An EU-Canada joint infrastructure
for next-generation multi-Study Heart research

Deliverable D2.2

Data distribution protocols and interfaces

Reference	D2.1_euCanSHare_BSC_30112019
Lead Beneficiary	BSC
Author(s)	Josep Ll. Gelpí, Laia Codó, Alejandro Canosa
Dissemination level	Public
Type	Report
Official Delivery Date	30th November 2020
Date of validation by the WP Leader	27/11/2020
Date of validation by the Coordinator	27/11/2020
Signature of the Coordinator	

euCanSHare is funded by the European Union's H2020 Framework under Grant Agreement 825903.

Version Log

Issue Date	Version	Involved	Comments
19th Nov	1	Josep Ll. Gelpí, Laia Codó, Alejandro Canosa	First draft
23/11/2020	2	Katharina Heil, Karim Lekadir	First comments
27/11/2020	3	Marcel Koek, Frederic Haziza, Laia Codó	Final updates
27/11/2020		Katharina Heil, Karim Lekadir	Final review
			Final version

Executive Summary

This document reviews the efforts dedicated to the development and implementation of interoperability channels among euCanSHare components, as well as the creation of uniformed connections with external consumers and services. A common OIDC-based euCanSHare-ID service is put in place to wrap in key eucanSHare platform components (**section 3.1**). A network of epidemiological databases is feeding the euCanSHare catalogue enabling multi-cohort harmonization studies, even in a federated manner (**section 3.2**). Additionally, European Genome-Phenome Archive (EGA) and Euro-Bioimaging (EuBi) are leveraged into the platform as permanent data repositories for omics and bioimaging datasets respectively (**section 2.1**). They provide efficient data distribution channels within euCanSHare by means of robust streaming REST APIs, or more complex systems like user-specific FUSE-enabled remote access to encrypted files over SFTP (**section 3.3**). Like that, euCanSHare modules as the analysis portal are able to offer researchers an integrated view of the repository's datasets, both public and protected, together with analysis tools and pipelines (**section 3.3.3**). For that, repositories' native authentication/authorization mechanisms are employed, which represents a challenge for achieving a complete integration of the system's components, and therefore, further developments are being carried out in the euCanSHare scope in order to offer a common and standardized way for accessing data (**section 4.**).

Table of Contents

1. Introduction

1.1. Motivation and strategy

euCanSHare aims to establish a unified environment for cardiovascular data sharing and analysis. euCanSHare will provide to the cardiac research community the first centralised, secure and easy-to-use platform to leverage European and Canadian cardiovascular research data and technologies, improve data discoverability, and lead to cutting-edge collaborative research in the domain of cardiovascular personalised medicine.

The initial computational framework to support the integrated euCanSHare platform was already established in the first year of the project (see Deliverable D2.1). The platform (Figure 1 https://eucanshare.bsc.es) will integrate a:

1. Data Catalogue, where relevant studies and cohorts can be identified,
2. Access Manager, in charge of managing the request for data access and the handling of the appropriate credentials
3. Data analysis and visualization portal where a number of tools and workflows will be available for the final research user.

The platform should provide means for data management, data communication among its components and with the data providers, management of data access credentials, and a single sign-on protocol for euCanSHare users authentication.

The euCanSHare platform has been conceived as a fully virtualized cloud-based infrastructure. Its components leverage different software solutions developed in other environments in order to offer an integrated user experience. The modular strategy assures that every function will be performed by specialized software giving a much more reliable solution than developing a new platform from scratch. However, this modular nature raises a number of interoperability issues as the original software products may use different protocols. We have performed an extensive analysis of the data management protocols of the modules and data providers involved in euCanSHare and devised a strategy to coordinate all of them. Deliverable D2.2 summarizes this strategy, the developed communication interfaces among components, and a roadmap for future developments. Wherever possible, standards and well-known protocols have been used to ease the compatibility with external entities, like ELIXIR (https://elixir-europe.org) or EOSC (https://www.eosc-portal.eu/).

