Co-funded by the European Union

D3.1 Report on the EUreka3D services and resource hub: design and implementation

Due date M4: 30/4/2023

Dissemination level: Public

Authors:
Michal Orzechowski, Bartosz Kryza, Lukasz Dutka (CYF)

Ignacio Lamata Martinez (EGI Foundation)

HISTORY OF CHANGES			
Version	Date	Author	Comments
0.1	14/04/2023	Ignacio Lamata Martinez (EGI), Michal Orzechowski, Bartosz Kryza, Lukasz Dutka (CYF)	First draft
0.2	21/04/2023	Hugo Manguinhas (EF)	Peer review
0.3	27/04/2023	Antonella Fresa, Valentina Bachi, John Balean (Photoconsortium), Giuseppe La Rocca, Gergely Sipos (EGI), Lianne Heslinga (EF)	Added suggestions and modifications
1.0	28/04/2023	Valentina Bachi (Photoconsortium)	Final version

TABLE OF CONTENTS

EXECUTIVE SUMMARY

The Digital Cultural Heritage discipline is becoming increasingly important in the scientific world, but the European Cultural Heritage (CH) domain lacks efficient mechanisms and a proper infrastructure to take advantage of recent technological advances to support the digitisation, preservation and presentation of CH data. EUreka3D aims to contribute to the current digital transformation of the CH domain with a range of services that are immediately available and tested in real-life environments.

In the scope of the EUreka3D project, the DataHub and services under development and described in this document will enable Cultural Heritage Institutions to access a virtual data space, use storage and computing resources to manage their 3D assets in a secure and easy to use and integrate manner. Additionally, this approach is intended to be an illustrative use case to provide an innovative cost-effective solution to data storage and the online delivery of heritage assets, by providing CHIs with a secure environment, which is technically more flexible and energy efficient.

This deliverable is the first technical deliverable of the project and presents the design and initial implementation of the EUreka3D services and resource hub (what can be considered the "EUreka3D infrastructure"). The deliverable presents the general concepts behind the services such as Cloud computing and data management, followed by description of the service architecture and design, used technologies and hardware resources provisioned for deployment of the service. The content is based on the initial set of requirements that have been captured from the project partners, which include cultural heritage institutions, Europeana Foundation and one Europeana accredited aggregator (i.e. Photoconsortium) - and on the technical offerings that are available from the technical partners, CYFRONET and EGI.

The final product will be extended during the project to meet its future technological infrastructure and capacity needs, so it will be fully described in Deliverable 3.3 at the end of the project (December 2024).

1. INTRODUCTION

The Digital Cultural Heritage (DCH) discipline is becoming increasingly important in the scientific world, being supported by EC Recommendations, such as 2011/711/EU or the most recent recommendation of 10.11.2021 on a common European data space for cultural heritage [1] on the digitisation and online accessibility of cultural material and its digital preservation, on assets such as monuments, sites, museum objects and other tangible heritage, especially those at risk. The EC has made this priority clear with the investment in the field of CH of more than €1 billion over the last two decades.

However, the European CH domain lacks efficient mechanisms and a proper infrastructure to take advantage of recent technological advances to support the digitisation, preservation and presentation of CH data. There is no agreement in the sector for a common framework to support CH data interoperability and workflows. Initiatives such as Europeana play a leading role in the presentation of European cultural heritage assets to the different scientific communities and the general public, but still lack some 3D functionality, have limited storage features, is not yet integrated with the European Open Science Cloud (EOSC) and rely on commercial solutions, that often operate outside of the EU, to render 3D objects. This is why a new, innovative and Europe-centred platform is needed. The European Commission is acknowledging the need and for this purpose it has launched, under Horizon Europe calls, the new initiative of a European Collaborative Cloud for Cultural Heritage (ECCCH) [2]. However, the ECCCH operational services are not yet ready, the calls are expected to produce their first concrete results not earlier than the next couple of years, and the implementation of the future interoperability of ECCCH with Europeana is still under investigation.

EUreka3D aims to contribute to the current digital transformation of the CH domain with a range of services that are immediately available and tested in real-life environments. The infrastructure described in the present deliverable will be established as a service and resource hub, which will serve as the foundation for the next generation of CH systems and services. This robust and innovative IT infrastructure will support the storage, management and presentation of CH data in a secure manner, will be compatible with the existing Europeana platform, and will be open for extension to new functionality and capabilities as the needs of the CH domain evolve. The IT infrastructure is used experimentally in a pilot that combines the work of WP2 (where digital contents are generated) and WP3 (where digital services are made available).

Specifically, the technical work of EUreka3D will focus on the following aspects:

- **Digitisation**, conducting a comprehensive approach to the needs and workflow of high-quality 3D digitisation of CH assets, based on the recommendations of EU funded VIGIE Study 2020/654 [3], including guidelines and standards to promote digitisation skills, workflows and methodologies, to balance the knowledge and skills of the different European CH institutions. This work is implemented in WP2.
- **Preservation**, using recent technological advances to store and manage data in a secure and efficient manner. Establishing the basis to lead the CH domain world-wide in terms of storage capabilities that are able to cope with CH data and its challenging data requirements. This work is implemented in WP3.
- **Presentation**, enabling both existing initiatives, such as Europeana, and future applications to use the 3D content digitised during EUreka3D and afterwards, in order to present the data in a suitable form to the end user (such as raw data, as a rendered interactive 3D object and so on). This area is crucial to ensure greater access to and use of cultural material. Contents generated in WP2 and

stored in WP3 will be ingested in Europeana under the support of Photoconsortium accredited aggregator. Demonstrations of applications will be prepared in EUreka3D as a combination of the efforts of WP2, WP3 and WP4.

In fact, these aspects contribute to the creation of a common European framework for the CH discipline.

This report describes the preliminary version of the EUreka3D services and resource hub, as a starting point for the platform that will build the capability needed for the project.

As the EUreka3D project aims to address the need for the digital transformation of Cultural Heritage Institutions (CHIs), this requires an overall rethinking of the underlying work processes and business models. The project's vision is to contribute to the deployment of a European Data Space for Cultural Heritage, which involves CHIs of all sizes embracing advanced digitisation, holistic representation of tangible objects, and re-use approaches. From a technical viewpoint, CHIs need to move away from former ICT generations and towards a comprehensive, integrated, cloud-based IT infrastructure that reaches outside the borders of the individual institute. The EUreka3D project will offer a knowledge centre and service and resource hub, based on a smart technical infrastructure whose services are registered on the European Open Science Cloud (EOSC). CHIs can access a virtual space of knowledge, use storage and computing resources to manage their 3D assets, and create, manage, archive, preserve and share digitised objects with a focus on 3D digitisation and knowledge modelling. Overall, the project will improve the digital capacity of the cultural sector and create high-value datasets directly beneficial to CHIs.

