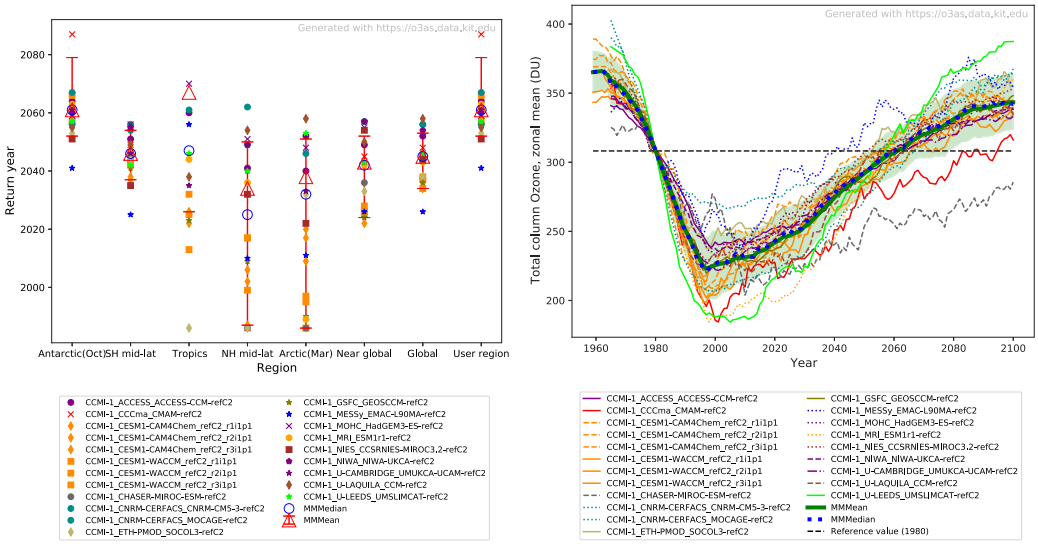

Fig. 4. Left: O_3 return dates for a recovery to ozone values in 1980. Right: Timeseries of total O_3 column data showing a decline of O_3 and the subsequent recovery of O_3 in the future.

7.1.2. Container based

O3as (using Kubernetes)

O3as service is composed of three main components (Fig. 3): o3skim to reduce the original data to the parameters of interest, o3api to provide API-based access to the skimmed data, and o3webapp for a user-friendly web interface (in development). O3skim runs on the local HPC system, while o3api and o3webapp are deployed in the Kubernetes system for container scalability, ensuring a fast enough response to user requests. If no Kubernetes system is available for the service providers, it can easily be instantiated by Infrastructure Manager (IM). Then the following steps are applied:

1. Install and configure cluster_issuer to handle Certificates for secure HTTP connections, e.g. from LetsEncrypt.
2. Initialize Ingress resource to route external traffic.
3. Deploy the o3api component as a container with the pre-configured PriorityClass.
4. Add Horizontal Pod Autoscaler (HPA) to respond with more containers on higher than usual loads.
5. Finally, instantiate o3api as a service.

If an existing Kubernetes cluster is used instead, steps 1–5 must be adjusted accordingly.

The output of the services is a set of projections of the ozone distribution. Example figures for climate models with projections for ozone until the year 2100 are shown in Fig. 4.

7.2. Resource management

Regarding resource management in cloud computing providers, we have identified two different approaches: (i) static infrastructures, where the number of nodes that compose the platform remains constant during the lifetime of the service, and (ii) dynamic infrastructures, where elasticity models are applied to the platform to adapt its size to the workload, thus allowing to reduce costs and wasting resources. As in the previous section, we have chosen two thematic services to exemplify adopting these two different solutions.

7.2.1. On-demand fixed infrastructures

Flexibility in choosing the computing platform was one of the objectives of LAGO TS. Beyond the elasticity or automatic management, the priority is providing resources that accomplish the needs of every specific calculation. So these requirements may be as variable as their parametrization. Some of these simulations may face intensive requirements, such as scratching up to several TBs of data; accessing many files through the Internet; continuously processing data in batch mode; or even sporadic calculations for demonstration purposes and scholars. Public clouds such as EOSC EGI Cloud Compute can tackle many of these tasks but require an upfront reservation of the resources by demand and fixing their environment. To face all these different approaches, three different services were integrated: the Infrastructure Manager (IM) service, the software encapsulation in standardized

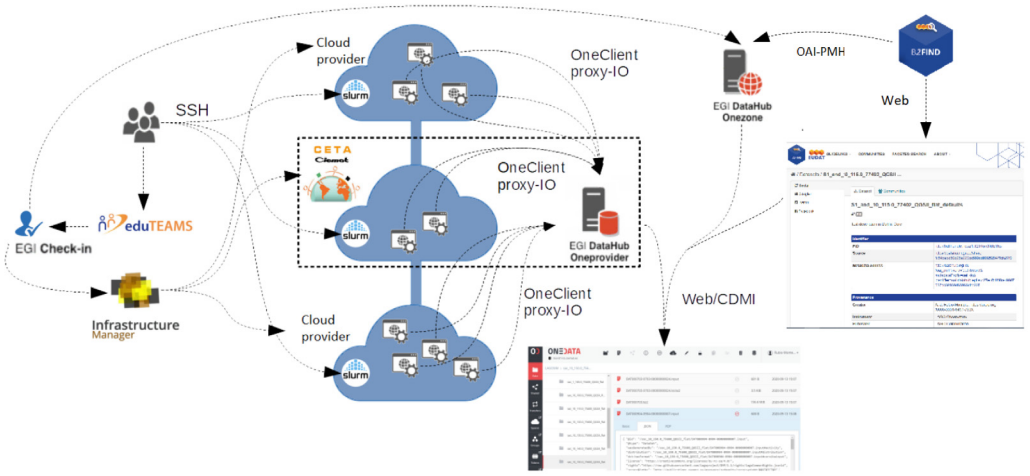

Fig. 5. On-demand deployment of fixed infrastructures and running LAGO software.

Docker images and the profiting from the ubiquity of the OneData cloud storage (EGI DataHub service). These technologies allow dynamically instantiating the virtual infrastructures needed, which are maintained fixed over days or weeks [19].

The resulting architecture is shown in Fig. 5. Researchers dynamically request virtual machines or batch clusters through the IM service to fix a temporal infrastructure on public clouds. Users can create any kind of cluster following not only their preferences, even the needs of specific calculations. In this sense, SLURM was a good choice as many LAGO collaborators commonly use it, and the behaviour could be similar to those used by other TS, like, e.g., Scipion, as it is described in Section 7.1.1. However, some scientists prefer to manage other implementations, such as Kubernetes, which is more suitable for Docker instances. Additionally, when tasks require scratching at the order of TBs, the user would not be allowed to spend the space accumulated by several computing nodes. In these cases, single virtual machines are the most suitable choice. IM deploys all these infrastructures after a few clicks on its website.

On the other hand, researchers can arbitrarily perform calculations by running the official LAGO docker images stored at DockerHub, which are periodically released by the CD/CI pipeline built on the JePL service [47]. Thanks to the virtualised approach, the software encapsulated in these images can be run on any platform supporting Docker. However, as the FAIR paradigm has to be fulfilled, all the LAGO software is always bound to the OneData cloud storage (DataHub), and it complies with the AAI procedure for the LAGO VO in every run. Thus, independently of the computing platform, PIDs always identify results, and they are browsable at the DataHub portal and findable by public harvesters such as B2FIND. Therefore, the on-demand provisioning of adaptable infrastructures supporting Docker through the IM service, jointly with the cloud storage via DataHub, allows users to accomplish their research without depending on other services.