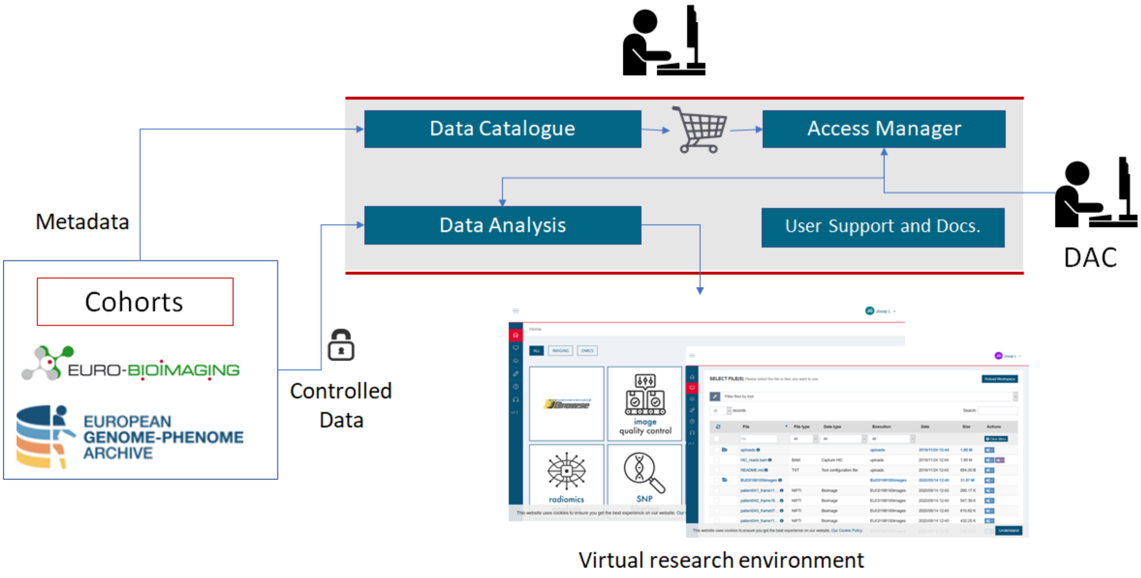


Figure 1: Overall schema of the euCanSHare data platform.

1.2. Data management requirements

Data management requirements for the platform can be summarized in the following points:

- **A single authentication and authorization procedure**. Most of the components of the platform require user authentication and authorization. We aimed to provide a Single-Sign-On procedure allowing euCanSHare registered users to access all components with a single login action. The system is compatible with other initiatives like ELIXIR-AAI.
- **Studies and cohorts' metadata available at the Data Catalogue and Data Analysis modules.** The euCanSHare Data Catalogue should contain the necessary metadata from the participating studies for users to be able to discover the available datasets. This requires communication channels for data providers to expose their metadata
- **Channel data access requests to the Access Manager.** Once the desired data variables or studies have been selected, the Access Manager component should be made aware of eventual Data Access Requests, redirect those requests to Data Access Committees (DAC's) and manage the granted access credentials to the Data Analysis modules.
- **Access to cohorts' data from the central repository**. Selected components of cohorts data under the necessary controlled access should be available in the central platform for further analysis.
- **Access to controlled data from the Data Analysis module**. The cloud-based Data Analysis platform would require accessing both public and controlled data to be incorporated into the analysis workflows. Access to controlled data will require to acknowledge the data access mechanisms of data providers and handle the appropriate credentials on behalf of the final user.

2. Infrastructure for data management

2.1. (Meta)Data providers

euCanSHare aims to offer a uniform access to datasets relevant for cardiovascular research by building an interoperability layer on top of an extensive network of multi-cohort data repositories and other consolidated data archives. The European Genome-Phenome Archive[1] (EGA) and Euro-Bioimaging[2] (EuBi) are the reference repositories for genomics and bioimaging data in the platform, while the euCanSHare network of associated research centres and biobanks host the cohort's datasets. Here we shortly introduce the particularities of these providers and their data.

Multi-cohort's data providers

Cohort's data providers comprise several Canadian and European institutions hosting multi-cohort epidemiological studies focused on cardiovascular research. Typically, each institution keeps their data protected under specific ethico-legal terms and within disconnected and geographically-dispersed local data infrastructures. euCanSHare FAIRification efforts are supported here by the implementation of a cataloguing and registration tool connected with a federated network of database infrastructures.

Cohort's data providers enhance discovery by populating the euCanSHare Data Catalogue. The metadata portal is powered by Mica[3], part of the oBiBa's software stack[4] and particularly designed for epidemiological data management, analysis and publication (see section 2.2. Data Catalogue). Additionally, and as long as the corresponding data terms-of-use allows so, data providers are also storing cohorts' datasets in their in-premises Opal[5] instances, the euCanSHare reference software for warehousing epidemiological data. The list of participants and their resources is listed in the Annex. Table I.

European Genome-Phenome Archive (EGA): the provider for Omics data

The European Genome-Phenome Archive (EGA) is a service for permanent archiving and sharing of euCanSHare Omics data and will allow a sustainable and secure data management beyond the duration of the project. CRG manages the Barcelona instance of the EGA archive. The archive's integration in euCanSHare includes the adoption of this infrastructure for depositing Omics datasets, making them accessible under a uniform authorization system covering different data access policies, and distributing them on using enhanced APIs to external applications/services such as the euCanSHare Data Analysis module (openVRE[6]). EGA distribution interfaces are discussed in section 3.3.1.

Euro-BioImaging (EuBi): imaging data provider

Euro-Bioimaging is the main European infrastructure for biological and biomedical imaging, which provides a wide range of services including the access to imaging instruments, expertise, training, and importantly, data management services. Euro-Bioimaging is managed by EMC. The euCanSHare platform aims to leverage the Euro-Bioimaging infrastructure in order to make the process of bioimaging data deposition and access easier in the scope of euCanSHare. The proposed mechanism around data deposition for euCanSHare users in the framework of the EuBi platform is thoroughly described in D3.3. Additionally, euCanSHare is currently developing brand new features for accessing imaging data from this repository, in a transparent and controlled manner, and making use of the currently available resources arising from EuBi (see section 3.3.2).