In a nutshell, the EUreka3D services and resource hub is based on EGI DataHub service powered by open-source Onedata data access and management platform. It enables CHIs to create, manage, archive, preserve and share digitised objects, in particular 3D digitization for the semantically enriched 3D records. In this light, the use cases identified in the project can be grouped into three very general categories:

Integration with content providers

- Content upload through several interfaces (POSIX, S3, REST)
- Metadata creation and curation in relevant formats, e.g. Europeana Data Model
- Content search and discovery based on the metadata

Training and capacity building

- Online access to stored objects using web based clients, as well as standard protocols
- Easy content sharing between users

Content aggregation in Europeana

- Access to the stored data objects and their metadata through standard interfaces including S3, REST and OAI-PMH

Detailed use cases and requirements will be collected in the next months and stored in a separate document in the project's knowledge space, which will be updated continuously and submitted to the EC with six-month interval reviews. The requirements document will work as the interface between IT and Content Providers/CH community in the EUreka3D project.

1.1 Role of this deliverable in the project

This report outlines the initial design and current status of the EUreka3D services and resource hub. The document is the first technical deliverable of the project. The content is based on the initial set of requirements that have been captured from the project partners, which include cultural heritage institutions, Europeana Foundation and one Europeana accredited aggregator (i.e. Photoconsortium) - and on the technical offerings that are available from the technical partners, CYFRONET and EGI. The final product will be fully described in **Deliverable 3.3** at the end of the project (December 2024), which includes:

- The **formal requirements**, initially collected among project partners and shared to the wider community of stakeholders, CHIs and Europeana partner organisations. These requirements will be derived from a set of use cases that will define the actors and the tasks they need to perform in the system. They will be formalised in a document that will serve as an "interface" between the technical partners and the content providers partners of the project. A first version of this document will be produced during the first six months of the project and will evolve during the course of the project, as necessary.
- The connection between the technical infrastructure and the **workflow** to store the 3D assets by the **content providers**.
- The connection between the technical infrastructure and the **workflow** to serve the 3D assets to the **end users**, including the publication of items in Europeana website according to the requirements of the Europeana Data Model that is currently in a process of upgrading and expansion to better accommodate the information (data, metadata and paradata) of 3D objects.

The design and the technical implementation of the EUreka3D service and resource hub will be refined iteratively during the next ~18 months, through interviews and interactive sessions organised within the project consortium, and with the broader landscape of 3D cultural heritage preservation institutes.

1.2 Relationship to other deliverables

This document is closely related to the following deliverables:

- **Deliverable 3.2** *"The EUreka3D AAI architecture"* (Project Month 22, October 2024), which describes the infrastructure and technologies implemented to perform the authentication and authorisation of users in the EUreka3D systems.
- **Deliverable 3.3** *"Final report on the EUreka3D services and resource hub"* (Project Month 22, October 2024), which is the update of this preliminary document at the end of the project.

There is also an evident link with the intermediary technical progress reports (D1.3, 1.4, 1.5, 1.6) and the integration reports (D1.2 and 1.7), where updates on the progress of WP3 tasks, including the connection and interoperability with Europeana, are provided.

1.3 List of acronyms

Acronym	Description
AAI	Authentication and Authorisation Infrastructure
API	Application Programming Interface
CH	Cultural Heritage
CHI	Cultural Heritage Institutions
EOSC	European Open Science Cloud
GUI	Graphical User Interface
IT	Information Technology
OS	Operating System
QoS	Quality of Service
VM	Virtual Machine
VO	Virtual Organisation

1.4 Structure of the document

The rest of this document is organised as follows:

- **Section 2** briefly explains some important concepts to understand the technology layer used in the project and referenced in the rest of the document.
- **Section 3** describes the general architecture of the EUreka3D infrastructure.
- Finally, **Section** 4 provides some conclusions.

2. Concepts and technologies

This section provides an introduction to the technical concepts that are useful to understand this document. **Cloud Computing** is explained in Section 2.1, which is relevant as EUreka3D's services and resource hub are built in a cloud environment. Section 2.2 introduces generic concepts related to **data management** in the context of digital assets of cultural objects. Section 2.3 describes some of the most relevant systems for **data storage** that are currently used in the IT industry. Finally, Section 2.4 gives a generic overview of the main technology behind this service: **EGI DataHub** and Onedata.

2.1 Cloud Computing

Traditionally, the computing hardware infrastructure that supports software applications has run on independent physical servers that were purchased, installed and managed by the organisations behind such applications. This process highlighted several problems, including:

- It requires a **costly and time-consuming** procurement process, especially in highly bureaucratic organisations, and delivery and installation often take considerable time. Also, local infrastructures have an environmental impact that also generates costs.
- Servers **require expertise** to be managed and maintained.
- Servers **become outdated** over time and need to be upgraded.
- **Hardware failure** affects all computing components sooner or later, and replacing or fixing server components means unacceptable application downtime. The solution is to duplicate servers for high availability, which is more expensive and involves additional servers to maintain.

Cloud technologies are used to address these problems. They have been made possible by recent advances in computing and have become a common trend in recent decades, as organisations have transitioned from physical servers to cloud environments. Cloud-based services proved essential in the context of covid-pandemic, when many organisations were forced to boost remote working and/or online interaction with users, customers and other stakeholders, and this acted as an accelerator of the digital transformation process of the organisations in various sectors (including cultural heritage sector). Cloud technologies are based on the following principles:

- Servers and other hardware components are no longer in the organisation's premises, but are hosted by a cloud provider, who rents them for the organisation. This avoids the need to purchase hardware and allows organisations to dynamically use the exact amount of hardware they need at any given time.
- Servers are maintained by the cloud provider, so the client organisation does not have to devote effort and resources to this task.
- From the customer's perspective, servers and hardware components are nor perceived and managed as physical elements, but as virtual ones. This means that servers can be requested and served almost instantaneously and on demand, which helps to easily overcome hardware failures.

This virtual allocation of servers is done through **Virtual Machines (VM**), which are software components that run over physical hardware and emulate and provide the functionality of a physical computer system. One of the advantages of cloud technology is that the virtual infrastructure can be created directly through software instructions, without the need to physically access the servers. Moreover, this virtualisation greatly

facilitates *elasticity*, a term used to describe the ability of a system to increase or decrease its resources to adapt to the current workload.

EUreka3D uses EGI's Cloud Computing to supply its infrastructure. As described above, this infrastructure is virtual, not physical, being flexible to change according to the project needs.

