**Proof of Concept of Model-based Cardiovascular Prediction**

**Call identifier:** FP7-ICT-2013-10 - **Grant agreement No.** 611232
**Thematic Priority**: ICT-2013.5.2: Virtual Physiological Human

# Deliverable 3.1

# Data Infrastructure and Information System (DIIS) deployment and configuration report

Due date of delivery: 03/10/2014

Actual submission date: 24/11/2014

Start of the project: 1st October 2013
**Ending Date**: 30th September 2016

**Start of the project:** 1st October 2013
**Ending Date**: 30th September 2016

Partner responsible for this deliverable: P9 - GNÚBILA

Version: 1.0.1

**Dissemination Level: Public**

**Document Classification**

| | |
|---|---|
| Title | Data Infrastructure and Information System (DIIS) deployment and configuration report |
| Deliverable | D3.1 |
| Reporting Period | M12 |
| Authors | Sébastien Gaspard<br>Jérome Revillard<br>David Manset |
| Work Package | WP3 |
| Security | Public |
| Nature | Document |
| Keyword(s) | Infrastructure, FedEHR |

**Document History**

| Name | Remark | Version | Date |
|---|---|---|---|
| Sébastien Gaspard | | 0.1 | 09/10/2014 |
| Sébastien Gaspard | | 0.2 | 14/10/2014 |
| Jérome Revillard | | 0.3 | 16/10/2014 |
| David Manset | | 0.4 | 17/10/2014 |
| Jérome Revillard | | 1.0 | 17/10/2014 |
| Sébastien Gaspard | | 1.0.1 | 27/10/2014 |

**List of Contributors**

| Name | Affiliation |
|---|---|
| Sébastien Gaspard | GNÚBILA |
| David Manset | GNÚBILA |
| Jérome Revillard | GNÚBILA |

**List of reviewers**

| Name | Affiliation |
|---|---|
| Ashley Waring | Fraunhofer MEVIS |
| Edwin Morley-Fletcher | LYNKEUS |

**DISCLAIMER:** The present document partially builds upon the relevant work performed within the MD-Paedigree project, for the implementation of the Infostructure which CARDIOPROOF is linking to and working with. This approach is detailed within WP3, where it is stated, among other things referring to MD-Paedigree, that "Starting from the platform available in MD-Paedigree, GNÚBILA will adapt it to fit the identified and prioritized CARDIOPROOF requirements".
In the PART B of the DoW, a specific section ("B1.2.3.1 Making use of the MD-Paedigree e-health platform") is dedicated to explaining such integration with MD-Paedigree. In this section it is clearly stated that: "Through MD-Paedigree, CARDIOPROOF can thus benefit from a highly advanced repository and Science Gateway from its inception, thanks to the available ground-truth infrastructure at OPBG […]".
The excerpts within this deliverable which are derived from the work already performed in MD-Paedigree, as well as those parts of the work which are updates and CARDIOPROOF adaptations of MD-Paedigree's implementation, are highlighted in the text and can be easily identified by reading the titles of the relevant sections.

# 1 Project summary

CARDIOPROOF aims to overcome knowledge and technology gaps on validity, usability, and comparative effectiveness by applying advanced VPH modelling methods in clinical trials of patients with aortic valve disease (AVD) and aortic coarctation (CoA). These two diseases, which both affect the linked system of the left ventricle (LV), aortic valve and the aorta, are of particular interest for CV modelling approaches.

Using the modelling methods developed in previous work, the primary objectives of CARDIOPROOF are to conduct validation trials in patients with AVD or CoA that reflect a *real-world approach* by covering and comparing the complete spectrum of cardiovascular treatment, provide first data about comparative clinical and cost effectiveness of VPH approaches compared to conventional diagnostic and treatment algorithms, and accelerate the deployment of CV VPH methods by improving their usability and interoperability in the clinical context. In addition to the validation of disease modelling, CARDIOPROOF will have access to an advanced model-driven digital repository (the cardiac part of the ongoing MD-Paedigree FP7 project) and to data sources (such as the CaseReasoner developed in the Health-e-Child FP6 project and the simulation tools developed in the Sim-e-Child FP7 project). Using a hybrid trial design that combines all these existing data sources with prospective data, CARDIOPROOF is going to make future outcome studies more easily feasible, allowing them to be conducted in a more time and cost-effective manner.

