euCanSHare Deliverable D3.1:

Data Management Plan v1

EGA-CRG

April 30, 2019

**Abstract**

EuCanSHare aims at harbouring the most comprehensive cardiovascular data catalogue ever assembled by integrating major cardiovascular data sources from Europe and Canada, and facilitating enhanced cross-border data sharing, discoverability and exploitation for personalized medicine research in cardiology.

Thus, FAIR data management is at the heart of euCanSHare project, as euCanSHare will store/centralize various and heterogenous data coming from a variety of cardiovascular studies/platforms, maintaining a secure and sustainable platform for sharing and analysis.

This Data Management Plan addresses the purpose and description of data handled within the euCanSHare project and the implementation of a model for data handling during and after the project, from data deposition/collection, long-term storage and security, preservation and curation, integration and interoperability, accessibility and exploitation, including standards and methodology applied, as well as compliance with FAIR and EOSC data principles.

**Table of contents**

[**1. Data summary**](#_w5rg9twn6cup) **3**

[1.1 Purpose of data collection](#_yaual9nysvfm) 3

[1.2 Description of data](#_l3xeujcyx235) 3

[1.3 Data security levels, confidentiality of potentially disclosive personal information](#_f86xjre86nqh) 5

[**2 FAIR Data management, documentation and curation**](#_d4k696yzl0sb) **5**

[2.1 Data collection / generation](#_8kb2hseskdas) 6

[2.2 Data storage](#_nadxd6nsnak) 6

[2.3 Data sharing and access](#_hgtxtrb30mly) 7

[2.4 Making data findable, including provisions for metadata](#_djnwii1wq149) 7

[2.5 Making data openly accessible](#_2brvr7y52978) 8

[2.6 Making data interoperable](#_hrh6s9c33wr) 10

[2.7 Increase data re-use (through clarifying licenses)](#_sd9ktt1aoy0o) 12

[**3 Allocation of resources**](#_yvk430t6ygse) **12**

[**4 Data security**](#_uzcs954s5pzk) **13**

[**5 Ethical aspects**](#_7zdv7ff72u89) **13**

[**6 Other issues**](#_3ptju9tv780v) **13**

# **1. Data summary**

## **1.1 Purpose of data collection**

• State the purpose of the data collection/generation

• Explain the relation to the objectives of the project

• Outline the data utility: to whom will it be useful

Q>

1-What is the purpose of the data collection/generation and its relation to the objectives of the project? 2-Will you re-use any existing data and how? If so, how will data will be re-used? Outline of protocols here

3-To whom might it be useful (’data utility’)?

Data collection is central to the general purpose of euCanSHare project, which aims at centralizing and securing cardiology medical/research data while providing a sustainable platform for cross-border data sharing and multi-cohort analysis. In the first phase of the project data from a variety of cohort studies from existing multi-cohort initiatives in health (SHIP, UK Biobank, the Hamburg City Health Study) and cardiology (MORGAM, BiomarCaRE, CAHHM) will be incorporated into euCanSHare (see Table 1 of Project Proposal).

Data collected in euCanSHare will include diverse data from prospective studies monitoring health classic and genetic risk factors as well as cardiovascular patient´s endpoint clinical and imaging data.

These studies allow assessing classic cardiovascular risk factors (including lifestyle, demographic data) and genetic risk factors from data collected over different periods of times and geographic regions. Also, for some of them, follow-up parameters from cardiovascular patients are collected including genetic, clinical and imaging data. For some cohorts, also biological samples are available for future analysis (see Table 1). This collection of data from different sources is central to the project´s goal of building the first centralized, sustainable and “FAIR” platform for cardiovascular data sharing and analysis. Also, the expertise of the different cohort owners on data collection, data management, and metadata generation will be leveraged for future cohort data incorporation in the process of building a highly comprehensive, robust, secure and scalable cardiovascular data platform.

As personalized medicine approaches are urgently needed in cardiovascular research to improve risk assessment and prevention, early diagnosis, as well as for treatment personalization and drug development, data included in euCanSHare along with their analysis available through euCanSHare’s Analysis Platform have a great potential interest for both researcher and medical staff.

## **1.2 Description of data**

Project proposal/ Each individual project contributing datasets

• Specify the origin of data generated/collected

• Specify the types and formats of data generated/collected

• State the expected size of the data (if known)

Q>

1-What is the origin of the data?

2-What types and formats of data will the project generate/collect?

3-What is the expected size of the data?