2.2 Data management

Data management and storage systems have evolved significantly over the recent years, in-line with the paradigm shifts in the computing platforms. The first high-performance computations relied solely on local file-systems, but with the advent of cluster computing, local network filesystems, such as Network File System (NFS)[1] have started gaining maturity and became the basis for data intensive applications running within single data centres. With the availability of high throughput world wide networks and the need to process and share data across multiple data centres, first data management frameworks, allowing data transfers between different sites have been created, for instance Globus Toolkit data management components such as GridFTP, which required user applications to manually pre-stage data to the data centre where their computations were taking place. Finally, with the adoption of cloud computing technologies by both enterprises as well as research communities, the trend shifts toward object storages, such as AWS S3 or the Hadoop Distributed File System (HDFS)[2], without POSIX access. The lack of transparent, POSIX access to data, poses a very significant limitation on users of existing applications, which need to be adapted to the storage interfaces offered by cloud providers, or written from scratch.

In the context of Cultural Heritage Institutions some of the main challenges at the moment with regards to data management and access are as follows:

- *Missing solutions for transparent access to data in hybrid cloud environments* - applications need a transparent data access solution, enabling the users to simply deploy the application in the public cloud, and let the application access the data as if it was available locally.
- *Lack of support for legacy applications* - applications assume they can access the data as if it was available over a local file system (POSIX) on Virtual Machines or containers in the Cloud, while Cloud providers only provide object-storage or a local network file system not available outside of the Cloud infrastructure.
- *Vendor lock-in with respect to data access* - applications cannot be easily moved between different Cloud providers due to incompatible data access and management interfaces.
- *Simultaneous access to the same data over different protocols* - complex applications, composed of multiple components, some legacy relying on file system based data access, and some novel, already adapted to object storages such as S3, need to access the same data, spanning different clouds, using different types of protocols, which is not possible with existing solutions.
- *Scalability* - cloud storage, in particular from smaller and more affordable providers, often lacks scalability characteristics required by some of the large scale data processing applications, resulting in a bandwidth ceiling reached when the sum of network interfaces to the storage is saturated by a distributed application.

[1] https://en.wikipedia.org/wiki/Network_File_System
[2] https://www.ibm.com/analytics/hadoop

- *High-throughput and low latency* - applications running in multi-cloud or hybrid-cloud deployments, still require high throughput and low latency data access, however existing solutions often require cumbersome pre-staging of data, which introduces large latency and complicates application logic.

In the context of the EUreka3D project, data management is necessary to provide an abstraction of various storage systems and facilitate easy access, sharing, processing and publishing of the stored cultural heritage objects in a distributed cloud environment.

2.3 Storage systems

Modern cloud storage technologies have transformed the way data is stored and accessed. One of the most popular options is object storage, such as Amazon S3[3] or Ceph[4], which allow for the storage of large volumes of unstructured data, such as images, videos, and log files. Object storage systems organise data as objects rather than files, and they are designed to be highly scalable, durable, and cost-effective. On the other hand, POSIX-based storage systems like NFS are designed for more traditional file-based data storage. These systems are better suited for structured data, such as databases or documents, that require more complex file structures and access patterns. The main difference between object and POSIX-based storage systems lies in their architecture and approach to data management. While object storage focuses on storing and retrieving data as objects, POSIX-based storage systems organise data as files and directories, and allow for fine-grained control over access permissions and file operations. Furthermore, some object storages such as S3 do not allow modification of parts of files, only replacing the files as a whole, while POSIX storages provide random-write functionality by default.

Other common storage technologies include GlusterFS[5], OpenStack Swift[6] and WebDAV[7].

While object storages are more scalable and easier to use in cloud environments, they require that user applications can interact with them using custom protocols. This is often an issue for legacy applications, which rely on the POSIX interface for data access.

With respect to EUreka3D, and in particular to storage of Cultural Heritage objects, object storage seems as the most appropriate solution due to their scalability, ease of deployment in cloud environments as well as the fact that CH objects will be normally accessed as typically read-only.

[3] https://aws.amazon.com/s3/
[4] https://ceph.com/en/
[5] https://www.gluster.org/
[6] https://docs.openstack.org/swift/latest/
[7] https://en.wikipedia.org/wiki/WebDAV

2.4 EGI DataHub

EGI DataHub[8] is a service for provisioning large reference open data sets, based on Onedata[9] distributed virtual file-system platform, available to end users over standard POSIX, REST and Cloud Data Management Interface (CDMI)[10] interfaces.

In the scope of the EUreka3D project, EGI DataHub will enable Cultural Heritage Institutions to access a virtual data space and use storage and computing resources to manage their 3D assets in a secure and easy to use and integrate manner. Additionally, this approach is intended to be an illustrative use case (pilot) to provide an innovative cost-effective solution to data storage and the online delivery of heritage assets, by providing CHIs with a secure environment, which is technically more flexible and energy efficient.

The service is described in detail in Section 3.2

3. General architecture

This chapter presents detailed architecture of the EUreka3D services and resources hub. Section 3.1 explains the details of the EUreka3D infrastructure, how the hardware is allocated and what are the current, initial project requirements that have been allocated. Section 3.2 describes EUreka3D's resource hub, implemented via EGI DataHub, an EOSC service that enables data management in a user-friendly manner. Section 3.3 briefly describes the deployment of applications and services in the infrastructure, with a short introduction of Infrastructure Manager[11] (IM), an EOSC tool that supports the deployment of elastic and virtual clusters on top of the cloud-based infrastructure. Finally, Section 3.4 discusses some of the security aspects relevant to protect the project data.

3.1 The EUreka3D Infrastructure

3.1.1 Hardware allocation

The allocation of hardware resources is done within a cloud environment, as explained in Section 2.1. The servers used are *virtual*, allocated on demand as the project needs in a matter of minutes. This allows the infrastructure to be as agile as software components, being able to adapt to rapidly evolving environments.

The infrastructure of EUreka3D is deployed in **CYFRONET-CLOUD**, one of the resource providers of the EGI Federation, which uses OpenStack[12] as its cloud technology. Amongst other things, there are three main resources that are managed through the OpenStack platform:

- **Computing layer**. The different servers needed by EUreka3D are implemented as Virtual Machines, each of which is managed independently like a physical server.

[8] https://datahub.egi.eu
[9] https://onedata.org
[10] https://www.snia.org/cdmi
[11] https://www.grycap.upv.es/im
[12] https://www.openstack.org/

- **Storage layer**. Disks are managed as virtual devices to provide the required storage capacity, and can be attached to the different Virtual Machines as if they were physical drives connected to physical servers. In EUreka3D, the main storage allocation is managed by a service called EGI DataHub[13], described in Section 3.2.
- **Networking layer**. Network connectivity is managed and configured to allow Virtual Machines to communicate. This includes traffic routing (to move data packets from a source to a destination) and firewall rules (to block undesired network traffic).

