


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

Deliverable D4.1:

Opal software integration to euCanSHare

Reference	D4.1_euCanSHare_MUHC_28112019
Lead Beneficiary	MUHC
Author(s)	Isabel Fortier, Sofiya Koleva, Carsten-Oliver Schmidt, Jordi Rambla De Argila, Kari Kuulasmaa, Ari Haukijärvi, Teemu Niiranen
Dissemination level	Public
Type	Data infrastructure, software
Official Delivery Date	November 30 th 2019
Date of validation by the WP Leader	November 27 th 2019
Date of validation by the Coordinator	November 28 th 2019
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
16/11/2019	V1	Isabel Fortier	1st draft shared with WP partners, PC and PM
18/11/2019		Katharina Heil (UB)	1 st review by PM
26/11/2019	V2	Karim Lekadir (UB)	Review by PC
27/11/2019		Isabel Fortier and team, Carsten-Oliver Schmidt, Kari Kuulasmaa and team	Additional input and review
28/11/2019	Final	Katharina Heil, Karim Lekadir (UB)	Final check and updates

Executive Summary

In order to support the data harmonization process to be achieved within euCanSHare as well as proper documentation of the harmonized datasets generated, it is essential to implement a data and metadata documentation, processing and management system. The OBiBa software suite (Opal, Mica and Agate) and harmonization and cataloguing resources (harmonization guidelines and metadata standards) developed by Maelstrom Research are used as key elements of the EuCanSHare system. OBiBa software infrastructures are implemented in Spain, Finland, Germany and Canada to form the EuCanSHare harmonization platform. The platform will be pilot tested in 2019-2020 and, where required, the software will be customized to serve the evolving needs of EuCanSHare. The deliverable is a software and this report is the written description and presentation of the software and its implementation in different environments to support EuCanSHare activities.



Table of Contents

Table of Contents	3
1 General objectives and rationale	4
2 Maelstrom Research software and methodological approach to harmonization	4
2.1 OBiBa Software Suite	4
2.2 Maelstrom Research guidelines and standards	6
3 Management infrastructure.....	7
3.1 Implementation of the Obiba software stack at central EuCanSHare servers	7
3.2 Implementation of the Obiba software stack at THL.....	8
3.3 Implementation of the Obiba software stack at University of Medicine in Greifswald ..	9

Acronyms

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

OBiBa: Open Source Software for Epidemiology

MUHC: Research Institute of the McGill University Health Center

SPSS: Statistical Package for the Social Science

CSV: Comma-Separated Values - a delimited text file that uses a comma to separate values

SQL: Structured Query Language - is a domain-specific language used in programming

UID: Unique Participant Identifier

MORGAM: MONICA Risk, Genetics, Archiving and Monograph Network

BiomarCaRE: Biomarker for Cardiovascular Risk Assessment across Europe

SHIP: Study of Health in Pomerania

BSC: Barcelona Supercomputing Center

THL: Finnish Institute for Health and Welfare



1 General objectives and rationale

Harmonizing data across cohort studies is a complex process requiring use of rigorous methodological approaches and access to secure and effective software infrastructures. The OBiBa software suite (Opal, Mica and Agate) developed by Maelstrom Research was selected to serve as a key element of the EuCanSHare harmonization platform. The objective of the Deliverable 4.1. was to implement the suite of software in Spain, but also in Finland and Germany (the suite was already implemented in Canada) to form the EuCanSHare harmonization platform. The platform will be pilot tested using a methodological approach to data harmonization and quality control combining processes used by the [Maelstrom Research](#), [MORGAMv](#), [BiomarCaRE](#) and [SHIP](#) initiatives, and customized to serve the evolving needs of euCanSHare.

2 Maelstrom Research software and methodological approach to harmonization

The OBiBa suite of software (Opal, Mica and Agate) and the methodological approaches developed by Maelstrom Research are used as key elements of the EuCanSHare harmonization infrastructure.

2.1 OBiBa Software Suite

The OBiBa suite consists of 3 main software: Opal, Mica and Agate.