4- Data documentation

### **1.2.1 Pre-existing cohort data joining euCanSHare**

Origin>

In an initial phase, an initial reasonably diverse set of cohorts (24 cohorts) will be integrated into the catalogue and data repositories, namely:

* 21 cohorts from MORGAM Project (TLH)
* 1 cohort from BiomarCare (UKE)
* 1 cohort from UK Biobank (QMUL)
* 1 cohort from SHIP (UMG)

These consist of prospective studies from different and geographically diverse sources including diverse baseline measures and follow-up of patient ´ clinical outcomes. These include data collected from hospitals, health care institutes o administrative record databases and data contributed by volunteers or patients themselves **(Table1 and annex Table 1).**

Data types and formats, size>

Data types collected include quantitative, qualitative data generated from surveys, clinical measurements, interviews, medical records, sensor data, administrative records, experimental records such as medical or biochemical tests, genotype data, images and tissue samples (Table1, Table 2, annex Table 1 and links thereof).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Data**  **provider** | **Data type** | **Dataset**  **type** | **Datasets**  **Formats** | **Access levels** | **Size/**  **Volume** |
| MORGAM Project/THL | all:socio-demographic; lifestyle; medical history; clinical outcomes;  some also: genotypic data; biosamples | self-report questionnaires; clinical parameters measures; | metadata:PostgreSQL database; genomic data: Plink  Rest of data: csv | controlled/ metadata and aggregate statistics open | 241 variables  / |
| SHIP/UMG |  | interviews; laboratory analyses; somatometric and clinical parameters/ measurements; dental, dermatological, cardio-metabolic and various ultrasound examinations; cardiopulmonary exercise tests, sleep monitoring and whole-body magnetic imaging; bio-samples (blood, urine, tongue swap) | PostgreSQL | controlled/ metadata and aggregate statistics open | 40 TB |
| QMUL /UK Biobank | socio-demographic; biological measurements, lifestyle indicators, biomarkers and imaging | questionnaires (socio-demographic, family history, lifestyle, medical history, etc); physical measures; electronic medical records; sensor data | genomics: PLINK, VCF, CRAM files, BED, BIM, FAM, BGEN, BGI, BIN, BIM, batch, txt; physical.  activity: cwa; | registered/ metadata and aggregate statistics open | 200 TB |
| UKE/BiomarCaRE | epidemiologic | lifestyle |  | controlled |  |
| CAHHM |  | physical measures |  |  |  |
|  |  | clinical outcome |  |  |  |

Table 2 shows a more specific definition of data collected in each individual project.

**MORGAM Project/THL:**

Purpose…

General data types, cohort description and methodologies

Data is mainained in repositories blab la.

Table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **MORGAM Project/THL** | | | | |
| **Data type** | **Dataset**  **type** | **Source** | **Measures** | **Required metadata** |
| clinical outcomes | medical event registration; | self-report surveys;  administrative records (hospital discharge register; register of causes of death) | non-fatal and fatal cardiovascular events; death | diagnostic classification  (ICD coding); specific considerations; informed consent |
| genomics | genotype data SNPs array | Laboratory measurements  (Metabochip or Exome+ SNP arrays) | 200 000- 450 000 SNPs | array |
| socio-demographic | socio-demographic record | self-report surveys; administrative records | age; sex; marital status; educational achievement; | protocol, |
| lifestyle/ environmental | lifestyle/ environmental record | self-report surveys | smoking (or second hand smoking) behaviour; drug intake; awareness and treatment of high cholesterol; menopausal status; | protocol; definitions; units |
| physical measurements | physical measures; biochemical measures | self-report surveys;  hospital discharge register; laboratory tests | weight; height; waist and hip circunferences; blood preassure; metabolites levels | protocol,  conditions during measurement |
| Bio-samples | Plasma/serum, DNA, urine, saliva | medical laboratory | - | protocol,  storage conditions |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **UK Biobank** | | | | |
| **Data type** | **Dataset**  **type** | **Source** | **Measures** | **Required metadata** |
| clinical outcomes | medical event history/records; | self-report surveys;  administrative records (hospital discharge register; register of causes of death) | non-fatal and fatal cardiovascular events;  death | diagnostic classification  (ICD coding); specific considerations; informed consent |
| genomics  (1911) | genotype data SNPs arrays | Laboratory measurements  (Metabochip or Exome+ SNP arrays) | 200 000- 450 000 SNPs | array |
| socio-demographic |  | self-report surveys; administrative records | age; sex; marital status; educational achievement; | protocol, |
| lifestyle/ environmental |  | self-report surveys; | smoking (or second hand smoking) behaviour; drug intake; awareness and treatment of high cholesterol; menopausal status; | protocol; definitions; units |
|  | physical measures; biochemical measures | self-report surveys;  hospital discharge register | weight; height; waist and hip circunferences; blood preassure; metabolites levels | protocol,  conditions during measurement |
| Bio-samples | Plasma/serum, DNA, urine, saliva | - | - | protocol,  storage conditions |

**UK Biobank**

Purpose…

General data types, cohort description and methodologies

Data is mainained in repositories blab la.

Table

**SHIP**

Purpose…

General data types, cohort description and methodologies

Data is mainained in repositories blab la.