For example, we deployed a SLURM virtual cluster counting on ten nodes with 16 Intel Xeon E7 cores and 250 GB of shared memory and disk. Then, we simulated the expected flux of the atmospheric radiation during the interaction of cosmic rays with the Earth's atmosphere for every LAGO detector [48]. We computed from 1 to 7 days of the expected flux at high altitude or Antarctic sites, reaching up to 1 year of the energetic flux that is accounted for volcanic risk studies. These new integration times are an enormous statistical improvement compared to previous results obtained in LAGO. Note that, e.g., a 24-h flux in one of the high latitude (Antarctic) detectors involves the simulation of $\sim 1.9 \times 10^9$ different cascades. Thus, we simulated the impressive figure of $>10^{12}$ particles spending >300 kCPU-hours, generating >5 TB of synthetic data and metadata.

Results shown in Fig. 6 allow estimating radiation doses at different altitudes, which are currently used for designing new detectors; shielding instruments (e.g., HPC facilities); calculating the reference HEP flux for underground laboratories, volcanic risk assessments and mining prospecting [49].

7.2.2. Elastic infrastructures

The static infrastructures might end up with resource waste for cloud applications with varying loads. In order to adapt to the dynamic demand for computing capacity, the MSWSS service uses the Elastic Cloud Computing Cluster (EC3) tool [39] to create an elastic virtual cluster on top of the EOSC computing resources. EC3 CLI tool provides a set of pre-defined templates which can be combined and customized. It also allows the definition of custom templates with integrated Ansible scripts. This creates a template with strengthened security settings and additional configuration commands specific to the MSWSS service.

Fig. 7 shows the architecture of the MSWSS service with the interaction of the service operator and users. The users interact with the service using the Galaxy portal, where they can manage their data and submit jobs to the elastic virtual cluster for processing. The output data are stored within the MSWSS service and can be downloaded for post-processing tasks.

The service operator deploys the service using the EC3 tool using the customized template. Once the MSWSS service is deployed, the CLUES service monitors the SLURM batch system, deploys new virtual worker nodes as needed, and automatically re-configures the batch system. The deployment and configuration of worker nodes are performed using the Infrastructure Manager (IM) [50] service. To speed up the deployment process, the worker nodes are instantiated from a snapshot of the fully deployed worker node (golden image). This allows to decrease the start-up time from 21 to 5 min. It also helps to solve the issue with pending security updates concerning the vanilla image and the potential need to reboot the worker node for the updates to be applied properly. The service operator maintains the golden image in an up-to-date state. Security is also important for communications inside the virtual cluster. OpenVPN system creates secure connections inside the cluster and protects the data transfers. It also allows spanning the virtual cluster over the resources from different Cloud providers.

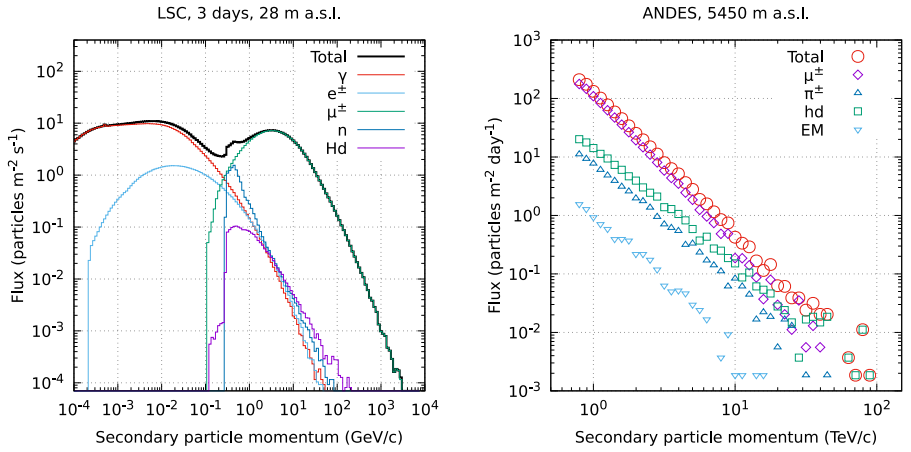

Fig. 6. Left: The energy spectrum of the flux of atmospheric radiation expected at La Serena (LSC) detector in Chile (at sea level) is used for designing, characterizing and calibrating new detectors and sites. Right: the expected flux of >TeV/c particles reaching the summit (5450 m a.s.l.) of the mountain where the ANDES underground laboratory will be installed. Since these particles are capable of traversing up to thousands of metres of rock, being the background signals for neutrino physics experiments and dark matter searches.

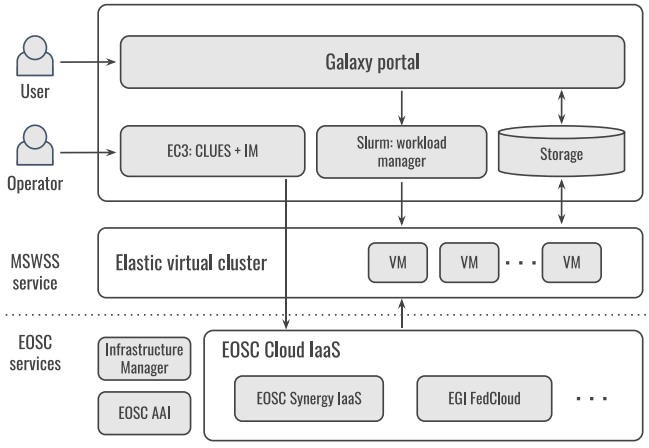

Fig. 7. Architecture of the MSWSS service.

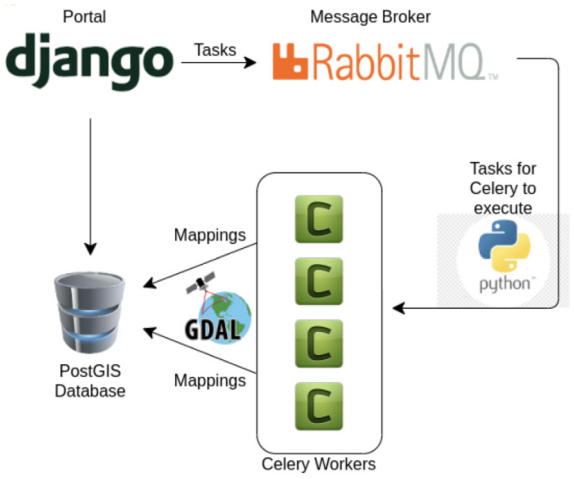

Fig. 8. First WORSICA's architecture for managing the data produced by the service.

7.3. Data storage

Finally, in the field of data storage, we have also observed three different approaches: (i) local storage, (ii) external storage and (iii) hybrid approach. The following subsections analyse the use cases of WORSICA, OpenEBench and UMSA to illustrate the three approaches.

7.3.1. Local storage

WORSICA uses the Dataverse application to manage the data produced by the service and disseminate it to the research community and the public in general.

The first version of WORSICA's data manager was developed as a local repository, using the architecture presented in Fig. 8. This approach raised several constraints to the adoption of a FAIR-compliant data management paradigm, such as (i) the lack of a unique global identifier for the datasets produced, (ii) the data was not accessible and stored in multiple places internal to the service, (iii) nonexistence of metadata for each dataset; and also (iv) the access to the data did not follow the controlled vocabularies that apply to FAIR principles.

In order to surpass the previous oversights, a Dataverse REST API that allows running all necessary operations efficiently, like:

1. the ability to implement in any language only being dependent on the provided interface without any library requirements;
2. the capability to easily maintain the WORSICA code in parallel with Dataverse service updates;
3. Moreover, provide the features required to share sensitive data with the public.

The current architecture of WORSICA's data manager evolved and can be seen in Fig. 9. The WORSICA service is now working with Dataverse to automate the data availability and use services for and distribute credit to the data creator. Dataverse allows the creation of multiple virtual archives called Dataverse collections. Each Dataverse collection contains datasets, and each dataset contains descriptive metadata and data files. Therefore, this version of WORSICA enables datasets to be linked in Dataverse to the appropriate ontologies to increase interoperability and data FAIRness. Variable names can also be included in datasets metadata in the native language (Portuguese) and get Universal Resource Identifier (URI) for those entities in controlled vocabularies (e.g., in the case of WORSICA, a DOI—Digital Object Identifier is created). Furthermore, standardized metadata fields are available in Linked Open Data Cloud through standard machine-to-machine interfaces available in Dataverse.