Keeping the OpenStack platform operational is a complex process involving many challenges, and most of this burden is handled automatically by the EGI Federation resource provider, which facilitates the allocation of hardware for the project.

Figure 1 depicts an oversimplified version of this hardware allocation: There is a pool of physical hardware resources available in the cloud provider. When a user requests a server, a Virtual Machine is created by the cloud software (OpenStack in this case) and some hardware resources are assigned to it, making a virtual allocation of hardware resources. Resources can be added and removed on demand as needed, which enables a more efficient use of hardware resources.

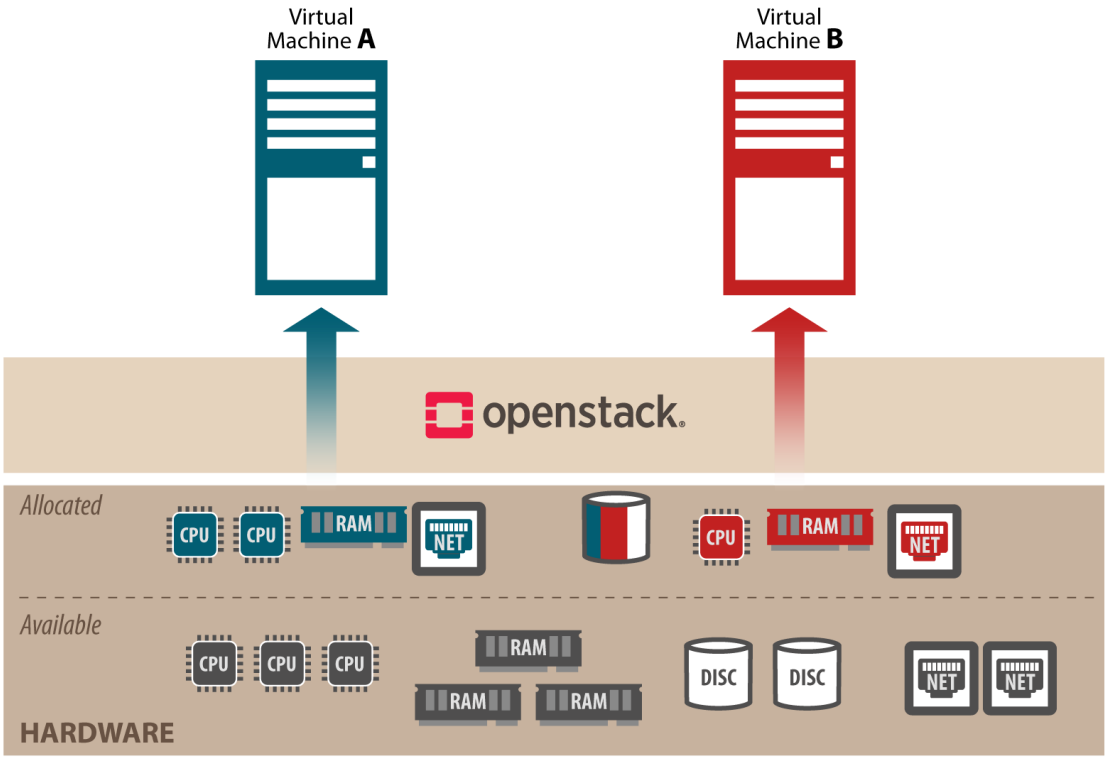

Figure 1: Simplified view of hardware allocation

[13] https://www.egi.eu/service/datahub/

3.1.2 Hardware Requirements

For the initial version of the EUreka3D platform the following is requested:

NAME	TYPE	vCPU cores	RAM	STORAGE	OS	COMMENT
Servers	2 x VM	32	128 GB	1 TB		
Data storage	S3 Object Store	-	-	20 TB	-	

3.1.3 Other infrastructure provisioning

In addition to the hardware allocation, two other elements are provisioned:

- **Domain Name System (DNS)**: The naming system for addressing machines in a human-friendly way. The EGI Dynamic DNS service[14] provides this functionality for EUreka3D.
- **Transport Layer Security (TLS) certificates**: The software artefacts containing cryptographic properties to enable secure network communication between users and applications. Certificates must be issued by a third-party that is globally trusted, and bespoke applications in EUreka3D use Let's Encrypt[15] for this.

3.2 EUreka3D resource hub

EUreka3D resource hub is based on EGI DataHub, a service of the EGI Federation implemented and managed by Cyfronet. It provides a Graphical User Interface (GUI) to access the service, as shown in Figure 2.

[14] https://docs.egi.eu/users/compute/cloud-compute/dynamic-dns
[15] https://letsencrypt.org/