2.2. (Meta)Data consumers

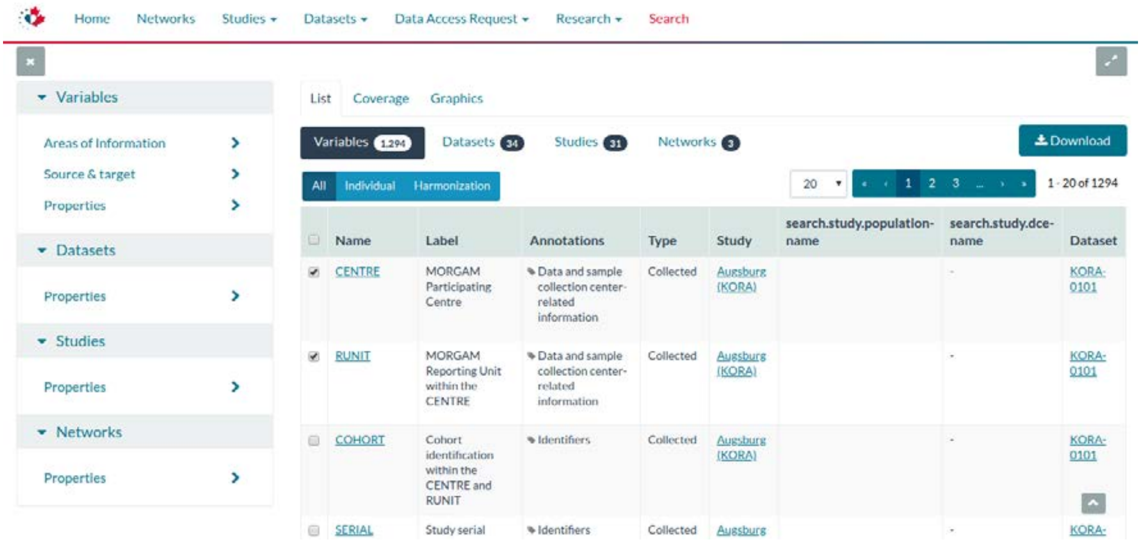
euCanSHare VRE

The euCanSHare VRE (Virtual research environment) is the data analysis portal designed to integrate reference datasets, results and analysis tools into a cloud-based workbench for cardio-vascular research. Based on a clean and bear BSC's openVRE installation (described in detail at D2.1), the euCanSHare VRE is being populated with a comprehensive set of pipelines, and currently it is featuring a growing list of bioinformatic tools (see D4.4) and a complete collection of bioimage-based analyses and pipelines (see D4.3).

In a similar way, a number of relevant reference repositories are brought into the platform implementing specific data and metadata interoperability interfaces with the above mentioned (meta)data providers (*i.e.* EGA, EuBi). The development of these VRE-pluggable data loader components as well as the resulting euCanSHare (metadata)data network is further discussed in section 3.3.

Data Catalogue

The euCanSHare Data Catalogue provides appropriate tools for cohort data management and ensures data-control mechanisms for accesses over the data deposited in the platform. oBiBa's Opal is the main tool for cohort metadata deposition, which can be configured for either working in a centralized way or a distributed manner. This component strongly depends on Mica, which consumes data from Opal, and provides an interface from which data managers can register studies linked to cohort data, and decide their degree of visibility. Finally, a cohort browser module has been deployed consisting of a web-portal built on Drupal, which represents public studies and data coming from Mica, and from where users can search, filter, and select interesting variables for their research (Figure 2).



□	Name	Label	Annotations	Type	Study	search.study.population-name	search.study.dce-name	Dataset
☑	CENTRE	MORGAM Participating Centre	Data and sample collection center-related information	Collected	Augsburg (KORA)		-	KORA-0101
☑	RUNIT	MORGAM Reporting Unit within the CENTRE	Data and sample collection center-related information	Collected	Augsburg (KORA)		-	KORA-0101
□	COHORT	Cohort identification within the CENTRE and RUNIT	Identifiers	Collected	Augsburg (KORA)		-	KORA-0101
□	SERIAL	Study serial	Identifiers	Collected	Augsburg		-	KORA-

Figure 2: The euCanSHare Cohort browser: Search, filter, and select variables data from a user interface.

Besides, the euCanSHare Data Catalogue will provide the infrastructure needed for requesting access to protected datasets, which will be tightly coupled to the Access Manager, another component of the platform.