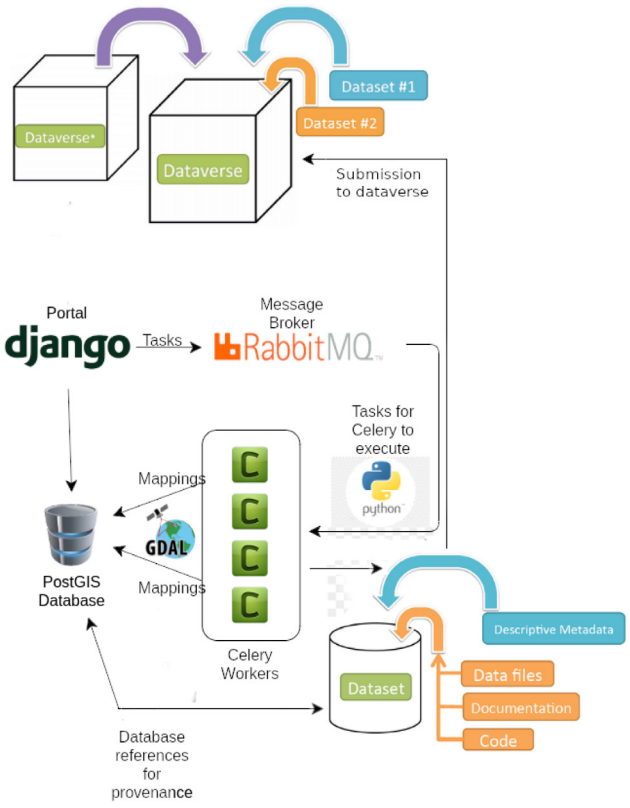

Fig. 9. Current architecture of the data manager implemented in the WORSICA service.

7.3.2. External storage

OpenEBench uses B2SHARE for the long-term availability and storage of scientific benchmarking datasets. Adopting EUDAT's technical standards, data models, and policies helps OpenEbench further enforce the FAIR-compliant data management of the platform. One of the significant capabilities gained through the integration with EUDAT is the minting of Digital Object Identifiers (DOIs) for the benchmarking data collections generated in OpenEBench.

A variety of dataset types are involved in the benchmarking workflows at OpenEBench. Those datasets cover the reference data used as gold-standard data, the predictions submitted by participants, new datasets like the actual results scoring and ranging participants, provenance reports and metrics plots. In this way, a compact and human-readable dataset is ready to be referred to in scientific publications, promoting transparency, reproducibility and data reuse. Furthermore, EUDAT registries provide rich metadata fields for easing data discovery, so thus submitted collections could be annotated with cross-links to OpenEBench for further insights. OpenEBench is one of the EUDAT-registered research communities benefiting from a particular extended metadata model and customized access rules. It facilitated a better integration with OpenEBench, implemented through a REST-based programmatic data publication workflow triggered from the platform Web GUI.

Within the OpenEBench ecosystem, benchmarking data is accessible on the Web or via specific REST and GraphQL APIs. Nextcloud and MongoDB are the technologies used to store datasets and metadata, respectively, continuously operated under a FAIR-compliant data governance plan that considers, for example, unique and accessible identifiers, document versioning, provenance preserving metadata, strict publication rules or a formal benchmarking data model. When a benchmarking events manager or developer participating in a given event is willing to publish its data outside the platform, they register its B2SHARE API token in OpenEBench and initiate the publication process to B2SHARE via the Web GUI:

- OpenEBench composes the data collection with a specific metadata form, validates it, and programmatically submits both the data and metadata to the B2SHARE server.
- Over HTTPS and on behalf of the user, the platform implements the full EUDAT data publication workflow through the B2SHARE REST API. The outcome is a DOI associated with the new registry, which is captured and saved in OpenEBench to keep both systems cross-linked. Eventually, published benchmarking datasets can be consumed using both B2SHARE and OpenBench platforms.

The OpenEBench data flow can be seen in Fig. 10.

7.3.3. Hybrid approach

UMSA leverages three different storage classes for different purposes. First, data acquired by the instruments (mass spectrometers) are stored on a traditional POSIX filesystem, implemented as RAID disk arrays and clusters of GPFS servers, re-exporting the volumes via NFS and CIFS to the clients. This technology has limited scaling (up to petabytes per filesystem). However, it has been proven for decades and is stable, suitable for the primary experimental data that are irrecoverable otherwise. This storage is also backed up weekly to a remote site provided by the CZ national e-Infrastructure.[1]

The UMSA service itself is based on Galaxy, which mounts the primary data filesystem (above) read-only, and it exposes the datafiles to the users via its *Data libraries*,[2] with appropriate access control and without the need to copy large files.

Standard Galaxy setup requires a shared filesystem mounted at both head and worker nodes (alternate setups require many file transfers for each job, which we cannot afford). It can be either a single-tier storage or it can serve as the first tier (cache) of object storage (the second tier). We use the two-tier setup, with an NFS-mounted SSD-only shared filesystem, with typical usage up to dozens of TB only, fast enough not to slow down data processing with I/O, and S3 object storage (provided by the national e-Infrastructure again), which scales up to many PB easily.

8. Related work

As discussed in Section 5, the need for the thematic services to evolve to an Open Science and Open Access paradigm focus on four dimensions: The use of federated and coherent AAI frameworks to seamlessly integrate services and provide mechanisms to restrict access to specific data and resources; the need for services to support the compliance to the FAIR principles, by providing long-term storage for data and metadata and discovery of datasets through persistent identifiers; the integration of workload management systems to deal with multi-tenant and concurrent workloads; and the support of resource management services to adapt infrastructure to the actual workload demand. The cloud computing paradigm offers several advantages, such as scalability, flexibility, cost-effectiveness, high-level platform services and easy access to a wide range of computing resources. This kind of computing can be very beneficial to tackle the above challenges, contributing to open science, and it is applicable for services coming from almost any scientific domain. In this section, we explore the most relevant

[1] http://du.cesnet.cz/.

[2] https://galaxyproject.org/data-libraries/.

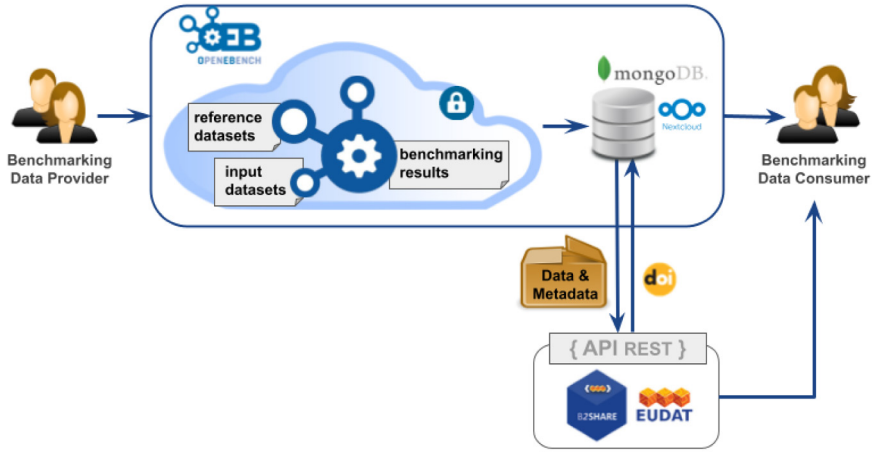

Fig. 10. OpenEBench data flow. Users execute evaluation benchmarking workflows, which use and produce datasets. That data is locally stored in MongoDB (for metadata) and Nextcloud. It can then be submitted to B2SHARE, using their API REST, which will lead to minting a DOI. Then, it is mapped and saved to the OpenEBench database for cross-linking purposes. Published benchmarking data can be consumed using both B2SHARE and OpenBench platforms.

state-of-the-art works dedicated to other approaches for service or application integration to a cloud-based service infrastructure in four scientific areas of the thematic services we have presented in the paper. Moreover, we compare these approaches with the ones carried out by our thematic services.