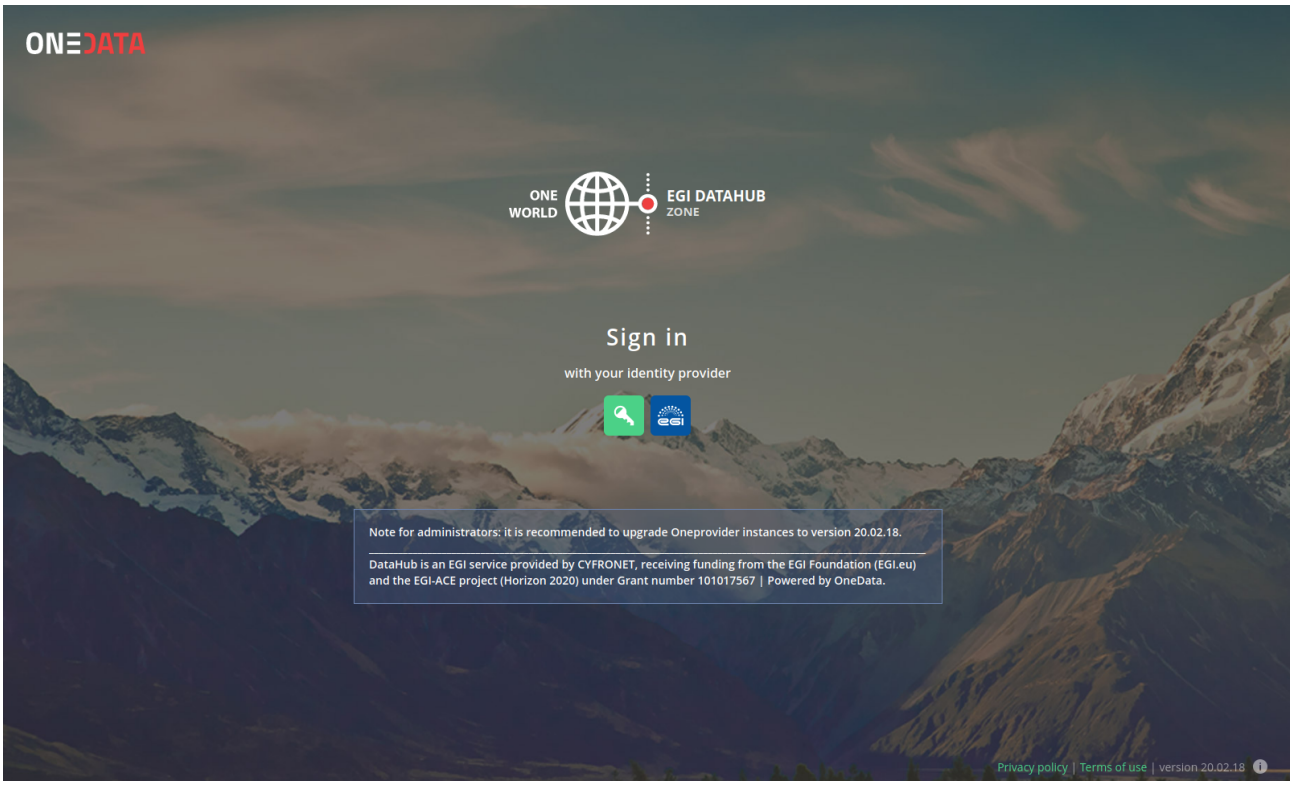

Figure 2. EGI DataHub login screen

DataHub is powered by Onedata[16], a globally distributed storage solution that integrates storage services from various providers using possibly heterogeneous underlying technologies, such as NFS or other POSIX-compliant file systems as well as Ceph[17], S3[18], GlusterFS, WebDAV, XRootD[19] and OpenStack Swift, and provides interfaces to clients based on CDMI, REST API, and virtually mounted POSIX filesystem.

DataHub utilises Onezone to federate access to various storage providers, as depicted in Figure 3, making it easier for users to manage and share their data.

[16] https://onedata.org/#/home
[17] https://docs.ceph.com/en/quincy/
[18] https://aws.amazon.com/it/s3/
[19] https://xrootd.slac.stanford.edu/

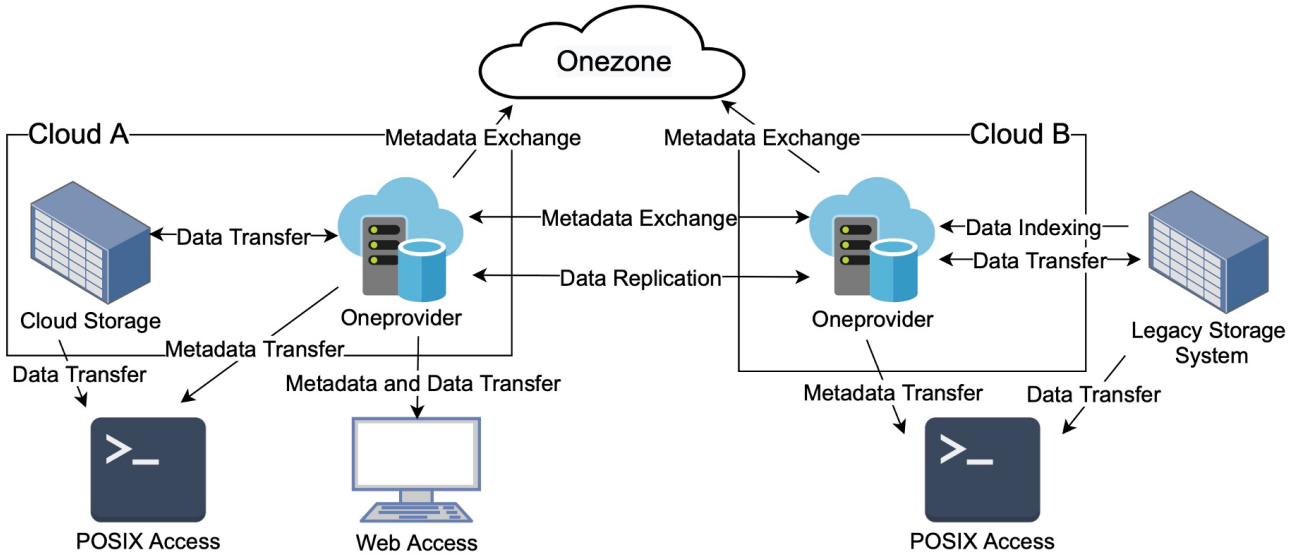

Figure 3. Overall EGI DataHub deployment diagram

Oneprovider is installed in data centres and serves as the main data management component of Onedata, as shown in Figure 4. It is responsible for provisioning the data and managing transfers. Oneprovider provides a unified interface to multiple file systems used in the data centre, and the system can scale to thousands of instances to improve performance. DataHub leverages Oneprovider to manage and provision data in a way that is transparent to the users.

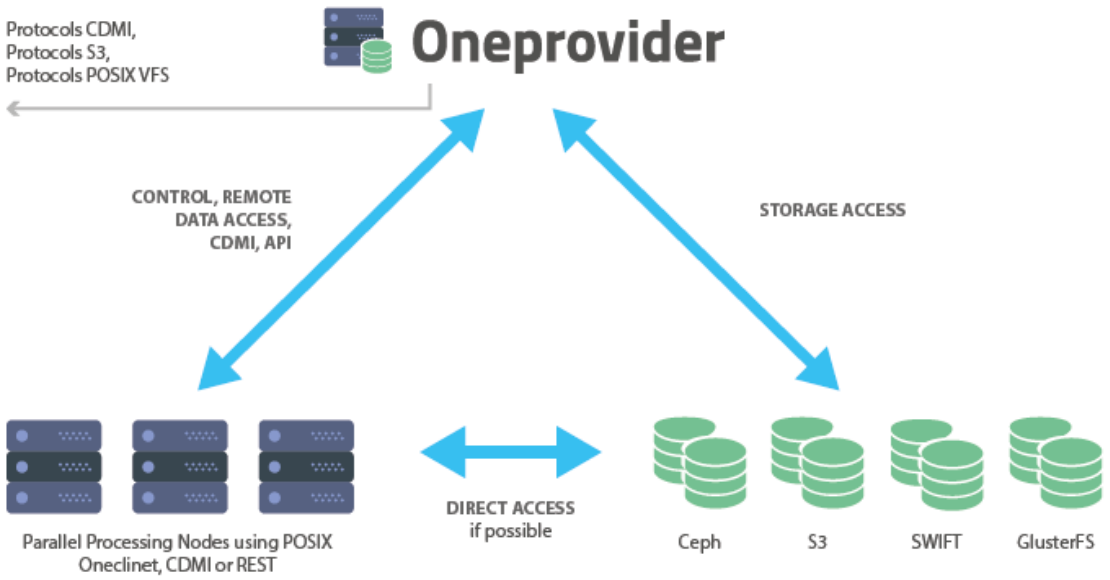

Figure 4. Deployment diagram of Onedata Oneprovider service

Oneclient is a command line tool that enables users to access the virtual filesystem on a VM or host directly via a Fuse mountpoint, as shown in Figure 5.