Access Manager

The euCanSHare Access Manager (associated to D2.3, due for M36) is going to be the application responsible for centrally administering and implementing euCanSHare data control mechanisms. Currently under development, the framework will act as a mediator between several platform's elements. Firstly, it will fetch and process researchers' data petitions from the Data Catalogue and dispatch them to the corresponding cohort's data provider. The Access Manager will consult catalogue data requests through Mica's REST API (Table 1), and will deliver these to the appropriate Data Access Committee (DAC), who in turn and upon approval, will grant permissions for the protected data following the authorization protocol dictated by the data provider. Secondly, the Access Manager will handle the particulars of each (meta)data provider for offering a uniform authorization model to euCanSHare clients - researchers, clinicians, and other data consumers within the platform like the analysis portal, the euCanSHare VRE.

APIs	Short Description	URL
Mica REST API	Export cohort's associated data: studies, projects, policies, DACs	https://mica.eucanshare.bsc.es

Table 1: Programmatic access to the euCanSHare data catalogue

3. Data communication channels

The euCanSHare portal and research environment is a modular infrastructure whose components use a diverse set of communication protocols and data management technologies. We dedicate our efforts to assure a coherent user experience bundling these components under a unique authentication layer, while implementing interoperability data channels for routing data to/from remote sources and reference data repositories in a secure and efficient manner.

3.1. Integrated authentication system

The euCanSHare platform implements a Single-Sign-On scheme based on OpenIDConnect[7], which allows a common authentication framework among the different platform components (Figure 3). The platform relies on an authentication server based on Keycloak[8], which provides support for multiple identity providers, such as Elixir-AAI (*i.e.* EOSC Life ID). Currently, oBiBa's authorization server (Agate) as well as the openVRE (Data Analysis module), delegates authentication in Keycloak server.

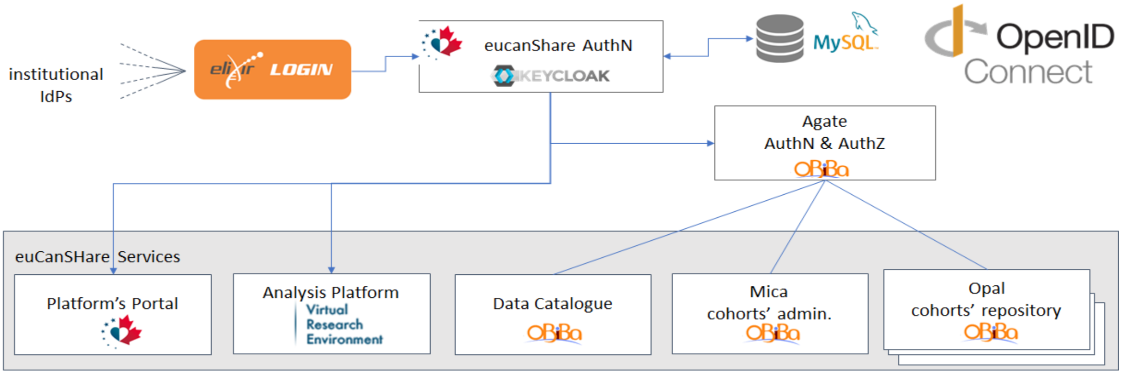


Figure 3: Data flow in the integrated euCanSHare authentication system.

European data repositories such as EGA or EuBi, currently manage their own local user list and authorization policies for data access. Nevertheless, efforts have been made for establishing a common Authentication and Authorization framework within euCanSHare's ecosystem, by implementing OpenIDConnect standards in such data repositories and also, by developing the Access Manager framework (task 2.3), which was already described in section 2.2.3.

3.2. Cohort Data transmission

oBiBa's Opal is a tool developed for facilitating data harmonization and secure data-sharing among several Opal instances. Collaborative projects are benefited when harmonizing cohorts data with Opal and Mica, as it provides mechanisms to run queries between multiple Opal instances, while ensuring data protection, as each Opal instance applies its own access rights policies on data. Opal and Mica platforms are the applications adding an interoperability layer among cohort's datasets, allowing harmonization studies and a programmatic (meta)data consumption by the epidemiological datasets, according to the user's authorization level.

Distributed cohorts data sharing

A main Opal instance has been deployed at the Barcelona Supercomputing Center, which relies on a NoSQL database (MongoDB), and that currently hosts public cohorts anonymized metadata from different data providers. Also, a distributed setup among different Opal instances have been successfully tested, enabling the connection with several cohorts data providers, such as THL and SHIP, demonstrating the system capability for handling cohorts data from sparse sources (see Annex. Table I). A secured communication among Opal instances was established via HTTPS transport layer.

3.3. Repository's data access

Efforts have been dedicated to implement specific metadata and data transfer paths to build the most convenient data network for euCanSHare. The modular and decentralized setup of the platform, and the need of handling sensitive, heterogenous and potentially heavy datasets are requirements highly conditioning the access to euCanSHare storage services. Meta(data) providers offer independent mechanisms to expose and share in a secure and controlled way their data. Here we detail the development or update of some of these protocols.