Opal supports data management and harmonization. It provides a centralized web-based data management system allowing study coordinators and data managers to securely

import/export a variety of data types (e.g. text, numerical, geolocation, images, videos) and formats (e.g. SPSS, CSV) using a point-and-click interface. Opal then converts, stores and displays these data under a standardized model. Opal can also read data directly from various data sources engines such as LimeSurvey or SQL databases. Once data have been imported, the Opal web application facilitates data curation and quality control procedures and allows basic descriptive statistics computation with graphical displays such as bar charts and scatter plots. The application also allows annotating variables with metadata to create searchable data dictionaries. For example, using the variable classification taxonomy developed by Maelstrom Research, each data item collected by a study can be annotated with a standard list of domains of information. This then facilitates metadata browsing and data discoverability using the Mica web portal. Thanks to its integration with the R software and R markdown, statistical analyses and reports on data stored in Opal can also be performed. To ensure privacy, Opal stores participant

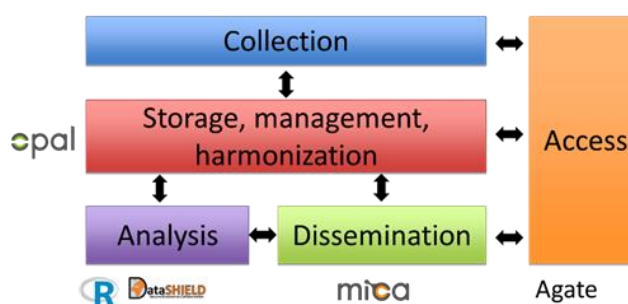


Figure 1: OBiBa software suite



identifiers in a distinct and secure database and provides administrators with tools to manage them. For example, to de-identify data disseminated to researchers, Opal can create new sets of unique participant identifiers (UIDs) when exporting datasets. Since mapping of UIDs remains in Opal, disseminated UIDs can be re-linked to internal UIDs at import. When used across multiple studies, Opal is an efficient tool to harmonize data. As such, it includes functionalities to: define variables targeted for harmonization, develop and implement processing algorithms used to derive common-format data, and efficiently document data harmonization decision making. Establishing a secure connection with an R client allows use of the R programming language to derive common format variables.

Mica is a metadata catalogue and data discovery tool. It is used to create websites and metadata portals for individual epidemiological studies or multi-study consortia, with a specific focus on supporting observational cohort studies. The Mica application helps to organize and disseminate information about studies and networks without significant technical effort. Mica is made up of a number of different modules developed to add and edit descriptive information pertaining to epidemiological research networks, studies and datasets. The network description module allows Mica users to disseminate information such as network description and the number, design and geographical coverage of participating studies. A study description module is used to assemble and publish information on epidemiological studies such as participant selection criteria, sample size and data collection timelines. A dataset module supports the display and dynamic search of study data dictionaries (i.e. codebooks). For a consortium of studies, a special dataset module allows documenting and presenting variables targeted for harmonization and harmonization results. All Mica modules are customizable, i.e. information fields can be added or modified to meet the specific needs of a study or research network. Further, although the software is preconfigured for English or French, its user interfaces can be translated in any International Organization for Standardization language without the need to modify the code. Mica also supports the creation of multilingual websites. Once populated with study and variable metadata, Mica includes a search engine which allows investigators to quickly find the information they need for their research projects. For example, users can quickly identify a list of studies with a given profile (e.g. cohorts recruiting middle aged participants), which collect data on a specific health outcome (e.g. stroke), risk factors of interest (e.g. physical activity) and confounding factors (e.g. age, sex, income, education, work status). Connecting Mica to one or more Opal database(s) allows users to search beyond the metadata by securely querying the actual study data hosted on remote servers.

Agate is used to manage users. Agate is a lightweight but highly secure central authentication server that offers user related services to the Opal and Mica (server and Drupal) applications: user authentication, user profile management, user notifications. Opal and Mica can use Agate as their user registry and user notification service.

Mica and Opal users belong to groups with different levels of application access managed by Agate.