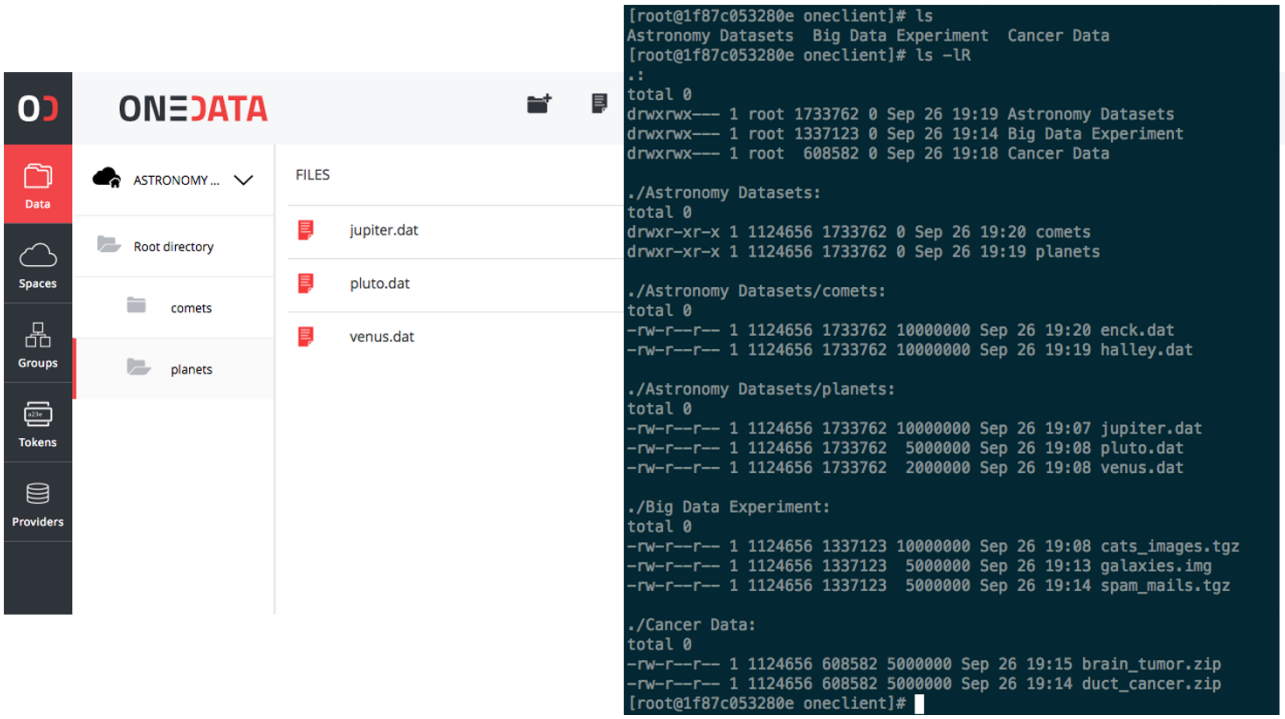

Figure 5. Example of data access using Oneclient

Data ingress and egress functionality will be achieved using the Onedata data access features. These include both automation of exposing existing data collections directly from legacy storage, automatic replication of data to the computational sites and replication of results back to the origin data centre.

Onedata provides several protocols for data access, including:

- POSIX - data can be accessed directly on any machine which can run Oneclient command line tool, and are visible as a regular POSIX filesystem to the legacy applications.
- CDMI - data can be uploaded and downloaded using HTTP based CDMI protocol, enabling easy integration with other services.
- S3 - Onedata can expose data using standard AWS S3 protocol.
- GUI - for smaller data sets, data can be also easily uploaded or downloaded using the GUI of Oneprovider.
- Data import - Oneprovider supports direct import of data from legacy storage, which can be automatically synchronised, i.e. data can be added and modified directly on the storage in the origin site, while the relevant metadata will be automatically updated and visible in the Cloud.

Additionally, Onedata has comprehensive support for metadata management, which can be accessed and modified using GUI (see Figure 6), CDMI or directly as extended attributes on the POSIX filesystem, as well as creation of custom indexes on the metadata to support efficient data discovery over large data sets.