3.3.1. EGA (meta)data access

While EGA metadata is publicly available, access to their datasets is restricted. The primary mechanism for downloading them is via the Data Transfer API, developed by CRG as part of Task T3.4, that automatically encrypts all files prior to stream them. The user (or the programmatic client) receives GA4GH-encrypted[9] files that should be decrypted using a user-specific SSH key like the one specified in the original download request. Additionally, EGA offers applications to wrap around this API like an interactive shell (EgaDemoClient[10]) or the EGA Globus[11]. Regarding data transfer protocols, EGA supports HTTPS and SFTP, as well as the IBM Aspera proprietary adaptive protocol.

CRG has released EGA metadata API version 2 (Table 2) as part of task 2.2. The updated interface provides a programmatic interface for accessing a more complete set of experiment-related information, access policy codes based on the Data Use Ontology (DUO), and as usual, the metadata associated with the Omics datasets. The API follows a RESTful schema, requires no authentication, and runs over HTTPS.

RESTful APIs	Short description	BASE URL
Data Transfer API	Data streaming REST API. Around it, several clients are developed to minimize data transfers	https://ega.ebi.ac.uk/ega/rest/access/v2
Metadata API	Publicly available information of EGA studies, samples, policies, DACs and datasets	https://ega-archive.org/metadata/v2/
Submitter API	Creation, edition and deletion of EGA validation objects and projects	https://ega-archive.org/submission-api
Permissions API	REST API interface for handling user level permissions in data repositories.	In progress (task 2.3)

Table 2: List of EGA REST APIs.

Regarding EGA data distribution, the access is primarily provided via the Data Transfer API above mentioned, yet, CRG has implemented an enhanced interface for providing a more efficient and case-specific access to EGA datasets. Conveniently designed for federated and distributed platforms like euCanSHare, the partner has developed a file-system application based on a FUSE (Filesystem in UserSpace) layer able to access files without having to decrypt them (Figure 4). The Crypt4GH application[12], allows accessing the EGA-outbox remote directory of the user as a POSIX-like file system where files are transparently decrypted, never being stored on disk in an unlocked format. The eventual data transfer is on SFTP, using a Crypt4GH-compatible key, here an SSH key ed25519[18] signed. Furthermore, the euCanSHare platform is hosted at BSC's StarLife[19] cloud, the cluster also allocating Barcelona's mirror of the EGA repository. This colocality might optimize even further the data transmission between euCanSHare and EGA.

For development purposes, CRG has submitted to EGA a toy sample dataset for testing the EGA-outbox connection from outbox.ega-archive.org with euCanSHare VRE (detailed in section 3.3.3).

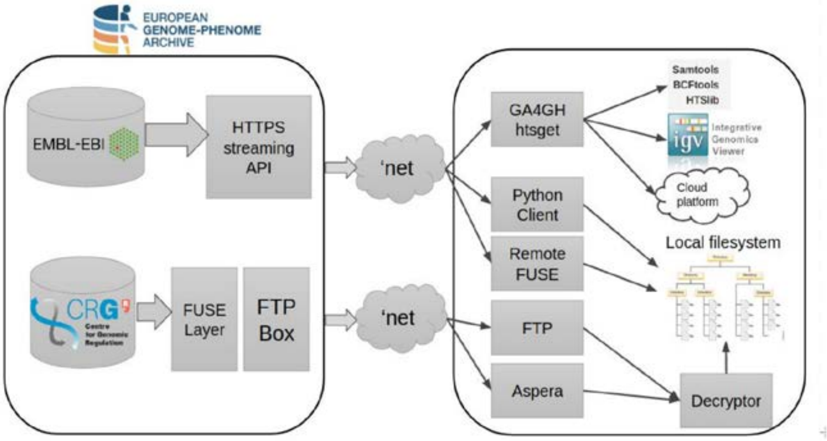


Figure 4: EGA data distribution channels into local and cloud environments.

3.3.2. XNAT (meta)data access

EMC has selected XNAT[13] as the technological solution for importing, archiving, visualizing, assessing quality, post-processing and securely distributing bioimaging data in EuBi. Programmatic interactions with XNAT are powered by a set of RESTful services, accessible by its common API (Table 3) on top of HTTPS. It facilitates searching and downloading of both, scan's data and metadata of imaging-based projects.

EuBi services are protected by a local authentication system based either on an LDAP archive or a PostgreSQL database exported on the XNAT REST interface using a Basic authentication schema. Additionally, XNAT provides a mechanism for allowing user's authentication on external applications via alias tokens. The token corresponds to a hashed combination of EuBI user ID and password accrediting the login until the token expires - 48 hours unless administrators set it otherwise. The authorization model (discussed in detail in D3.3) achieves a fine grain control of the projects and their data.