# 2 Executive summary

Gnúbila, with consultation from clinical and technical partners has put in place a first version of the infrastructure of the CARDIOPROOF project. The infrastructure has been built on top on the ongoing FP7 project MD-Paedigree, as an extension it, in order to optimise the work and cost. This document describes the current and future software infrastructure and site deployments.

A presentation of the gnúbila's FedEHR product will be done. This is the key part of the solution. The platform will be decomposed in two parts in order to be presented: a backend and a frontend. A detailed architectural view of each part will be done. Also, after a complete FedEHR features presentation, the actual deployed infrastructure as well as the future one will be shown.

# 3 Introduction

Installed on top of the existing MD-Paedigree infrastructure, the current solution provides two nodes having the ability to host science gateways, some central services, and a web portal. The CARDIOPROOF project shall add at least one node to the infrastructure. The architecture is composed of hardware described in MD-Paedigree D14.1.

Most of the solution provided uses MD-Paedigree functionalities and resources to start at a high level of value. Improvement designed for CARDIOPROOF will profit directly and indirectly to the usability of MD-Paedigree with the adjunction of new functionalities and collection of different data. Using such a solution provides a significant cost reduction in terms of hardware infrastructure, software evolution, and maintenance.

This document will present the different elements of the infrastructure, specifying if it comes from MD-Paedigree as is, extended, or if it is new for CARDIOPROOF.

# 4 Software components

In CARDIOPROOF, the data infrastructure and information system is based on gnúbila's FedEHR product. As the BARC study "*Big Data Survey Europe*" from 2013[1] states, organizations are taking a serious view on big data, recognizing the critical success factors and issues associated with handling enormous data volumes. Like medical imaging in the 80s, big data is indeed about to reorganize medical practice. Big data not only is a major challenge for ICT and health care professionals, but also is a great opportunity. The use of massively available medical data may allow clinicians to simulate potential outcomes and ~~so~~ prevent patients from undergoing ineffective treatments or improve current treatments. In other words, accumulating and using data to develop a greater understanding of pathophysiological processes will result in significant healthcare improvements.

**Developed in collaboration with renowned medical centres in Europe, FedEHR is a patient-centric Electronic Health Records (EHR) big data solution, supporting this long-term goal.** FedEHR, stands for Federated EHR. It leverages on the cloud elasticity to provide a scalable vendor-neutral database, which is able to cope with massive multi-modal and heterogeneous medical information, data and knowledge integration. FedEHR takes its roots from leading edge technologies developed and tested in computationally and data intensive environments at the European Organization for Nuclear Research (CERN[2]). Over the last 7 years, FedEHR maturated from its application to diverse fields of medical science, from advanced biomedical research, to translational and clinical medicine. FedEHR was installed in reference hospitals internationally, including Necker Enfants Malades in Paris, France, the Great Ormond Street Hospital in London UK, Ospedale Bambin Gésu in Rome, ~~in~~ Italy, and the Johns Hopkins University hospital in Baltimore, USA. The solution also was awarded at major events from its inception, including the Gold Medal at the International Inventions Exhibition in Geneva[3] in Switzerland, in May 2007; the Technology Transfer Award from CNRS,[4] in December 2010; and the Best Exhibit Award of Europe's largest ICT conference, ICT 2013[5], last November 2013.

## 4.1 Overview

The system is composed of two main components:

- A **backend**, composed by all the Web Services being the ground of the platform with the addition of external applications;
- A **frontend**, offering a user-friendly web interface allowing user communities to easily interact with the platform.

The backend and the fronted are decoupled; it is possible to have multiple frontends accessing the same services with heterogeneous and/or specialized interfaces. The figure 1 gives a simplified overview of the platform architecture.

[1] http://www.pmone.com/fileadmin/user_upload/doc/study/BARC_BIG_DATA_SURVEY_EN_final.pdf
[2] Health Surveillance. How Knowledge Transfer Changed Biology, Medicine and Health Care. D. Manset. WILEY, 2014. In Press.
[3] Gold Medal at the International Inventions Exhibition in Geneva http://cds.cern.ch/record/1035139 "MAAT Gknowledge took a gold medal for MammoGrid, a GRID-based mammogram analysis system to be implemented in the Extremadura region in Spain".
[4] Technology Transfer Award from CNRS https://gnubila.fr/en_GB/awards-recognitions
[5] Best Exhibit Award at ICT 2013 http://ec.europa.eu/digital-agenda/en/news/meet-winners-best-exhibitors-ict-2013 MD-Paedigree.