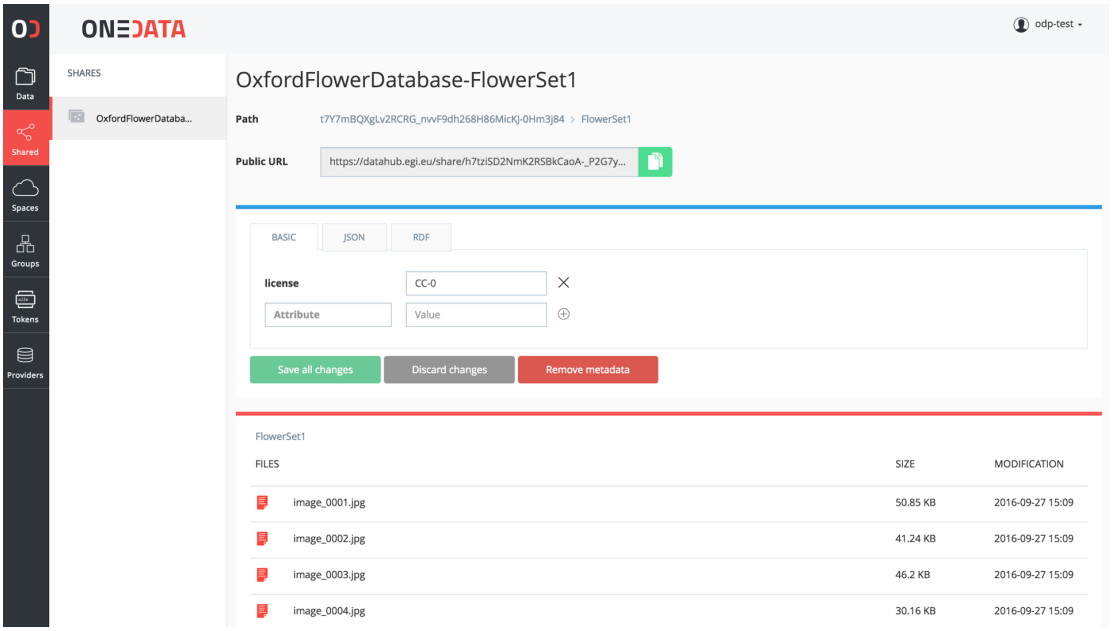

Figure 6. Onedata metadata editor

It also supports enabling OAI-PMH[20] endpoint for a selected dataset, thus allowing harvesting of the stored collocation metadata by various services. Furthermore, Onedata provides a complete replication and transfer management REST API, allowing scheduling of transfers of both single files as well as entire folders between different Oneprovider deployments. The replica invalidation can be automated easily using GUI or REST API, enabling control using low and high thresholds which trigger invalidation of replicas in a given site based on their popularity.

All Onedata components have REST APIs[21] defined using the OpenAPI specification[22], enabling easy integration and automatic generation of client libraries for most existing programming languages and frameworks. The APIs provided by Onedata include:

- **Onezone API**, to allow control and configuration of local Onezone service deployments, in particular: management of users, groups, spaces, shares, providers, services, handles and clusters.
- **Oneprovider API**, to enable access to data through CDMI- compatible endpoints, as well as data management related tasks such as data replication.
- **Onepanel API**, to allow administrators to control deployment of other Onedata components, modifying their configuration - *e.g.* adding more nodes or new storage resources.

[20] https://www.openarchives.org/pmh/
[21] https://onedata.org/#/home/api
[22] https://www.openapis.org/

For additional information, please refer to Onedata[23] and EGI DataHub[24] user documentation.

3.3 Application deployment

3.3.1 EGI Data Hub

DataHub is a service that is already implemented and deployed, and no further action is required to be used. It is also integrated with EGI Check-in[25], the solution proposed for Authentication and Authorisation.

However, DataHub will have to be configured to attend to the specific needs of the EUreka3D project, specially in what is relevant to the authorisation of users to access the data. This will be discussed in more detail in Deliverable 3.2.

3.3.2 Bespoke applications

At this current stage, it is not clear what custom applications will be used in EUreka3D, so it is not possible to provide specific details on the deployment of such applications. However, to support this activity, an EOSC service called Infrastructure Manager (IM [4]) will be used.

Infrastructure Manager is a free and open-source framework to assist users with the creation of infrastructure in a cloud environment and the deployment of software. The framework has been been developed by the Grid and High Performance Computing Group (GRyCAP)[26] at the Instituto de Instrumentación para Imagen Molecular (I3M)[27] from the Universitat Politècnica de València (UPV)[28], and it is also partially funded by the European Commission. Some of the key features of IM include:

- In a user-friendly manner, IM orchestrates the deployment of cloud resources, the installation of software packages, its configuration and the monitoring and potential update of the virtual assets created.
- It supports different cloud providers, such as EGI, AWS, Google Cloud[29] and Azure[30], amongst others. Also, it can create complex virtual infrastructures across multiple providers too.
- It uses TOSCA [5] and RADL [6] templates to define the virtual infrastructure, making it possible to manage virtual infrastructure through source code.

IM can be used from different interfaces, including the Command Line (CLI), a Web Graphical User Interface (Web GUI) and a REST Application Programming Interface (API).

In EUreka3D, IM can assist on the following tasks:

[23] https://onedata.org/#/home/documentation/doc/user_guide.html
[24] https://docs.egi.eu/users/datahub/
[25] https://www.egi.eu/service/check-in/
[26] http://www.grycap.upv.es/
[27] http://www.i3m.upv.es/
[28] http://www.upv.es/
[29] https://cloud.google.com/gcp
[30] https://azure.microsoft.com/

- Provisioning of the virtual hardware according to the required specifications and the available images.
- Installation and configuration of the Operating System based on a predefined image of the EGI cloud.
- Installation and configuration of additional software, such as a Kubernetes cluster or other applications that might be required in the project.

More information on IM can be found in its documentation[31], while more details about EUreka3D's bespoke applications will be given in the future Deliverable 3.3.

3.4 Security

Security is an important and centric part of the technology behind the "EUreka3D platform", as important as the development, implementation, user experience and performance of the services provided. Security must be carefully considered from the design phase of the project, and will be monitored and developed, where necessary, throughout the project and during the lifetime of the data. This section does not intend to present a full security assessment, but to highlight some of the security mechanisms implemented to protect the major asset of the project: the data.

3.4.1 Data in transit

As usual on the Web, data in transit are protected by using the TLS protocol, which guarantees data confidentiality (the data cannot be read by an external actor) and integrity (the data cannot be altered by an external actor without being detected) during the transit of data over a network. Figure 7 shows an example of this communication. A user connects through a Web browser to a system within the EUreka3D ecosystem, such as a bespoke Web application running on a Virtual Machine or the DataHub service. This data exchange is encrypted and protected in a TLS secure channel, with the help of the certificates used by the respective applications. To validate that these certificates are legitimate, they have to be certified by an external party, in this case Let's Encrypt. The validation process takes place in the user's environment, who is able to identify whether the certificates have been certified by a trusted third party.

[31] https://docs.egi.eu/users/compute/orchestration/im/

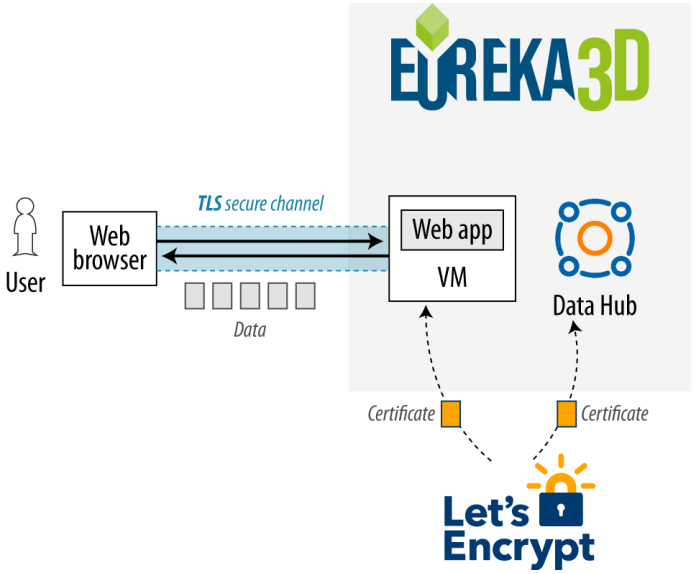

Figure 7: Encrypted data communication

3.4.2 Data at rest

Data at rest are not encrypted by EGI DataHub, so the confidentiality and integrity of these data are delegated to the security limits presented by the access to the virtual machine(s) with access to the disk where the data are stored. Users can upload data in an encrypted form, but this is not a requirement in EUreka3D and will just over complicate the system. However, due to the federated architecture of Onedata, users can explicitly specify QoS rules defining where the data can be stored (for instance by specifying geographical or network constraints), thus ensuring that data are only stored in trusted data centres.