In the field of EO, a cloud-based approach can help to improve the efficiency and effectiveness of applications and services while increasing accessibility and facilitating share the data with the community. Works like [51] carefully review and compare seven different EO platforms that store and analyse big datasets and have adopted the cloud computing paradigm. Usually, in this domain, the services require a robust and scalable computational infrastructure to serve their users. As we have seen, this is the case with WORSICA, a service aiming to study the evolution of coastal zones and water in inland areas for better coastal, estuarine, and natural resource management. Other services study these subjects, such as [52–55]; however, the WORSICA service has the particularity of being freely accessible to the scientific community and performing simulations on-demand for the entire globe.

Changing the scope, but inside this domain, we can focus on vegetation and agriculture services that use algorithms to estimate evapotranspiration values to study the evolution of forest masses and crops, like [56,57]. The execution of these algorithms supposes high computational demands, both CPU and memory-intensive and the archival of the output data generated consumes a substantial amount of storage resources [58]. Typically in this field, the software packages and libraries that implement ET estimation, such as the ILWIS software [59] or the 'Water' procedure [60], are publicly available for individual use by researchers. However, those implementations have been developed as standalone artefacts typically executed on personal computers, which considerably complicate processing data sharing and might impose limitations on the locally available computing capacity. A prevalent approach to the cloud is taken by the Google Earth Engine (GEE) platform [61], an initiative to facilitate the implementation and execution of scientific workflows that consume satellite imagery as input. However, this service is under service conditions imposed by Google that might only sometimes fit the community needs [62,63]. Conversely, in combination with EOSC services and resources, SAPS follows an innovative approach based on deploying on-demand SAPS instances on top of federated resources. The service is highly configurable, allowing the selection of the algorithms used to estimate ET, and provides standard output data and metadata that can be easily shared among researchers of the area.

Continuing in this domain, but moving now to the EO missions, we can find programs like Copernicus that envisage the use of cloud computing to change the current on-premises processing approach to a cloud-based approach and offer the environments as a service. Entities such as EUMETSAT (European Organization for the Exploitation of Meteorological Satellites) envisaged a new design of a multimission processing infrastructure using the cloud advantages to extend the resources and make them available to several missions. GCore follows these approaches to extend the functionality with the use of cloud resources to break the bottleneck that an on-premise classical system can obtain during the re-processing tasks associated with a particular mission, for example. During this task, massive use of processing power is needed in the nominal platform to affect the mission's nominal operations. The capacity of GCore to deploy additional processors on the cloud is used to reduce the impact of such a task. This approach can also be used for data archiving and cataloguing the products resulting from the processing in the cloud, making them available to perform higher processing levels directly on the cloud using processors as a service specifically defined and published previously in a marketplace.

Moving to the bioinformatic domain, with a cloud-based approach, bioinformatic applications and services can store, process, and analyse large volumes of data in real-time, without the need for expensive and complex on-premise infrastructure, contributing to the open science approach at the same time by facilitating data sharing between researchers and institutions. This is the case of the ScipionCloud thematic service, which offers a ready-to-use infrastructure in the cloud to users that aim to process their CryoEM data. A similar approach is followed by Stion [64], a web application that provides on-demand access to GPU instances on AWS for biomedical researchers to process Cryo-EM data. This solution automatically deploys instances (virtual machines) but does not integrate a batch system. It also sets up an auto-shutdown mechanism to power off instances after a certain period, which is risky and insufficient from our point of view. On the contrary, integrating ScipionCloud with EC3 guarantees a much better way to optimize the use of cloud resources than Stion's implementation. Besides, to use Scipion, providing your own AWS account is mandatory, which might be a drawback for many users in terms of complexity and cost. Another approach found in the CryoEM world is the work done by Cianfrocco's lab at the University of Michigan, which offers several tools in this area. The first interesting tool [65] is integrating an 'AWS batch system' in one of the most popular software packages

used for CryoEM processing; Relion. This batch system allows sending Relion commands to AWS instances, which includes deploying, running and shutting down the instance. This approach was only integrated into an old version of Relion which might imply that it was not a big success in the CryoEM community. The second tool is COSMIC [66], a freely available web platform for submitting cryo-EM jobs through the cloud to Comet, a cluster at the San Diego Supercomputer Center. Although COSMIC has no cost access and this tool is available for everybody, users might prefer to work on their server instead of preparing their processing workflows to be sent step by step to a cluster. Regarding the use of external storage for the sustainability of data, our OpenEBench thematic service is not the only platform designed around FAIR principles that ensures the availability of its data and metadata through EUDAT services or similar data infrastructures. WorkflowHub [67] is applying a similar strategy for publishing scientific workflows, wrapping them into enriched Research Object Crates (RO-Crate 1.1 specification), which include data resources, semantic annotations, and all additional information that guarantees the workflow is reusable and reproducible. In this case, Zenodo is the infrastructure minting the DOI, and the publication process follows DataCite [68] guidelines, one of the broadest cross-domain metadata standards available. Among EOSC resources, other platforms like ROHub [69] propose using EUDAT services as an internal storage solution. ROHub is a platform that enables the management, sharing and preservation of research data as Research Objects. It integrates B2DROP [70] as the underlying technology for the researcher's personal storage space and B2SHARE for DOI minting and sharing.

Regarding astrophysics, it is less common to find cloud-based approaches than in other fields. However, as LAGO's thematic service aims, it could be useful for storing and sharing large amounts of astronomical data, such as data collected from telescopes or simulations. It could also provide remote access to high-performance computing resources for large-scale simulations and data analysis, like the Theoretical Astrophysical Observatory (TAO) [71], an online virtual laboratory houses mock observations of galaxy survey data. ARTI manages software needed for LAGO simulations [48], a data-intensive and highly-complex framework designed to calculate the expected flux of signals in any site around the World and under realistic geomagnetic, atmospheric and detector conditions. As the simulated phenomena, i.e., the interaction of cosmic rays with the atmosphere and the detector response to the resulting flux secondary particles, is, essentially, a sequence of stochastic processes, the simulation performed needs to integrate the flux over long periods to reduce the impact of statistical fluctuations. Such large times, from several hours to days and even years for some applications such as volcano risk assessments, require large computing facilities and storage. Suitable simulations will typically spend tens of weeks in current CPUs and output several TB of data. Examples are the results presented in Section 7.2.1, and deeply in [19,49]. Moreover, all these data must be properly identified, catalogued and curated to accomplish the FAIR principles. However, previous attempts were made to adapt ARTI to its use in distributed high-performance computing infrastructures [72] and for the adoption of data curation standards [73], it was only thanks to the development of OnedataSim [19] within the LAGO TS that it was possible to achieve sufficiently long integration times while complying FAIR principles.

Environmental sciences applications and services can also benefit from a cloud-based approach. The ability to quickly and efficiently process and analyse large amounts of data is crucial in environmental science research, and cloud computing can provide a cost-effective solution for this. For example, applications can manage and analyse data from environmental monitoring systems, such as air and water quality sensors, and share results among the scientific community, contributing to better environmental management and policy-making. This is the case of O3as, a service inspired by the need to have a large plethora of climate model data consistently. Because of the ozone assessment[3] that culminates in a publication every four years [74], scientists need to have a quick way on looking at ozone data from climate data to estimate the time when the amount of ozone in the stratosphere has reached a level pre-ozone hole. Often the models must be collected, the analysis program was redone, and the values recalculated [75,76]. In the field of water and its management, we can find several approaches for modelling water supply systems offering various features. Integrated Tool for Water Supply Systems Management [77] puts together the QGIS database, Epanet hydraulic model, and Google Maps. A web-based EPANET model catalogue and execution environment [78] focuses on model sharing and model viewing. Another example of integrating GIS and the hydraulic model is a tool for an integrated water and wastewater management system in municipal enterprises [79]. According to our knowledge, these services do not provide cloud-based elastic computational back-end, and they do not implement FAIR principles for data sharing, two basic capabilities that the MSWSS thematic service offers, thanks to using resources from the EOSC cloud infrastructure and FAIR data repositories.