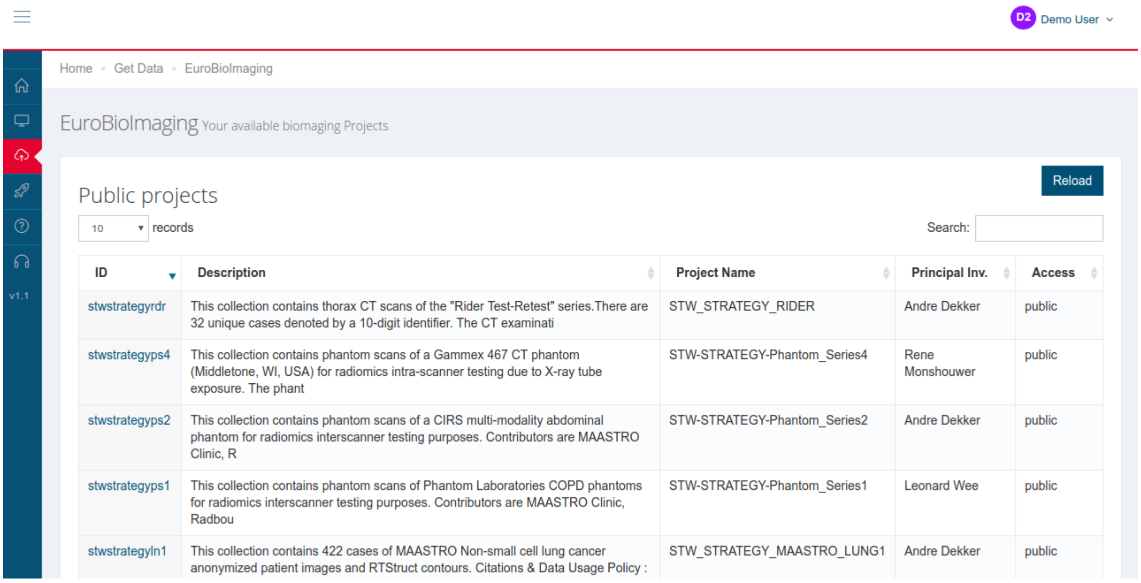
For development purposes, EMC has prepared in EuBi a private sample project for testing the access to protected EuBi scans' data from euCanSHare VRE (detailed in section 3.3.3). Furthermore, as part of the use-case led by QMUL in WP5, another protected EuBi project is being submitted for depositing bioimaging data that will eventually be imported and analysed on the euCanSHare analysis portal.

RESTful APIs	Short Description	BASE URL
XNAT API	Data streaming of imaging data in a RESTful manner. Public and private imaging associated data: projects, samples, scans, policies Creation of new imaging-based projects	https://xnat.bmia.nl/data/archive/projects/

Table 3: Euro-Bioimaging REST access.

3.3.3.Accessing (meta)data from the euCanSHare VRE

BSC, in collaboration with CRG and EMC, has prepared custom views on the VRE web interface so that with one click, registered users are able to import from euBI and EGA the particular datasets they have been granted access to. Dynamic queries to the repositories' metadata APIs feed the HTML tables listing the data. On the VRE backend, data channels have been implemented for accessing the controlled datasets by impersonating user's accounts on the original repository, either sharing identification keys with the VRE or delegating authentication into temporal identity tokens. This information is locally stored on the VRE MongoDB database as part of the user's profile data, although in the future the Access Manager will be storing and delivering interlinking authentication and authorization data.



EuroBioImaging Your available biomaging Projects

Public projects

10 records

Search:

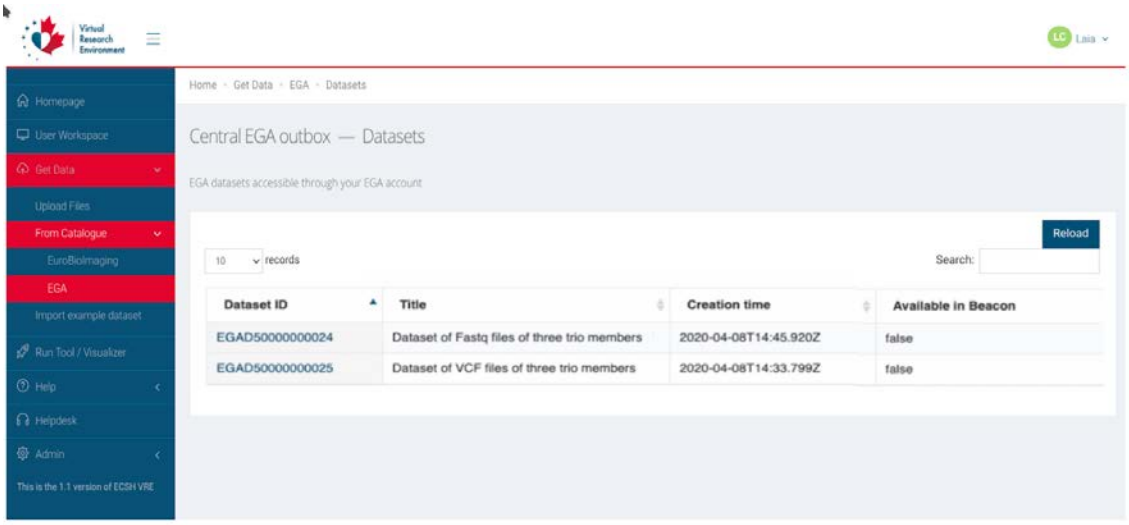
ID	Description	Project Name	Principal Inv.	Access
stwstrategyrdr	This collection contains thorax CT scans of the "Rider Test-Retest" series.There are 32 unique cases denoted by a 10-digit identifier. The CT examinati	STW_STRATEGY_RIDER	Andre Dekker	public
stwstrategyps4	This collection contains phantom scans of a Gammex 467 CT phantom (Middletone, WI, USA) for radiomics intra-scanner testing due to X-ray tube exposure. The phant	STW-STRATEGY-Phantom_Series4	Rene Monshouwer	public
stwstrategyps2	This collection contains phantom scans of a CIRS multi-modality abdominal phantom for radiomics interscanner testing purposes. Contributors are MAASTRO Clinic, R	STW-STRATEGY-Phantom_Series2	Andre Dekker	public
stwstrategyps1	This collection contains phantom scans of Phantom Laboratories COPD phantoms for radiomics interscanner testing purposes. Contributors are MAASTRO Clinic, Radbou	STW-STRATEGY-Phantom_Series1	Leonard Wee	public
stwstrategyln1	This collection contains 422 cases of MAASTRO Non-small cell lung cancer anonymized patient images and RTStruct contours. Citations & Data Usage Policy :	STW_STRATEGY_MAASTRO_LUNG1	Andre Dekker	public