Table

**BiomarCare**

Purpose…

General data types, cohort description and methodologies

Data is mainained in repositories blab la.

Table

### **1.2.1 New cohort data joining euCanSHare**

Data types and required metadata of interest for cardiovascular research expected to enter euCanSHare will be defined on based of the initial cohort’s datasets described in 1.2.1 and the expert opinion of cardiovascular researcher participating in the project (WP).

In general, the initial cohort data **so far** includes the following data types:

* **socio-demographic** (including own and parent´s country/region of birth, country/region of residence, ethnic group, housing conditions, community safety, work settings, income, educational level),
* **lifestyle,** (including physical activity, sun exposure, smoking and alcohol consumption habits, diet, medication intake)**,**
* **environmental** (including second-hand smoking and physical environment measures such as residencial air pollution, residential noise pollution, greenspace, distance to the coast)**,**
* **medical history** (including cardiovascular events, diabetes and hypertension, family history of cardiovascular events or heart disease)**,**
* **physical measures** (including blood pressure measures, ECG, height, weight, waist and hip circumferences, cardiopulmonary exercise tests)**,**
* **clinical assays results** (including biochemical parameters, diagnostic of diabetes)**,**
* **genetic data** (including genotypes, exome-sequencing, genome sequencing)**,**
* **cardiac imaging** (heart MRI, carotid artery US)and
* **clinical outcomes** (fatal and non-fatal cardiovascular events, heart disease diagnose of heart disease or diabetes, death during follow-up).

**(Table x shows a list of the data types and their common formats in initial cohort data)**.

**Also, accepted formats of data has to be defined by data managers (WP,** 2.21.1)**.**

The definition of the data types and common formats of data sets expected to enter the project will grow on the base of this info as it is incorporated in the project and in relation to the storage and data manager infrastructures and should allow to define and serve as a guideline for submission requirements (e.g cohorts should include two or more of a set of data types of interest for cardiovascular research defined in WP) defined in a submission documentation (See submission).

Maelstrom tools will be used by data contributors later on to harmonize the diverse definitions among different projects and reach a homogenous consensus.

# **2 FAIR Data management, documentation and curation**

Angel; Jordi Rambla, Aad van der Lugt, Marcel Koek, each cohort contributor using pre-existing platform

Data management workflow is depicted in the diagram in figure 7 of Deliverable D7.1.

The data management will deal with integrating established infrastructures and technologies for secure data storage (~~local repositories of data owners and~~ central repositories in ELIXIR/EGA, BBMRI and euro-BioImaging)

and access (EGA DAC tools, Maelstrom tools), data integration (e.g. Opal, Mica), quality control (Square2), data analysis (ELIXIR, euro-BioImaging, FASTR, Rocket) for the storage and management of the heterogenous cardiac data.

This paths of data flow will include:

* Data sharing/ data collection: Submission of data, metadata and access conditions.

This involves the procedures for submitting cohort data on the centralized submission portal, the documentation on the minimal required cohort data and metadata including access conditions, the evaluation of submission requests and the transfer of submitted data to the corresponding repositories.

* Data secure storage: Encription and storage of metadata and raw data in certified repositories

This involves the procedures for securely storing in the corresponding repositories for long term re-use as well as the storage of access conditions in the ADAM/DUO infrastructure.

* Data findability/ discoverability through a centralised web-portal linking metadata across the different repositories (-omics, clinical, imaging and bio-samples) to assure the transversal consistency of the specifications, thus presenting the heterogeneous cardiac data to the users in an integrated manner.

This involves the communication through APIs of the metadata.

* Data accessibility through securely connecting the data to the relevant euCanSHare users and/or data processing environments or harmonization tools
* Data interoperability through leveraging tools for the harmonization of data
* Data reusability through leveraging infrastructure for data long-term storage and tools for the analysis of data.
* This involves the transfer of data to working platform and use of harmonization tools/ analysis tools

## **2.1 Data collection / generation**

Jordi Rambla general ideal scheme, each cohort contributor using pre-existing platform

• Specify the methodology of data deposition for user

• Define protocols for depositing new cohort raw data into the appropriate repositories and methods to provide rich metadata to foster a quality re-use of raw data for newly coming research projects.

Two phases of data incorporation are foreseen: an initial phase of incorporation of cohort data of pre-existing initiatives, described in 1.2.1 that will lead the development of infrastructure and tools and a later phase where de novo cohort data or data coming from new initiatives will be incorporated (after the data catalogue and platform is developed).

### **2.1.1 Pre-existing cohort data**

The integration of pre-existing cohort´s datasets to euCanSHare will be achieved by deposition of metadata in the centralised metadata repository in EGA and raw data in their corresponding centralised repositories. This involves the accomplishment with platforms requirements (See storage, conditions for data types and platforms), and in some cases the adaptation of platforms to fit new data types and formats and new submission procedures.