3.4.3 Data access

Data access is protected with EGI Check-in[32], an identity management system that uses OIDC[33]/OAuth 2 (amongst other technologies) to protect Web resources. EGI Check-in acts as an *authentication proxy*[34] and as an attribute management system, delivering authorisation information to applications to enable them to make authorisation decisions.

A dedicated resource pool (alias Virtual Organisation) has been created to configure the authorisation attributes of users and to organise the community in EOSC. The resource pool allocated to the EUreka3D project is named ***culturalheritage.vo.egi.eu***[35], and is intended to be used to represent the Cultural Heritage community beyond the project.

Users authenticated in the EGI Check-in can then proceed to the EGI DataHub, where they can login using their credentials and access their storage resources through the Onedata Web interface, or generate an

[32] https://www.egi.eu/service/check-in/
[33] https://openid.net/connect/
[34] An intermediate system between the user and an organisation that authenticates the user.
[35] https://operations-portal.egi.eu/vo/view/voname/culturalheritage.vo.egi.eu

access token and use any of the supported data access mechanisms. The data access can be controlled using fine grained settings available in the web GUI or through REST API. Figure 8 presents the authorisation settings as they are shown in the GUI.

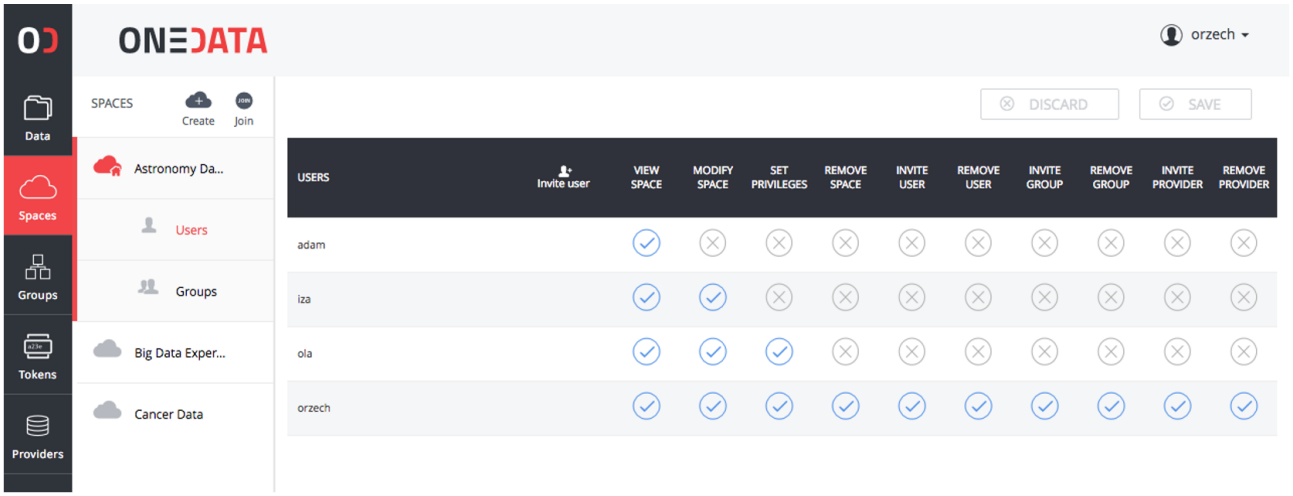

Figure 8. Onedata data access authorisation settings

For more detailed information about this infrastructure refer to upcoming **Deliverable 3.2** *"The EUreka3D AAI architecture"*.

4. Conclusions

This deliverable presented the preliminary architecture and technologies of EUreka3D services and resource hub and explained some of the concepts necessary to understand the underlying technology layer, such as Cloud Compute or data management concepts and challenges that, in the framework of the project, will contribute to implement the Data Space for the Cultural Heritage domain. The deliverable also follows the accomplishment of MS8 “The EUreka3D infrastructure is operational”, due at M3 and timely delivered on 31/3/2023.

A core component of the EUreka3D services and resource hub is based on EGI DataHub, an EGI Federation service deployed and managed by Cyfronet AGH and based on the Onedata data access platform, also developed by Cyfronet. A detailed architecture of the service has been presented, focusing on aspects relevant to EUreka3D requirements, such as data and metadata access and management, data sharing and security.

As stated before, this deliverable has described the initial state of EUreka3D’s architecture, which will be adapted and extended during the project’s lifetime according to the project needs, and whose final state will be documented in a future deliverable.

REFERENCES

[1] Commission Recommendation of 10.11.2021 on a common European data space for cultural heritage https://digital-strategy.ec.europa.eu/en/news/commission-proposes-common-european-data-space-cultural-heritage (last accessed Apr 2023)

[2] A European Collaborative Cloud for Cultural Heritage https://research-and-innovation.ec.europa.eu/research-area/social-sciences-and-humanities/cultural-heritage-and-cultural-and-creative-industries-ccis/cultural-heritage-cloud_en (last accessed Apr 2023)

[3] VIGIE 2020/654 (2020) "Study on quality in 3D digitisation of tangible cultural heritage: mapping parameters, formats, standards, benchmarks, methodologies, and guidelines" https://digital-strategy.ec.europa.eu/en/library/study-quality-3d-digitisation-tangible-cultural-heritage (last accessed Apr 2023).

[4] Caballer M., Blanquer I., Moltó G., de Alfonso C. (2015) "Dynamic Management of Virtual Infrastructures" J. Grid Comput., vol. 13, no. 1, pp. 53–70, Mar. 2015, doi: 10.1007/s10723-014-9296-5.

[5] TOSCA (2016) "TOSCA Simple Profile in YAML Version 1.0". https://docs.oasis-open.org/tosca/TOSCA-Simple-Profile-YAML/v1.0/TOSCA-Simple-Profile-YAML-v1.0.html (last accessed Apr 2023).

[6] RADL (2023) "Resource and Application Description Language (RADL) — IM Documentation 1.0 documentation". https://imdocs.readthedocs.io/en/latest/radl.html (last accessed Apr 2023).