Figure 5: Euro-BioImaging datasets available at euCanSHare VRE.

As shown in Figure 5, imaging scans from Euro-BioImaging are currently listed and available for its use in the euCanSHare VRE. Via REST requests to the EuBI XNAT server, the VRE connects to the repository' metadata service and builds a single-page application (SPA) that dynamically generates a searchable table. It enables an intuitive browsing among EuBi metadata structures (Project > Subject > Experiment > Imaging scan). Scans correspond to downloadable archived and compressed files containing primary and secondary imaging data in different formats (i.e. DICOM or NIFTI).

The VRE triggers a cURL-enabled transfer over HTTPS for downloading scan's data into the VRE user's workspace. The network transfer is implemented as an asynchronous job queued at the VRE backend. While importing the data, the VRE also maps EuBi's file-related metadata fields into the analysis portal. In this way, imported files are automatically annotated with information on the file format, the file content, the persistent URL from the reference repository, and other provenance metadata to trackback the outsourced data. Like so, scan data becomes eligible on the workbench and users can feed it into bioimaging analysis tools and pipelines.

The list of EuBi projects presented on the VRE tables correspond to both, publicly available datasets, together with those private projects the euCanSHare user account has access to. The authenticated access to XNAT is enabled using the user's XNAT alias token, which temporarily will permit a delegated login in a secure way. Researchers generate the token at the EuBi interface and later, they add it to the VRE user's profile. The VRE will validate and use it under a Basic authentication schema when interacting with the XNAT REST API.



Central EGA outbox — Datasets

EGA datasets accessible through your EGA account

10 records

Search:

Dataset ID	Title	Creation time	Available in Beacon
EGAD50000000024	Dataset of Fastq files of three trio members	2020-04-08T14:45.920Z	false
EGAD50000000025	Dataset of VCF files of three trio members	2020-04-08T14:33.799Z	false

Figure 6: Protected EGA datasets for a particular user available in the euCanSHare VRE.

Furthermore, EGA omics datasets have also been made accessible for its use from the euCanSHare VRE (Figure 6). Unlike EuBi, there are no open-access datasets in EGA, so VRE browsable tables are listing only the EGA data for which the user account has been granted access to. Again, the VRE is connecting to the data repository on behalf of the user, even though now, the data is only going to be transferred through the network when potentially an analysis tool requires the byte-access to the remote EGA dataset - or part of it.

Taking advantage of the EGA FUSE-based file system described above (section 3.1.1), the VRE mounts the user's EGA outbox on its local file system via SFTP in read-only mode. The VRE establishes the connection using the EGA SSH private key (and passphrase), which the user has previously stored on the VRE. As such, the VRE can transparently list user's available datasets at EGA, and display them on the HTML browsable table. Additional queries to the EGA Metadata API help displaying the information in a more comprehensive manner. When the user selects a particular file to be imported into the VRE's workspace, the file is registered as a new entry on the analysis portal and becomes eligible for any analysis tool. Yet, and unlike with EuBi data, the file is not actually downloaded to the local VRE storage but it simply points to the remote resource.

The VRE makes sure that this EGA remote device is also available to the different VRE workers where analysis tools are actually executed. But unlike the central node only willing to list and register EGA datasets, workers' file system requires to transparently decrypt the data to feed the analysis tools locally running there. For this reason, the working nodes, corresponding to independent virtual machines, make use not only of the SFTP-based EGA outbox, but also the Crypt4GH-enabled local FUSE file system.