Additionally, an extensive set of metadata is being built by data owners by leveraging on Maelstrom tools that would help adapt euCanShare´s infrastructures to cardiovascular datasets by re-using MuGVRE´solutions for interoperability among storing and analysing infrastructures (See interoperability), implement data quality control and curation work and build a comprehensive data catalogue (WPs, see findability).

2.1.1.1 Process for data providers: ?????

The protocol for data deposition is yet to be defined?

The protocol for initial cohort data deposition will involve the submission of datasets directly to the corresponding platforms (See storage) as described in the sections above.

2.1.1.1.1 Submission to EGA

Data submitted to the EGA is done through a submission account on EGA.

Metadatais submitted through Submitter Portal (<https://ega-archive.org/submission/>) and should accomplish the requirements specified in EGA submission documentation linked therein. Briefly, the submission project must involve human data, provided the informed consents and have a biomedical interest and the submitter should provide this info by completing the submission statement document. The submission request is evaluated by a committee of expert curators, and after approval, submitters will be guided to comply with the specifications of data submission in EGA, including the procedures for deposition of required and additional metadata and for the download/installation and use of software for encryption of raw data. Submitters should upload the encrypted raw data to EGA through FTP clients or Aspera, enter each sample metadata through the EGA submit platform or uploaded it in the specified form as csv file, ad Associate each data file to a registered sample and study by [**Linking files to samples**](https://ega-archive.org/submission/tools/submitter-portal)

This can be alternatively done programmatically through an xml or json based programmatic submission.

Table x shows required metadata for each EGA data type.

|  |  |
| --- | --- |
| **Data type** | **Required metadata** |
| General (all data types) | Subject\_ID, BioSample\_ID, case/control status,  gender, phenotype/disease (of study) |
| Phenotypic data |  |
| Genetic data |  |
| analysis-based submission is defined below (for your BAM/BAI pairs, variation -VCF - and phenotype files) | Platform, instrument model,  Experiment type,  Pair end,  File format,  Samples link |
| un-based submission (for raw files - fastq - and aligned data - BAM/CRAM) |  |

2.1.1.1.2 Submission to euro-BioImaging

2.1.1.1.3 Submission to BBMRI

2.1.1.1 Process for infrastructures

EGA (CRG) and euro-BioImaging (EMC) will implement mechanisms for linking metadata across the different repositories, to assure the transversal consistency of the specification of studies, and to enable presenting them in an integrated way. This part of the project will realise the recently published joint-strategy of ELIXIR and euro-Bioimaging to link their infrastructures (imaging and biomolecular)17.

EMC will adapt its research infrastructure for euCanSHare to allow for standardised large-scale data storage for cardiac imaging and advanced automated image analysis. The euCanSHare infrastructure will be built around the state-of-the-art open source XNAT software (www.xnat.org), which will facilitate secure storage and

management of the cardiac imaging data. Not only raw (DICOM format) images will be stored, but also image-

derived data like segmentations and meshes, and other associated quantitative data.

### **2.1.2 New cohort data entering euCanSHare**

Process of submission by future User?

The protocols to deposit new cohort raw data into the appropriate repositories will be defined, together with the methods to provide rich metadata to foster a quality re-use of raw data for newly coming research projects.

Bio-samples will be deposed according the guidelines and quality assurance defined by BBMRI.

This task will ensure that request of information is performed according to specific formats as means to check quality control before data deposition.

New data entering euCanSHare will follow the General data deposition procedure depicted in the diagram.

A definition of minimal conditions of cohort data and required metadata for a cohort data submission request to be approved in euCanSHare (WPs) will be documented and accessible from the submission platform.

This definition will comply with the individual requirements and licence agreements of data storing platforms. These include:

* For EGA: human genotypic data (sequence, array
* For euro-BioImaging:
* For EGA: BBMRI

Data contributors will submit cohort data and metadata complying with the euCanSHare required conditions (See 1.2.1) through the centralized submission platform and euCanSHare committee will approve the submission.

Process of submission AFTER CENTRILISED PLATFORMS IS DEVELOPED – Data manager infrastructure

From there, metadata and raw data are distributed to their corresponding central repositories (See 2.2 Data storage **Table 4**), where the data will be encrypted, assigned unique identifiers, linked together with metadata and securely and permanently stored.

## **2.2 Data storage**

Jordi Rambla, Aad van der Lugt, Marcel Koek, BBMRI people, each cohort contributor storing data in pre-existing repositories

• Specify the methodology of data storage

• Data workflow diagram

Cohort data entering euCanSHare will be distributed according to their nature to be physically stored in dedicated centralised distributive repositories **Table 3**.