Regarding air quality forecasting and, specifically, dust storm forecasts, there are several web-based approaches to offer this information to their users, like METOFFICE,[4] Breezometer,[5] WINDY[6] and Plume Labs.[7] Although most of them have an attractive design, they only offer visual products of their data, not numerical binaries. The amount of data they provide seems to be not big as our thematic service SDS-WAS service. Moreover, some have advertisements and banners that significantly affect the user experience. Cloud-based approaches can also be found in the literature, such as [80,81], whose architectures are interesting. However, they are more focused on the analysis of air pollution parameters. SDS-WAS provides, apart from a huge quantity of materials on dust storm forecast, services for data storage, download and data analysis based on EOSC cloud services (B2SAFE, B2HANDLE). As far as we know, there are no comparable services related to dust storms in terms of collecting a bunch of numerical model outputs, observational data (in situ and satellite) and providing derived graphical (plots) and numerical (skill scores) products with an interactive dashboard application.

Regarding mass-spectrometry data and its application to the environmental domain, numerous tools to process MS data exist, as sub-area-specific reviews analyse [82,83]. They range from proprietary software by laboratory equipment vendors third-party commercial software, and open-source tools of varying quality and maturity. UMSA is dedicated to storing and processing MS data, focusing on GC-MS and untargeted analysis of low-abundant compounds. We can find here the MassBank project [84], an effort to create a public repository for publishing and storing the results of the analyses. Based on Galaxy, UMSA can leverage dozens of mass-spec-related tools published by the community. An ongoing effort is to provide Galaxy-based environments focusing on specific MS applications, e.g. [85]. According to our knowledge, no such effort matches the UMSA purpose. Another approach can be seen with GNPS [86], a web-based computational environment and data treatment environment centred

[3] https://csl.noaa.gov/assessments/ozone/index.html [2022-02-09].
[4] https://www.metoffice.gov.uk/.
[5] https://breezometer.com/.
[6] https://www.windy.com/.
[7] https://air.plumelabs.com/.

around*molecular networking*—visual display of chemical spaces and relationships among compounds. However, most methods leverage MS^2 data, which differs from the focus of UMSA.

Finally and concerning Authentication and Authorization Infrastructure (AAI), over the past two decades, research collaboration has been a driving force for the evolution of the AAI in Europe and around the world. With the work in the national academic federations, eduGAIN and later AARC [87], the foundations have been laid to build a strong research and education AAI that can meet the requirements of scientific collaboration and open science. Communities have relied on federated mechanisms that facilitate using authentication against a federation of institutional IdPs and managing membership centrally through LDAP directories (such as Virtual Organizations [88] and group memberships through Keycloak [89]), and infrastructure and service providers have adopted such technologies to provide interoperability. The thematic services use solutions built on top of those technologies and approaches.

As we have seen, the open science community is strongly interested in adopting cloud-based solutions that offer greater efficiency and scalability, regardless of the scientific domain. This approach also enables collaboration and data sharing among researchers, regardless of location. However, transitioning from a traditional desktop application to a cloud-based architecture can be challenging, even with the assistance of institutions like EOSC. The main objective of this work is to guide on this path for any application seeking to leverage cloud services for the benefit of open science, no matter the science domain.

9. Conclusions

EOSC-Synergy is building Open Science capacities by developing ten data-intensive thematic services oriented to four different scientific disciplines. The adaptation, improvement, and quality assessment of those services on a federated data infrastructure strongly aligns with the objectives of EOSC [90]. A key factor for the success of Open Science and the EOSC initiative [91] is performance, i.e., how EOSC as an ecosystem operates and how the resources are used and acknowledged by the users. All the services consume services from the EOSC catalogue, which will provide feedback on the usability and relevance of the model. It is essential to state that EOSC is defined as "A web of FAIR data and related services" and relies on three principles: Increasing the availability of FAIR data, supporting the creation of federated infrastructures to support Open Science and the promotion of Open Science as the new normal. Other worldwide Open Science initiatives share these principles.

However, characterizing new applications to join is key to evaluating the rightmost services to be used. This is why the EOSC-Synergy project is developing best practices and experiences to promote the adoption of Open Science by the research communities by expanding and building knowledge on common interfaces, standards, and best practices. This paper presented an application modelling proposal based on the previous analysis of gaps and bottlenecks performed by ten different services from four disciplines.

The ten thematic services have been analysed concerning four dimensions (authentication and authorization; resource management and offering model; workload management including containerization; and data storage and preservation), identifying gaps and bottlenecks. Those requirements are common to many other scientific services.

The findings on these four categories can be summarized as follows:

- Authentication must consider federated AAI solutions based on institutional Identity Providers, either general-purpose research and academic solutions (such as EduGAIN) or thematic instances (such as the Life Sciences AAI). Those solutions not only provide higher user acceptance and convenience but also strongly facilitate the integration of multiple third-party solutions. Moreover, the usage of institutional IdPs facilitates the processes of identity verification and trust by the service providers, which could be key in access-restricted environments (Open Science does not mean unconditionally open, and restrictions are necessary when dealing with sensitive data).
- Data management. Compliance with the FAIR principles is an effort-consuming task. First, data should be collected and prepared to be annotated appropriately and follow a standard schema for data and metadata (Interoperable and Reusable). Second, data should be preserved in an Accessible and long-term lasting archive. Third, data should be Findable by registering the metadata and accessible endpoint in a popular metadata harvester.
- Workload management. Workload management should consider concurrent and computing-intensive workloads when exposing scientific data and applications as a service. Thematic services should consider batch queues, container management solutions, workflow engines or combinations of them to run bags of jobs dealing with datasets efficiently.
- Deployment of resources. Open access services require resources to deal with the workloads. Cloud infrastructures could provide resources on demand, but this will require adapting the application architecture, using Infrastructure as Code descriptions to some extent to configure resources on the fly and elastic workload management services that could interact with the cloud resource management systems.

We selected ten services from the EOSC Marketplace to address these requirements. In a nutshell, AAI solutions such as EGI Check-in, eduTEAMS, B2Access and life-science AAI are mature enough to provide a coherent authentication model for a whole application. Applications that require a dynamic backend or on-demand deployment found Infrastructure Manager (IM) and Elastic Compute Clusters in the Cloud (EC3) as reasonable solutions to describe their infrastructure as code and deploy resources according to their workload. Depending on the workload type, job management is driven by SLURM batch queues or Kubernetes services. Preservation of data is obtained through EGI DataHub or EUDAT's B2SAFE and B2SHARE storage services, registering persistent identifiers through B2Handle. Finally, services needing local storage used Dataverse as an OAI-PMH on-premise storage.

The impact of EOSC in the thematic services of EOSC-Synergy is mainly composed of three main facts. Firstly, the capacity expansion through the federation of computing, storage, and data resources aligned with the EOSC and FAIR policies and practices. Secondly, software and service quality evaluation of the thematic services is critical to improve robustness and reliance and increase user experience. EOSC-Synergy also focuses on transverse training to facilitate the adoption of technologies and the use of thematic services. Finally, the cross-fertilization between different thematic areas has allowed the collaboration between thematic services to take advantage of the developments, solutions, experiences and best practices on AARC [87]. The foundations have been laid to build an intense research and education AAI that can meet the requirements of scientific collaboration and open science. Communities have relied on federated mechanisms that facilitate using authentication against a federation of institutional IdPs and managing membership centrally through LDAP directories (such as Virtual Organizations [88] and group memberships

through Keycloak [89]), and infrastructure and service providers have adopted such technologies to provide interoperability. The thematic services use solutions built on top of those technologies and approaches.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

This work was supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No 857647, EOSC-Synergy, European Open Science Cloud - Expanding Capacities by building Capabilities. Moreover, this work is partially funded by grant No 2015/24461-2, São Paulo Research Foundation (FAPESP). Francisco Brasileiro is a CNPq/Brazil researcher (grant 308027/2020-5).