3.4. euCanSHare as a data provider

Data produced and hosted within the euCanSHare platform should be also accessible for being consumed by other applications and services, or for being deposited to permanent archives like EGA or EuBi. The primary source of internal data generation is the analysis module, where researchers generate their results that are kept on private VRE workspaces. The BSC has implemented a REST access to these files in a programmatic manner, improving the interoperability with other euCanSHare components, like data submission portals. The API (Table 4) implements OAuth2 for authenticating the access and it is integrated under the euCanSHare SSO.

APIs	Short Description	BASE URL
openVRE REST API	- Expose metadata of input and results - Download over HTTPS VRE files	https://vre.eucanshare.bsc.es/api/v1 (API specification)

Table 4: Programmatic access to the Data Analysis module data.

4. Roadmap

The detailed features will be available in the first prototype release of euCanSHare data portal. Here are listed a number of objectives ahead of us covering distinct interconnectivity aspects of the platform development:

- Common Authentication system on the euCanSHare platform.
    - BSC will extend OIDC identity providers managed by KeyCloak to include institutions not covered by ELIXIR AAI.
    - EuBi will implement an OpenIDConnect layer on top of the OAuth 2.0 protocol for XNAT.
    - EGA as an identity provider for euCanSHare.
- Common Authorization framework.
    - EGA will develop an Access Manager and the interfaces for data consumption (Data Catalogue) and permissions delivery (DACs & Data Repositories) (task 2.3)
- Repository's data access.
    - The euCanSHare VRE will consolidate EGA access via SFTP for multiple and simultaneous users.
    - The euCanSHare VRE tools will consolidate the use of Crypt4GH and/or other on-the-fly decryption tools with the integration of genomics tools consuming EGA data.
- Access to euCanSHare analysis data.
    - BSC will align the VRE REST API[15] with the GA4GH standards by implementing a Data Repository Service[14] and/or a Task Execution Service.
- Federated Data Analysis.
    - DataSHIELD[16,17] module setup will be explored as a strategy for federated cohorts analysis on top of the network of Opal instances in the euCanSHare platform.

5. References

[1] https://www.ebi.ac.uk/ega/home

[2] https://www.eurobioimaging.eu/

[3] http://micadoc.obiba.org/en/latest/

[4] https://www.obiba.org/

[5] http://opaldoc.obiba.org/en/latest/

[6] https://vre.eucanshare.bsc.es/vre/home/

[7] https://openid.net/connect/

[8] https://www.keycloak.org/

[9] https://www.ga4gh.org/news/crypt4gh-a-secure-method-for-sharing-human-genetic-data/

[10] https://www.ebi.ac.uk/ega/about/your_EGA_account/download_streaming_client#client

[11]https://www.ebi.ac.uk/ega/about/your_EGA_account/download_streaming_client#API_wrapper_globus

[12] https://github.com/EGA-archive/crypt4ghfs

[13] https://wiki.xnat.org/documentation/the-xnat-api

[14] https://ga4gh.github.io/data-repository-service-schemas/preview/release/drs-1.0.0/docs/

[15] https://github.com/euCanSHare/VRE_data_api

[16] https://academic.oup.com/ije/article/39/5/1372/804410

[17] http://opaldoc.obiba.org/en/latest/r-user-guide/

[18] https://tools.ietf.org/html/rfc8032

[19] https://www.bsc.es/es/marenostrum/star-life

6. Abbreviations

BSC	Barcelona Supercomputing Center
CRG	Location of euCanSHare multi-cohort studies
DAC	Data Access Committee
DUO	Data Use Ontology
EGA	European Genome-Phenome Archive
EMC	Erasmus Universitair Medisch Centrum Rotterdam
FUSE	Filesystem in UserSpace
EuBi	Euro-BioImaging
MUHC	McGill University Health Centre
QMUL	Queen Mary University of London
SHIP	Study of Health in Pomerania
THL	Finish National Institute for Health and Welfare
VRE	Virtual Research Environment

7. Annex

Table I. Location of euCanSHare multi-cohort studies

Multi-cohort studies	Institution	Cohort's location	
		central	on-premises
Hamburg City Health Study UKE Clinical Cohort Studie AtheroGene StenoCardia	UKE	👥? Opal server	👥 (other)
Study of Health in Pomerania (SHIP)	Greifswald		👥 👥? Opal server
UK Biobank	UK Biobank		👥 (other)
MOnica Risk, Genetics, Archiving and Monograph (MORGAM)	THL	👥? Opal server	👥 Opal server
Canadian Alliance for Healthy Hearts & Minds (CAHHM)	MUHC		👥 👥? Opal server

Table I. Location of euCanSHare multi-cohort studies .Considering for data warehouses: the database application (Opal instance or 'other'), the location ('central Opal' at BSC or 'on-premises' installation) and type of hosted data (Anonymized data 👥? *or individual data).* 👥