**Table 3**. Repositories of data and metadata in euCanSHare.

|  |  |  |
| --- | --- | --- |
| **Genral data type** | **Centralised repository** | **Data formats accepted** |
| Metadata | EGA | all formats |
| Genetic data | EGA | sequence: |
| Phenotypic (socio-demographic, lifestyle, etc) | EGA | All formats |
| Cardiac imaging | euro-BioImaging |  |
| Biosamples | BBMRI |  |

### **2.2.1 Metadata, genetic and phenotypic data workflow in EGA.**

Metadata, genetic and phenotypic data will be stored in repositories of the European Genome-phenome Archive (EGA), a service specialized in permanent archiving and sharing of all types of personally identifiable genetic and phenotypic data resulting from biomedical research projects.

CardiaC imaging data (or links to them) will be stored in a central repository based on XNAT archives and will be handled by euro-BioImaging.

### **2.2.2 Cardiac imaging data workflow in euro-BioImaging**

### **2.2.3 Biosamples data workflow in BBMRI**

Biosamples data will be stored in BBMRI.

## **2.3 Data sharing ~~and access~~**

• Each addition of a new cohort to the ensemble represents a tedious process, which discourages data exchange and enrichment.

Make more Accessible thanks to an integrated system of Data Access Committees (DACs) and new technologies based on the blockchain technology linked to state-of-the-art infrastructures (ELIXIR, EGA, euro-BioImaging, BBMRI)

• Specify if existing data is being re-used (if any)

• Specify technology for data access and transfer

• Specify methodology of data re-use (analysis)

Researcher will be able to work on a workspace from their systems without data download or software installation?? Software on platform?

(iii) APIs for securely connecting the data to the relevant euCanSHare users and to data processing environments (e.g. ELIXIR).

Additionally, EGA is collaborating with the BROAD Institute, Harvard to develop a tool similar to their Data Use Oversight System (DUOS) that is planned to be implemented a as system to pre/process applica ons before submitting them to the Data Access Committees (DAC), to improve/facilitate? current access procedure ask Jordi. New functionalities for data management in EGA are to be developed within the Excelerate project, specially for dealing with phenotypic data? Excelerate project figure 9 Deliverable D7.1

See Data security for methodology of secure data storage and transfer of sensitive data.

## **2.4 Making data findable, including provisions for metadata**

more easily **Findable** thanks to the largest and most detailed cardiovascular data catalogue ever

assembled;

metadata- Carsten Oliver Schmidt, each cohort contributor using pre-existing platform

* Outline the discoverability of data (metadata provision)
* Outline the identifiability of data and refer to standard identification mechanism. Do you make use of persistent and unique identifiers such as Digital Object Identifiers?
* Outline naming conventions used
* Outline the approach towards search keyword
* Outline the approach for clear versioning
* Specify standards for metadata creation (if any). If there are no standards in your discipline describe what type of metadata will be created and how
* Project-Describe relevant models, tools and infrastructures for (meta) data sharing and distribution within the network

Q>

1- Are the data produced and/or used in the project discoverable with metadata, identifiable and locatable by means of a standard identification mechanism (e.g. persistent and unique identifiers such as Digital Object Identifiers)?

2- What naming conventions do you follow?

3- Will search keywords be provided that optimize possibilities for re-use?

4- Do you provide clear version numbers?

5- What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

Data Browser : access to cohort data, and metadata with different security levels

Mica solution will allow network integration with external data custodians and network coordinators.

In WP2 and WP3, euCanSHare will extend and integrate the Mica solution (www.obiba.org/pages/products/mica) developed by MUC for building the largest and most comprehensive easy-to-use multi-cohort catalogue ever put together in the cardiovascular domain. This will help data custodians and network coordinators such as MORGAM, BiomarCaRE and CAHHM to efficiently organise and disseminate

## **2.5 Accesibility**

• Access request mechanisms remain highly traditional, manual and lengthy, which greatly reduce the much- needed efficiency in data-driven biomedical research.

Make more **Accessible** thanks to an integrated system of Data Access Committees (DACs) and new technologies based on the blockchain technology linked to state-of-the-art infrastructures (ELIXIR, EGA, euro-BioImaging, BBMRI)

API: Marcel Koek, DAC: Aad van der Lugt

each cohort contributor using pre-existing platform

* Specify which data will be made openly available? If some data is kept closed provide rationale for doing so
* Specify how the data will be made available
* Specify what methods or software tools are needed to access the data? Is documentation about the so ware needed to access the data included? Is it possible to include the relevant so ware (e.g. in open source code)?
* Specify where the data and associated metadata, documentation and code are deposited
* Specify how access will be provided in case there are any restrictions

Q>

1- Which data produced and/or used in the project will be made openly available as the default? If certain datasets cannot be shared (or need to be shared under restrictions), explain why, clearly separating legal and contractual reasons from voluntary restrictions.

Note that in multi -beneficiary projects it is also possible for specific beneficiaries to keep their data closed if relevant provisions are made in the consor um agreement and are in line with the reasons for opting out.