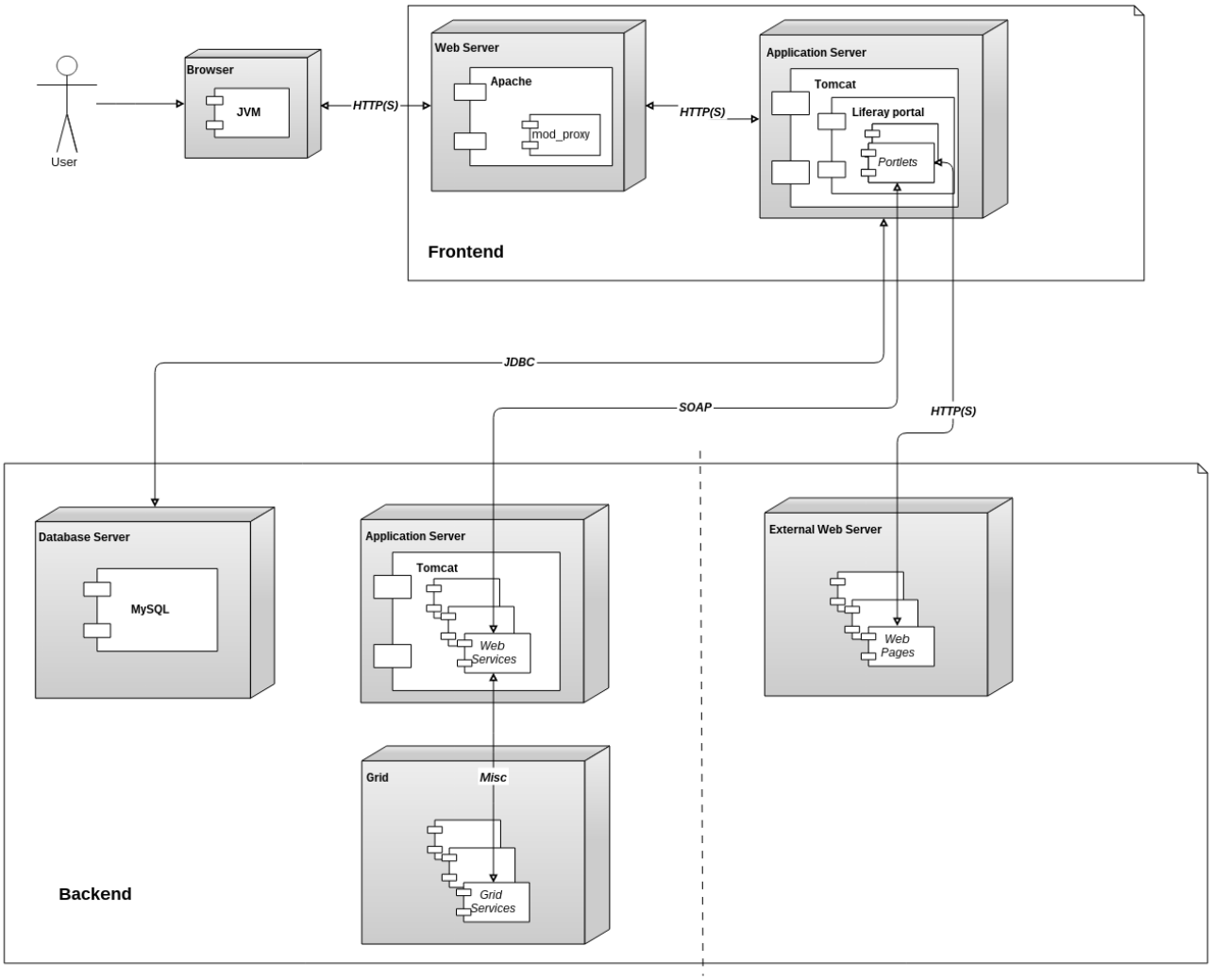



Figure 1: System Overview

### 4.1.1 The System Backend (inherited from MD-Paedigree)

As introduced previously, the CARDIOPROOF system is composed of two parts, the frontend and the backend. The backend, depicted in Figure 2, is the foundation of the system. It is a Service Oriented Architecture (SOA) composed of different components interacting together or with external entities. Most of the platform features are exposed as Web Services, allowing an easily decoupling of the user interface rendering from the business logic. The user Interfaces can offer powerful features by aggregating the Web Services. The Liferay frontend, which will be presented later, does this by using portlets.