The existing suite of software needed to be improved/customized to serve the needs of EuCanSHare. A first round of customization was achieved before implementation of the

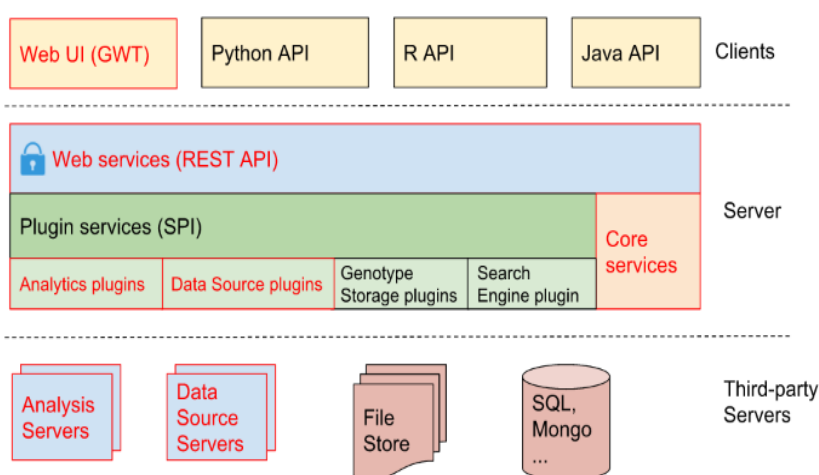


Figure 2: Opal software architecture

infrastructure (see section 3 below). Current functionalities will be tested in 2020 and could lead, where required, to additional customization before the end of the euCanSHare project. Figure 2 displays (in red) features already added to Opal layers.

The improved applications minimize the time and technological efforts to be dedicated achieving data transfer, documentation, cleaning, processing, dissemination and analysis. Some examples of improvement are listed below:

- Opal software extensibility was improved by building a plugin architecture - migrate data source connectors from Opal's core to its plugin layer. This will allow software developers from euCanSHare to create and utilize data source plugins. For example, THL will be using Mica Python API or R API to avoid manually importing data to their Mica server.
- Opal was better integrated to R studio to facilitate data documentation and harmonization and allow researchers to create tibbles (data frame) of data and data dictionary in R studio, merge them, and push the data to Opal.

2.2 Maelstrom Research guidelines and standards

Maelstrom Research developed procedures and standards to guide users in the development of high-quality data harmonization and metadata cataloguing¹². Existing procedures developed by Maelstrom will be aligned with procedures used by the MORGAM, BiomarCaRE and SHIP initiatives to serve the needs of EuCanSHare. These procedures will guide usage of the euCanSHare harmonization platform. For each harmonization project to be achieved procedure will include:

¹ Bergeron J, Doiron D, Marcon Y, Ferretti V, Fortier I. Fostering population-based cohort data discovery: The Maelstrom Research cataloguing toolkit. PLoS ONE 13(7)

² Fortier I, Raina P, Van den Heuvel ER, Griffith LE, Craig C, Saliba M, Doiron D, Stolk RP, Knoppers BM, Ferretti V, Granda P, Burton P. Maelstrom Research guidelines for rigorous retrospective data harmonization. Int J Epidemiol 2017; 46 (1): 103-105



Step 1: Assembling information using the euCanSHare study catalogue

Ensure appropriate knowledge and understanding of each targeted study design and data content.

Step 2: Defining core variables to be generated and evaluate harmonization potential

Outline the list of core variables to be generated across studies to answer the research questions addressed and determine the possibility for each study to construct each core variable.

Step 3: Process data using R and Opal

Generate study-specific harmonized dataset(s) to be used for data analysis.

Step 4: Estimate quality of the harmonized dataset(s) generated

Understand characteristics and utility of the harmonized dataset(s) generated to ensure adequate usage.

Step 5: Disseminate/preserve final harmonization products using Mica, R and Opal

Generate documentation required to support analysis and provide access to harmonized data to EuCanSHare investigators

3 Management infrastructure

3.1 Implementation of the Obiba software stack at central EuCanSHare servers

EuCanSHare central servers at the Barcelona Supercomputing Center (BSC) are being set up to provide the main euCanSHare project services. They are organized on top of a cloud-based infrastructure that provides data storage and computational framework for data management and analysis.