2- How will the data be made accessible (e.g. by deposition in a repository)? What methods or so ware tools are needed to access the data?

Individual data is accessible upon request. Different conditions apply to different cohort data and even data types.

For datasets coming from MORGAM Project..

For datasets coming from BiomarCare is under controlled access.

Datasets coming from UK Biobank are open to any researcher upon registration.

Datasets coming form CAHHM will be.

In all cases metadata or at least variable names and description is available.

3- Is documentation about the software needed to access the data included? Is it possible to include the relevant so ware (e.g. in open source code)?

When software or pipeline is needed as in the case of UK Biobank, it is provided.

4- Where will the data and associated metadata, documentation and code be deposited? Preference should be given to certified repositories which support open access where possible.

Data is deposited in certified repositories.

5- Have you explored appropriate arrangements with the identified repository?

6- If there are restrictions on use, how will access be provided?

7- Is there a need for a data access committee?

8- Are there well described conditions for access (i.e. a machine readable license)? How will the identity of the person accessing the data be ascertained?

For controlled data, access is provided upon approval of the concerned data access committee or data granting entity (see annex Table 1). Conditions for access of identifiable data are specified under individual project agreements. They are summarized in annex Table 1.

From proposal:

euCanSHare will implement a subportal dedicated to ease the process of acquiring access credentials according to the access rules of the respective cohorts. The task will leverage strategies already being assayed in the EGA infrastructure (e.g. DACs tool). In close collaboration with WP1 (ethics and legal analysis), the subportal will i) disseminate access policies and procedures aligned to Global Alliance for Genomics and Health Standards (DUC: Data Use Condition and ADA-M), ii) provide a simple framework to apply for data access, iii) facilitate the procedure of granting and managing granted credentials by the committees (DACs tools). For selected cohorts, the possibility of automatic credentials assignment based on applications and policies metadata will be explored through the blockchain technology in T2.5.

Three access levels> public data, registered data (need authentication), controlled data (need access permission). Rules should apply also to users registered rights (documentation) Mahsa Shabani

Data access API will be via many endpoints Marcel Koek

DAC will grant permission specific to every cohort Aad van der Lugt

Access policies (in close collaboration with WP1) will be defined and stored in the Automatable Discovery and Access Matrix (ADA-M) introduced by the GA4GH.

acces ADA-M? to ease the process of requesting access to the different cohorts and acquiring access credentials. The task will leverage strategies already being assayed in the EGA infrastructure, namely the Data Access Commi ees (DACs: www.ebi.ac.uk/ega/dacs).Through MCGILL, one of the original drivers of the Global Alliance for Genomics and Health (GA4GH), this subportal will disseminate access policies and procedures aligned to GA4GH using the Automat- able Discovery and Access Matrix (or ADA-M: www.github.com/ga4gh/ADA-M). The Access Manager component will be provided as a simple user interface for researchers to apply for data access, as well as for cohort owners to facilitate the procedure of gran ng and managing granted credentials. For selected cohorts, the possibility of automatic credentials assignment based on applica ons and policies metadata will be explored through a blockchain technology developed by our SME member LYN in the MyHealthMyData EU project (www.myhealthmydata.eu).

## **2.6 Making data interoperable**

• There are no catalogues that to enable easy access to information on available data, their precise characteristics and potential for cardiovascular research studies.

**Interoperable** by leveraging emerging solutions for data exchange, harmonisation and quality control;

include data harmonisation (Opal by MUHC: www.obiba.org/pages/products/opal), data quality control (Square2 by UMG7), cardiac data analysis (Rocket developed by UPF during VP2HF and CardioFunXion EU projects: FATSR by EMC: www.fastr.readthedocs.io) and selected bioinformatics capabilities (UKE, UMG). These tools will be made interoperable within customised pipelines through the well-established workflowing environment Galaxy (www.usegalaxy.org) and the Docker container technology (www.docker.com). euCanSHare’s data-analyser will constitute a major incentive for cohort owners to join the platform as it will provide them with state-of-the-art capabilities to harmonise and analyse data across disparate sources, in particular for those that are experts in one domain (*e.g.* -omics) but lack the tools for integrating other complimentary cardiac data (*e.g.* imaging).

, interoperability of the tools will be maintained through a rich set of metadata allowing the system to associate tools and data in a transparent manner.

euCanSHare will address interoperability issues such as cross-referencing between repositories (ELIXIR/EGA and euro-BioImaging), and the technical solutions for enabling efficient access to data using the technologies developed by ELIXIR, EGA and euro-BioImaging.

Josep Gelpi, Jordi Rambla, Aad van del Lugt, ..Isabel Fortier?

each cohort contributor using pre-existing platform

Harmonisation: Isabel Fortier

euCanSHare will centralize data from several multi-site studies in a unique web-portal that will accelerate access to information and facilitate new data harmonization.

integration:

euCanSHare also will integrate highly heterogeneous data such as -omics and cardiac imaging. To do so, euCanSHare will exploit the expertise of all consortium members, to achieve the robust integration of all of these technologies.