References

[1] Lisa O'Conner (Ed.), 2021 IEEE 17th International Conference on EScience (eScience), IEEE, Innsbruck, Austria, 2021, http://dx.doi.org/10.1109/eScience51609.2021.
[2] Foster, Open science definition, 2016, https://www.fosteropenscience.eu/foster-taxonomy/open-science-definition.
[3] E. Commission, E-infrastructures definition, 2016, https://ec.europa.eu/digital-single-market/en/e-infrastructures.
[4] I. Blanquer, G. Brasche, D. Lezzi, Requirements of scientific applications in cloud offerings, in: Proceedings of the 2012 Sixth Iberian Grid Infrastructure Conference, IBERGRID, Vol. 12, 2012, pp. 173–182.
[5] FAIR Data Maturity Model. Specification and Guidelines, The FAIR Data Maturity Model Working Group, 2020, http://dx.doi.org/10.15497/rda00050.
[6] E. Commission, EOSC European partnership proposal, 2020, https://bit.ly/2RKYmak.
[7] EOSC-Synergy, EOSC-synergy portal, 2020, https://www.eosc-synergy.eu/.
[8] E. Enhance, E. Future, EOSC portal catalogue & marketplace, 2021, https://marketplace.eosc-portal.eu/services.
[9] WORSICA, LNEC portal - water monitoring sentinel cloud platform, 2021, http://worsica.lnec.pt.
[10] WORSICA, Water monitoring sentinel cloud platform, 2021, https://worsica.incd.pt/index/.
[11] J. Cunha, T.E. Pereira, E. Pereira, I. Rufino, C. Galvão, F. Valente, F. Brasileiro, A high-throughput shared service to estimate evapotranspiration using landsat imagery, Comput. Geosci. 134 (2020) 104341, http://dx.doi.org/10.1016/j.cageo.2019.104341.
[12] M. Rodriguez, https://u.i3~m.upv.es/b7g7m. 2018.
[13] GCore, Gcore overview at EOSC synergy, 2022, https://www.eosc-synergy.eu/thematic-services/g-core/.
[14] J. de la Rosa-Trevín, A. Quintana, L. del Cano, A. Zaldívar, I. Foche, J. Gutiérrez, J. Gómez-Blanco, J. Burguet-Castell, J. Cuenca-Alba, V. Abrishami, J. Vargas, J. Otón, G. Sharov, J. Vilas, J. Navas, P. Conesa, M. Kazemi, R. Marabini, C. Sorzano, J. Carazo, Scipion: A software framework toward integration, reproducibility and validation in 3D electron microscopy, J. Struct. Biol. 195 (1) (2016) 93–99, http://dx.doi.org/10.1016/j.jsb.2016.04.010.
[15] Lessons learned: Recommendations for establishing critical periodic scientific benchmarking, 2017, http://dx.doi.org/10.1101/181677, BioRxiv.
[16] ELIXIR, ELIXIR European intergovernmental organisation, 2021, https://elixir-europe.org/.
[17] I. Sidelnik, H. Asorey, et al., LAGO: The latin American giant observatory, Nucl. Instrum. Methods Phys. Res. A 876 (2017) 173–175, http://dx.doi.org/10.1016/j.nima.2017.02.069.
[18] LAGO Collaboration, Latin American giant observatory, 2021, http://lagoproject.net. Accessed.
[19] A.J. Rubio-Montero, R. Pagán-Muñoz, R. Mayo-García, A. Pardo-Díaz, I. Sildelnik, H. Asorey, A novel cloud-based framework for standardized simulations in the latin American giant observatory (LAGO), in: 2021 Winter Simulation Conference, WSC, IEEE, Phoenix, USA, 2021, pp. 1–12, http://dx.doi.org/10.1109/WSC52266.2021.9715360.
[20] S. Basart, S. Nickovic, E. Terradellas, E. Cuevas, C. Pérez García-Pando, G. García-Castrillo, E. Werner, F. Benincasa, The WMO sds-WAS regional center for northern africa, middle east and europe, in: E3S Web of Conferences, in: E3S Web of Conferences, vol. 99, 2019, p. 04008, http://dx.doi.org/10.1051/e3sconf/20199904008.
[21] S. Basart, E. Terradellas, E. Cuevas, O. Jorba, F. Benincasa, J.M. Baldasano, The Barcelona dust forecast center: The first WMO regional meteorological center specialized on atmospheric sand and dust forecast, in: EGU General Assembly Conference Abstracts, in: EGU General Assembly Conference Abstracts, 2015, p. 13309.
[22] UMSA, UMSA overview at EOSC-synergy, 2022, https://www.eosc-synergy.eu/thematic-services/umsa/.
[23] MSWSS, Modelling service for water supply systems, 2021, https://mswss.ui.savba.sk:8443.
[24] O3AS, O3AS portal - ozone assessment cloud platform, 2022, https://o3as.data.kit.edu.
[25] G. King, An introduction to the dataverse network as an infrastructure for data sharing, Sociol. Methods Res. 36 (2) (2007) 173–199, http://dx.doi.org/10.1177/0049124107306660.
[26] M. Viljoen, Ł. Dutka, B. Kryza, Y. Chen, Towards European open science commons: The EGI open data platform and the EGI DataHub, in: 2nd International Conference on Cloud Forward: From Distributed To Complete Computing, in: Procedia Computer Science, vol. 97, 2016, pp. 148–152, http://dx.doi.org/10.1016/j.procs.2016.08.294.
[27] D. Lecarpentier, P. Wittenburg, W. Elbers, A. Michelini, R. Kanso, P. Coveney, R. Baxter, EUDAT: a new cross-disciplinary data infrastructure for science, Int. J. Digit. Curation 8 (1) (2013) 279–287, http://dx.doi.org/10.2218/ijdc.v8i1.260.
[28] A.B. Yoo, M.A. Jette, M. Grondona, SLURM: Simple linux utility for resource management, in: D. Feitelson, L. Rudolph, U. Schwiegelshohn (Eds.), Job Scheduling Strategies for Parallel Processing (JSSPP 2003), in: Lecture Notes in Computer Science, vol. 2862, Springer Berlin Heidelberg, Berlin, Heidelberg, 2003, pp. 44–60, http://dx.doi.org/10.1007/10968987_3.
[29] J. Goecks, A. Nekrutenko, J. Taylor, Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences, Genome Biol. 11 (8) (2010) R86, http://dx.doi.org/10.1186/gb-2010-11-8-r86.
[30] EGI, EGI cloud compute service, 2021, https://www.egi.eu/services/cloud-compute/.
[31] EGI, EGI high-throughput compute, 2021, https://www.egi.eu/services/high-throughput-compute/.
[32] EGI, EGI workload manager, 2021, https://www.egi.eu/services/workload-manager/.
[33] EGI, EGI DataHub, 2021, https://www.egi.eu/services/datahub/.
[34] EUDAT, B2safe, keep research data safe via data management policies, 2021, https://sp.eudat.eu/catalog/resources/5d81cb5b-3640-4430-b46e-fc652e06a4db.
[35] M. Caballer, I. Blanquer, G. Moltó, C. de Alfonso, Dynamic management of virtual infrastructures, J. Grid Comput. 13 (1) (2015) 53–70, http://dx.doi.org/10.1007/s10723-014-9296-5.
[36] EGI, Dynamic DNS for VMs in EGI cloud, 2021, https://docs.egi.eu/users/cloud-compute/dynamic-dns/.
[37] V. Tran, Fedcloud client documentation, 2021, https://fedcloudclient.fedcloud.eu/.
[38] EUDAT, B2FIND official webpage, find research data, research data portal, 2021, https://sp.eudat.eu/catalog/resources/33bc21d5-f53d-4eed-9a15-56f98f5c7f69.
[39] A. Calatrava, E. Romero, G. Moltó, M. Caballer, J.M. Alonso, Self-managed cost-efficient virtual elastic clusters on hybrid cloud infrastructures, Future Gener. Comput. Syst. 61 (2016) 13–25, http://dx.doi.org/10.1016/j.future.2016.01.018.
[40] EGI, EGI check-in, 2021, https://www.egi.eu/services/check-in/.
[41] EUDAT, Official webpage B2ACCESS identity & authorisation, 2021, https://sp.eudat.eu/catalog/resources/d04af0f5-2253-4ee4-8181-3a5a961ccd49.
[42] GEANT, Eduteams web site/, 2021, https://eduteams.org/.
[43] M. Linden, M. Procházka, I. Lappalainen, et al., Common ELIXIR service for researcher authentication and authorisation, in: F1000Research, Vol. 7, ELIXIR, 2018, http://dx.doi.org/10.12688/f1000research.15161.1.
[44] T. Binz, U. Breitenbücher, O. Kopp, F. Leymann, in: A. Bouguettaya, Q.Z. Sheng, F. Daniel (Eds.), TOSCA: Portable Automated Deployment and Management of Cloud Applications, Springer New York, New York, NY, 2014, pp. 527–549, http://dx.doi.org/10.1007/978-1-4614-7535-4_22.
[45] EGI, Check-in guide for service providers, 2022, https://docs.egi.eu/providers/check-in/sp/.
[46] INSTRUCT-ERIC, Instruct ERIC structural biology web site, 2021, https://instruct-eric.eu/.