One of the projected EuCanSHare services is a comprehensive data catalogue including cohort data, genomic data hosted at the European Genome-Phenome Archive (<http://ega-archive.org>) and image data provided by Euro-BioImaging (<http://www.eurobioimaging.eu/>). Support for cohorts' data management within the catalogue is being based on the Obiba software stack. BSC have deployed (Figure) servers for Agate (<http://agate.euCanSHare.bsc.es>), Opal (<http://opal.euCanSHare.bsc.es>), and Mica (<http://mica.euCanSHare.bsc.es>), and a provisional front end based on Mica Drupal client (<http://studies.euCanSHare.bsc.es>). Agate server will be connected with the EuCanSHare central authentication system to provide a single-sign-on system for the project and be able to accept identity providers like the forthcoming European Life Sciences ID systems. Opal server at BSC will host EuCanSHare public datasets, while controlled access datasets will be declared at central Mica server but kept in distributed Opal servers (data already available from MORGAM, and SHIP cohorts).

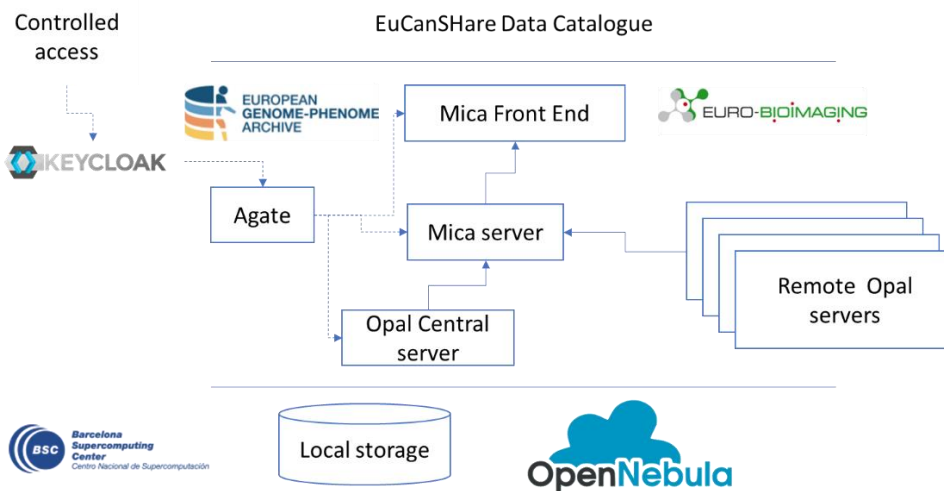


Figure 3: Schema of EuCanShare data central servers (Solid: Data channels. Dotted: Authentication channels)

3.2 Implementation of the Obiba software stack at THL

In Finland, THL are currently running the full OBiBa software suit on their own server. The THL Opal holds all MORGAM variable definitions (metadata) as well as all individual level survey data. THL imports the data on MORGAM studies and cohorts to the BSC Mica server and BSC Mica can get the variable data from THL Opal server if allowed by the data access policy and contracts. (See Figure3: remote Opal servers.) THL also produce "dummy" (anonymous) data where the value of the variable is binary - available or unavailable for MORGAM cohorts in order to make some statistics publicly available for the BSC Mica. This anonymous data is placed at Opal server at BSC.

The scientific benefit of this configuration is to increase MORGAM data visibility. The available data will be described via BSC system in detail without need for new and complex data sharing agreements from a large number of MORGAM participating centres which are the data owners.

In addition, THL imports the data on MORGAM studies and cohorts in THL Mica server. This is done because THL's Mica installation allows to show statistics from the real survey data residing in THL Opal. Data imports and updates in both BSC and THL Opal and Mica servers will be implemented using the R API built by Maelstrom Research. This will eliminate the need of unnecessary manual and/or duplicate work regarding data imports and updates.



3.3 Implementation of the Obiba software stack at University of Medicine in Greifswald

The team of the SHIP Study, conducted by the University of Medicine in Greifswald in Germany, has installed the full OBiBa software stack on three behind-the-firewall servers on the university network in order to create a cohort browsing application demonstrator for the euCanSHare project. The population-based cohort study investigators are using their own Opal repositories. Since the servers have no direct access to the Internet due to local security guidelines in Greifswald all updates will be done manually. Maelstrom research is working closely with the SHIP IT team to ensure smooth systems setup and continuous integration with future software releases.

In addition, the SHIP study team and the Maelstrom Research Harmonisation team will be enhancing the OBiBa software stack with a new feature that will allow the use of dummy variables. In this feature the collected data will be masked with binary values - true and false, where true will represent the presence of collected data and false – the lack of such. This will facilitate investigators and allow them to request data access only to relevant patient data.