* Assess the interoperability of your data. Specify what data and metadata vocabularies, standards or methodologies you will follow to facilitate interoperability.
* Specify whether you will be using standard vocabulary for all data types present in your dataset, to allow interdisciplinary interoperability? If not, will you provide mapping to more commonly used ontologies?
* Project-Outline methodology for cross-referencing between repositories (ELIXIR/EGA and euro-BioImaging), and the technical solutions for enabling efficient access to data using the technologies developed by ELIXIR, EGA and euro-BioImaging.
* Project-Outline harmonization methodology

Q>

1-Are the data produced in the project interoperable, that is allowing data exchange and re-use between researchers, institutions, organisations, countries, etc. (i.e. adhering to standards for formats, as much as possible compliant with available (open) so ware applications, and in particular facilitating recombinations with different datasets from different origins)?

2-What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?

3-Will you be using standard vocabularies for all data types present in your data set, to allow inter-disciplinary interoperability?

4-In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies?

Interoperability of the tools will be maintained through a rich set of metadata allowing the system to associate tools and data in a transparent manner. Easy-to-use modules for authentication (based on KeyCloak) and secure data management (Oauth2 protocol for all encrypted data transfers) will be also integrated. The protocols to plug-in the tools (data browsers, visualisers, or analysis tools) on top of the main infrastructure will re-used from MuGVRE. Finally, execution scheduling will be based on a traditional queueing system to handle demand peaks in applications of fixed needs, and an elastic and multi-scale programming model (pyCOMPSs, controlled by the PMES scheduler) for complex workflows requiring distributed or multi-scale executions schemes. In euCanSHare, this will enable to derive tools execution to remote cloud infrastructures (through the OCCI protocol, www.occi-wg.org) and also to HPC environments within ELIXIR, EGA and euro-BioImaging.

Mica solution (www.obiba.org/pages/products/mica) developed by MUC for building the largest and most comprehensive easy to-use multi-cohort catalogue ever put together in the cardiovascular domain. This will help data custodians and network coordinators such as MORGAM, BiomarCaRE and CAHHM to efficiently organise and disseminate information about their cardiovascular studies and networks without significant technical effort.

EMC and MUHC will define metadata fields for the project, as collected by the euCanSHare cohorts and commonly used in cardiovascular research. Selection of the fields will be informed by existing standards adapted to serve the specific needs of the project. A Working Group will be established and convened at consensus meetings to define standard metadata for imaging, omics, epidemiological, clinical, and bio-sample data. EMC (euroBioImaging) will also develop and implement the models to support cardiac imaging metadata (imaging modalities, protocols, parameters and biomarkers), while MCGILL will focus on cataloguing cohort-specific access policies and consent requirements. Metadata fields will be populated with information obtained from participating cohorts (see Table 1 for the initial cohorts, while that new cohorts will be added through awareness campaigns). The final Mica-powered euCanSHare platform will include Maelstrom´s powerful search engine for allowing investigators to quickly find the information, variables and data they need for implementing cardiovascular research projects.

harmonization- from project proposal

The project will leverage on the state-of-the-art technologies developed by the Maelstrom Research at MUHC (www.maelstromresearch.org), as well as the harmonisation models implemented during the MORGAM and BiomarCaRE projects. However, since it is impossible to perform a single harmonisation that will satisfy all future study requirements, we propose an original solution to facilitate future harmonisations by re-using previous harmonisation efforts in a more systematic manner. Specifically, we will store the harmonisation algorithms in a standardised electronic database such that any harmonisation effort can be easily searched and located in the database and re-used in new multi-cohort research studies, when relevant. With this approach, future harmonisations benefit from previous ones and new harmonisa on rules/algorithms are stored to further populate euCanSHare?s harmonisa on database. This approach will reduce cost and me of future mul -cohort research studies, while providing transparency on harmonisa on processes. In this case, the harmonised dataset does not need to be stored; only the harmonisa on rules and algorithms are saved in a standardised easy- to-search format and any harmonised data is generated on euCanSHare´s cloud by the so ware in real- me. In euCanSHare, the harmonisa on database will be ini ally populated based on the BiomarCaRE and CAHHM harmoni- sa on experiences, as well as based on use cases that will be inves gated to test the pla orm. For implementing the proposed itera ve harmonisa on solu on, MUHC?s software Opal20 will be adapted to provide a centralised web- based harmonisa on management system allowing study coordinators and data managers to securely store/export a variety of data types and harmonisa on rules in different formats using a point-and-click interface. Opal includes func onali es to define variables targeted for harmonisa on, develop and implement processing algorithms used to derive common-format data, and efficiently document data harmonisation decision making. To facilitate algorithm development, Opal also includes a comprehensive JavaScript library of functions commonly used to create harmonised variables. Establishing a secure connection with an R client also allows the use of the R programming language to derive common format variables. Opal then is executed to harmonise, store and display these data under the selected standardised model (e.g. BiomarCaRE?s or CAHHM?s model). Addi onally, in euCanSHare, automated data quality control will be enabled through the Square 2 tool recently developed by UMG, which will enable to check for the level of normalisa on, ambiguity and overall quality of both new sources slated for integra on and the overall set of sources already integrated in the system. The resul ng integrated harmonisa on system will provide a highly flexible and efficient semi-automated process to on-board and harmonise new databases within the infrastructure.