[47] S.B. Pablo Orviz, Jenkins pipeline library - official documentation, 2022, https://indigo-dc.github.io/jenkins-pipeline-library/2.0.0/index.html.
[48] H. Asorey, L.A. Núñez, M. Suárez-Durán, Preliminary results from the latin American giant observatory space weather simulation chain, Space Weather 16 (5) (2018) 461–475, http://dx.doi.org/10.1002/2017SW001774.
[49] A.J. Rubio-Montero, R. Pagán-Muñoz, R. Mayo-García, A. Pardo-Díaz, I. Sildelnik, H. Asorey, The EOSC-synergy cloud services implementation for the latin American giant observatory (LAGO), in: 37th International Cosmic Ray Conference (ICRC2021). PoS(ICRC2021)261, in: Proceedings of Science (PoS), vol. 395, SISSA, Berlin, Germany, 2021, p. 261, http://dx.doi.org/10.22323/1.395.0261.
[50] M. Caballer, I. Blanquer, G. Moltó, C. Alfonso, Dynamic management of virtual infrastructures, J. Grid Comput. 13 (1) (2015) 53–70, http://dx.doi.org/10.1007/s10723-014-9296-5.
[51] V.C.F. Gomes, G.R. Queiroz, K.R. Ferreira, An overview of platforms for big earth observation data management and analysis, Remote Sens. 12 (8) (2020) http://dx.doi.org/10.3390/rs12081253.
[52] G. Australia, DEA coastlines, 2022, https://cmi.ga.gov.au/data-products/dea/581/dea-coastlines. Accessed.
[53] R. Bishop-Taylor, S. Sagar, L. Lymburner, I. Alam, J. Sixsmith, Sub-pixel waterline extraction: Characterising accuracy and sensitivity to indices and spectra, Remote Sens. 11 (24) (2019) http://dx.doi.org/10.3390/rs11242984.
[54] R. Bishop-Taylor, R. Nanson, S. Sagar, L. Lymburner, Mapping Australia's dynamic coastline at mean sea level using three decades of landsat imagery, Remote Sens. Environ. 267 (2021) 112734, http://dx.doi.org/10.1016/j.rse.2021.112734.
[55] Copernicus, DIAS services, 2022, https://www.copernicus.eu/en/access-data/dias. Accessed.
[56] Q. Mu, M. Zhao, S.W. Running, Improvements to a MODIS global terrestrial evapotranspiration algorithm, Remote Sens. Environ. 115 (8) (2011) 1781–1800, http://dx.doi.org/10.1016/j.rse.2011.02.019.
[57] Z. Wan, K. Zhang, X. Xue, Z. Hong, Y. Hong, J. Gourley, Water balance-based actual evapotranspiration reconstruction from ground and satellite observations over the conterminous United States, Water Resour. Res. 51 (8) (2015) 6485–6499, http://dx.doi.org/10.1002/2015WR017311.
[58] S. Goodman, A. BenYishay, Z. Lv, D. Runfola, GeoQuery: Integrating HPC systems and public web-based geospatial data tools, Comput. Geosci. 122 (2019) 103–112, http://dx.doi.org/10.1016/j.cageo.2018.10.009.
[59] M. Abouali, J. Timmermans, J.E. Castillo, B.Z. Su, A high performance GPU implementation of surface energy balance system (SEBS) based on CUDA-C, Environ. Model. Softw. 41 (2013) 134–138, http://dx.doi.org/10.1016/j.envsoft.2012.12.005.
[60] G. Olmedo, S. Ortega-Farias, D. Fonseca-Luengo, D. de la Fuente-Saiz, F. Peñailillo, Water: actual evapotranspiration with energy balance models, 2017, R Package Version 0.6.
[61] G.E.E. Team, Google earth engine: A planetary-scale geo-spatial analysis platform, 2022, https://earthengine.google.com/.
[62] J. Padarian, B. Minasny, A. McBratney, Using google's cloud-based platform for digital soil mapping, Comput. Geosci. 83 (2015) 80–88, http://dx.doi.org/10.1016/j.cageo.2015.06.023.
[63] M. Amani, A. Ghorbanian, S.A. Ahmadi, M. Kakooei, A. Moghimi, S.M. Mirmazloumi, S.H.A. Moghaddam, S. Mahdavi, M. Ghahremanloo, S. Parsian, Q. Wu, B. Brisco, Google earth engine cloud computing platform for remote sensing big data applications: A comprehensive review, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 13 (2020) 5326–5350, http://dx.doi.org/10.1109/JSTARS.2020.3021052.
[64] S. Bhatkar, Stion – a software as a service for cryo-EM data processing on AWS, 2021, https://aws.amazon.com/blogs/hpc/stion-a-saas-for-cryo-em-data-processing-on-aws.
[65] M.A. Cianfrocco, I. Lahiri, F. DiMaio, A. Leschziner, Cryoem-cloud-tools: A software platform to deploy and manage cryo-EM jobs in the cloud, J. Struct. Biol. 203 (2018) 230–235, http://dx.doi.org/10.1016/j.jsb.2018.05.014.
[66] M.A. Cianfrocco, M. Wong-Barnum, C. Youn, R. Wagner, A. Leschziner, COSMIC2: A science gateway for cryo-electron microscopy structure determination, in: Practice and Experience in Advanced Research Computing 2017, 2017, pp. 1–5.
[67] R. Ferreira, et al., WorkflowHub: Community Framework for Enabling Scientific Workflow Research and Development, IEEE, 2020, http://dx.doi.org/10.1109/WORKS51914.2020.00012.
[68] DataCite, Locate, identify, and cite research data with the leading global provider of DOIs for research data, 2023, https://datacite.org/.
[69] Raúl PalmaEmail, et al., Rohub — A digital library of research objects supporting scientists towards reproducible science, 2014, http://dx.doi.org/10.1007/978-3-319-12024-9_9.
[70] EUDAT, B2DROP: Sync and share research data, 2021, https://eudat.eu/services/userdoc/b2drop.
[71] M. Bernyk, D.J. Croton, C. Tonini, L. Hodkinson, A.H. Hassan, T. Garel, A.R. Duffy, S.J. Mutch, G.B. Poole, S. Hegarty, The theoretical astrophysical observatory: Cloud-based mock galaxy catalogs, Astrophys. J. Suppl. Ser. 223 (1) (2016) 9, http://dx.doi.org/10.3847/0067-0049/223/1/9.
[72] H. Asorey, L.A. Núñez, M. Suárez-Durán, L.A. Torres-Niño, M. Rodríguez-Pascual, A.J. Rubio-Montero, R. Mayo-García, The latin American giant observatory: A successful collaboration in latin america based on cosmic rays and computer science domains, in: 16th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid), IEEE, Cartagena, Colombia, 2016, pp. 707–711, http://dx.doi.org/10.1109/CCGrid.2016.110.
[73] M. Rodríguez-Pascual, G. LaRocca, C. Kanellopoulo, C. Carrubba, G. Inserra, R. Ricceri, H. Asorey, A.J. Rubio-Montero, E. Núñez-González, L.A. Núñez, O. Prnjat, R. Barbera, R. Mayo-García, A resilient methodology for accessing and exploiting data and scientific codes on distributed environments, in: 18th IEEE International Conference on Computational Science and Engineering, CSE, IEEE, Porto, Portugal, 2015, pp. 319–323, http://dx.doi.org/10.1109/CSE.2015.27.
[74] World Meteorological Organization (WMO), Scientific Assessment of Ozone Depletion: 2022, GAW Report No. 278, 2022, p. 509 pp, https://csl.noaa.gov/assessments/ozone/.
[75] S.S. Dhomse, D. Kinnison, M.P. Chipperfield, R.J. Salawitch, I. Cionni, M.I. Hegglin, N.L. Abraham, H. Akiyoshi, A.T. Archibald, E.M. Bednarz, S. Bekki, P. Braesicke, N. Butchart, M. Dameris, M. Deushi, S. Frith, S.C. Hardiman, B. Hassler, L.W. Horowitz, R.-M. Hu, P. Jöckel, B. Josse, O. Kirner, S. Kremser, U. Langematz, J. Lewis, M. Marchand, M. Lin, E. Mancini, V. Marécal, M. Michou, O. Morgenstern, F.M. O'Connor, L. Oman, G. Pitari, D.A. Plummer, J.A. Pyle, L.E. Revell, E. Rozanov, R. Schofield, A. Stenke, K. Stone, K. Sudo, S. Tilmes, D. Visioni, Y. Yamashita, G. Zeng, Estimates of ozone return dates from chemistry-climate model initiative simulations, Atmos. Chem. Phys. 18 (11) (2018) 8409–8438, http://dx.doi.org/10.5194/acp-18-8409-2018.
[76] J. Keeble, B. Hassler, A. Banerjee, R. Checa-Garcia, G. Chiodo, S. Davis, V. Eyring, P.T. Griffiths, O. Morgenstern, P. Nowack, G. Zeng, J. Zhang, G. Bodeker, S. Burrows, P. Cameron-Smith, D. Cugnet, C. Danek, M. Deushi, L.W. Horowitz, A. Kubin, L. Li, G. Lohmann, M. Michou, M.J. Mills, P. Nabat, D. Olivié, S. Park, Ø. Seland, J. Stoll, K.-H. Wieners, T. Wu, Evaluating stratospheric ozone and water vapour changes in CMIP6 models from 1850 to 2100, Atmos. Chem. Phys. 21 (6) (2021) 5015–5061, http://dx.doi.org/10.5194/acp-21-5015-2021.
[77] J. Pérez-Padillo, J.G. Morillo, E.C. Poyato, P. Montesinos, Open-source application for water supply system management: Implementation in a water transmission system in southern Spain, Water 13 (24) (2021) 3652, http://dx.doi.org/10.3390/w13243652.
[78] T. Bayer, D.P. Ames, T.G. Cleveland, Design and development of a web-based EPANET model catalogue and execution environment, Ann. GIS 27 (3) (2021) 247–260, http://dx.doi.org/10.1080/19475683.2021.1936171.
[79] W. Kruszyński, J. Dawidowicz, Computer modeling of water supply and sewerage networks as a tool in an integrated water and wastewater management system in municipal enterprises, J. Ecol. Eng. 21 (2) (2020) 261–266, http://dx.doi.org/10.12911/22998993/117533.
[80] C.-T. Yang, Y.-W. Chan, J.-C. Liu, B.-S. Lou, An implementation of cloud-based platform with R packages for spatiotemporal analysis of air pollution, J. Supercomput. 76 (2020) 1416–1437, http://dx.doi.org/10.1007/s11227-017-2189-1.
[81] C. Zhang, J. Yan, Y. Li, F. Sun, J. Yan, D. Zhang, X. Rui, R. Bie, Early air pollution forecasting as a service: An ensemble learning approach, in: 2017 IEEE International Conference on Web Services, ICWS, IEEE, 2017, pp. 636–643, http://dx.doi.org/10.1109/ICWS.2017.76.
[82] C. Cand, H. Jand, J. Tanner, J. Cheng, Bioinformatics methods for mass spectrometry-based proteomics data analysis, Int. J. Mol. Sci. 21 (8) (2020) 2873, http://dx.doi.org/10.3390/ijms21082873.
[83] L. Yi, N. Dong, Y. Yun, B. Deng, D. Ren, S. Liu, Y. Liang, Chemometric methods in data processing of mass spectrometry-based metabolomics: A review, Anal. Chim. Acta 914 (2016) 17–34, http://dx.doi.org/10.1016/j.aca.2016.02.001.
[84] H. Horai, et al., MassBank: a public repository for sharing mass spectral data for life sciences, J. Mass Spectrom. 45 (7) (2010) 703–714, http://dx.doi.org/10.1002/jms.1777.
[85] Y. Guitton, M. Tremblay-Franco, G.L. Corguillé, J.-F. Martin, M. Pétéra, et al., Create, run, share, publish, and reference your LC–MS, FIA–MS, GC–MS, and NMR data analysis workflows with the Workflow4Metabolomics 3.0 galaxy online infrastructure for metabolomics, Int. J. Biochem. Cell Biol. 93 (2017) 89—101, http://dx.doi.org/10.1016/j.biocel.2017.07.002.
[86] M. Wang, J.J. Carver, V.V. Phelan, L.M. Sanchez, N. Garg, Y. Peng, D.D. Nguyen, et al., Sharing and community curation of mass spectrometry data with global natural products social molecular networking, Nature Biotechnol. 34 (8) (2016) PMID: 27504778.
[87] AARC blueprint architecture 2019, 2019, http://dx.doi.org/10.5281/zenodo.3672785.
[88] R. Alfieri, R. Cecchini, V. Ciaschini, L. dell'Agnello, Á. Frohner, A. Gianoli, K. Lõrentey, F. Spataro, VOMS, an authorization system for virtual organizations, in: F. Fernández Rivera, M. Bubak, A. Gómez Tato, R. Doallo (Eds.),

Grid Computing, Springer Berlin Heidelberg, Berlin, Heidelberg, 2004, pp. 33–40, http://dx.doi.org/10.1007/978-3-540-24689-3_5.

[89] J. Munke, M. Hayek, M. Golasowski, R.J. García-Hernández, F. Donnat, C. Koch-Hofer, P. Couvee, S. Hachinger, J. Martinovič, Data system and data management in a federation of HPC/Cloud centers, in: O. Terzo, J. Martinovič (Eds.), HPC, Big Data, and AI Convergence Towards Exascale Challenge and Vision, CRC Press, 2022, pp. 60–77, http://dx.doi.org/10.1201/9781003176664-4.

[90] European Open Science Cloud Partnership, Draft proposal for the European open science cloud (EOSC) partnership, 2023, https://eosc.eu/partnership. Accessed.

[91] E. Commission, D.-G. for Research, Innovation, Solutions for a sustainable EOSC : a FAIR Lady (olim Iron Lady) report from the EOSC Sustainability Working Group, Publications Office, 2020, http://dx.doi.org/10.2777/870770.