## **2.7 Increase data re-use (through clarifying licenses)**

**Re-usable** thanks to open-source data analysis tools that will allow new knowledge discovery and disease quantification with validation enabled on rich and large multi-country datasets, integrating -omics, imaging, environmental and other clinical data.

## 

* Specify how the data will be licensed to permit the widest reuse possible
* Specify when the data will be made available for reuse. If applicable, specify why and for what period a data embargo is needed
* Specify whether the data produced and/or used in the project is useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why
* Describe data quality assurance processes
* Specify the length of time for which the data will remain re-usable

Q>

1-How will the data be licensed to permit the widest re-use possible?

2-When will the data be made available for reuse? If an embargo is sought to give me to publish or seek patents, specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

3-Are the data produced and/or used in the project useable by third par es, in par cular a er the end of the project? If the re-use of some data is restricted, explain why.

4-How long is it intended that the data remains re-usable? Are data quality assurance processes described?

Data included in euCanSHare will be reusable through a centralized Web Portal to the platform, which, in addition to the data repositories or their links will integrate a range of established technologies developed by our consortium members (see Table 2 of project proposal) to offer comprehensive computing infrastructures to its users, encompassing software (for cataloguing, harmonization, co-analysis and storage of study-specific and harmonized data); methods (supporting complex data integration models); high-performance computing and compute cloud (for data processing and storage); training and user support resources (e.g. guidelines, tutorials and workshops informing users on methods, resources and policy tools offered), and a central web portal (allowing secure and user-friendly access to the platform catalogues and functionalities). Figure 2 of project proposal represents the main components and functionalities of the platform.

In general, data collected/centralized in euCanSHare will become accessible through a centralized Web Portal and interoperable interfaces integrating the tools from Table 2 in project proposal. This will be done by re-using the basis framework and IT solutions developed by BSC during the MuG project. Some data from some cohorts are restricted though. How to come about this. Every contributor should answer which part is restricted?

Finally, execution scheduling will be based on a traditional queueing system to handle demand peaks in applications of fixed needs, and an elastic and multi-scale programming model (pyCOMPSs, controlled by the PMES scheduler) for complex workflows requiring distributed or multi-scale executions schemes. In euCanSHare, this will enable to derive tools execution to remote cloud infrastructures (through the OCCI protocol, www.occi-wg.org) and also to HPC environments within ELIXIR, EGA and euro-BioImaging.

# **3 Allocation of resources**

• Estimate the costs for making your data FAIR. Describe how you intend to cover these costs.

• Clearly identify responsibilities for data management in your project.

• Describe costs and potential value of long-term preservation

Q>

1- What are the costs for making data FAIR in your project?

2- How will these be covered? Note that costs related to open access to research data are eligible as 3- part of the Horizon 2020 grant (if compliant with the Grant Agreement condi ons).

4- Who will be responsible for data management in your project?

5- Are the resources for long term preserva on discussed (costs and potential value, who decides and how what data will be kept and for how long)?

# **4 Data security**

**A secure communication interface**, which will integrate state-of-the-art protocols and technologies for ensuring interoperability and communication between components, secure data distribution, and cloud- based execution of tools. It will include easy-to-use modules for authentication (based on KeyCloak) and secure data management (Oauth2 protocol for all encrypted data transfers)

Easy-to-use modules for authentication (based on KeyCloak) and secure data management (Oauth2 protocol for all encrypted data transfers) will be also integrated.

• Address data recovery as well as secure storage and transfer of sensitive data

Q>

1- What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data)?

2- Is the data safely stored in certified repositories for long-term preservation and curation?

# **5 Ethical aspects**

• To be covered in the context of the ethics review, ethics sec on of DoA and ethics deliverables. Include references and related technical aspects if not covered by the former

Q>

Are there any ethical or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Descrip on of the Ac on (DoA).

Is informed consent for data sharing and long term preserva on included in ques onnaires dealing with personal data?

# **6 Other issues**

• Refer to other national/funder/sectorial/departmental procedures for data management that you are using (if any)

Q>

Do you make use of other na onal/funder/sectorial/departmental procedures for data management? If yes, which ones?