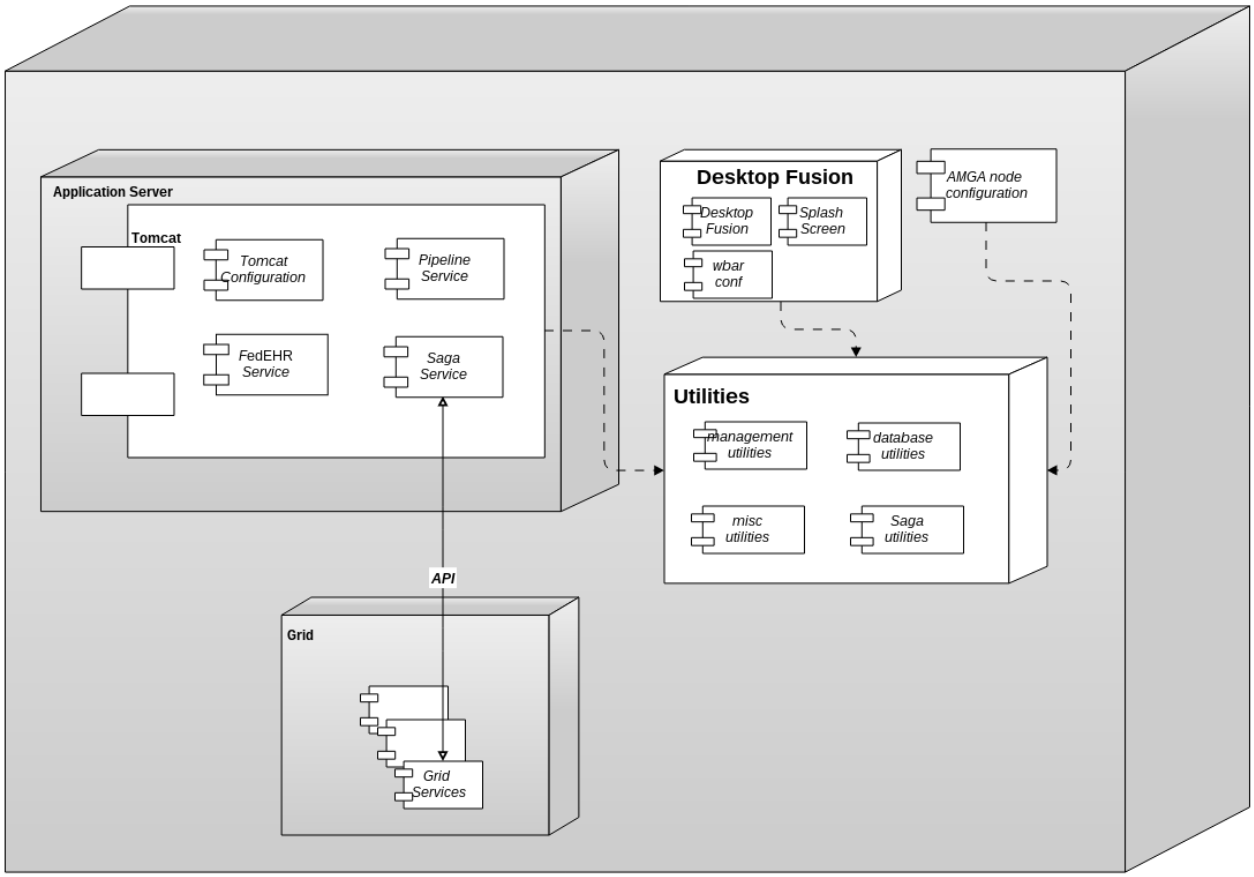



Figure 2: backend deployment diagram

This backend is provided by the gnúbila Pandora[6] solution and can be divided like this:

- **Pandora Desktop Fusion (inherited from MD-Paedigree):** Desktop Fusion is a remote desktop infrastructure that provides tools without requiring users to install them on their computers using ~~a~~ remote access to a pre-installed system.
- **Pandora Amga Node Configuration (extended from MD-Paedigree):** The Pandora system uses a distributed architecture where the whole information system is kept in interconnected AMGA databases. This information system contains the gateway information, which allows them to inter-operate. The CARDIOPROOF elements have been added to the node definitions to collaborate with the MD-Paedigree infrastructure.
- **Pandora Pipeline Service (inherited from MD-Paedigree):** The Pipeline Service is a web service responsible for executing tasks inside the Pandora Gateway. It organises task into simple workflows that can run in grid-configured environment.
- **Pandora Saga Service (inherited from MD-Paedigree):** The Saga Service is a Web Service provides an abstraction layer allowing interaction with various grid or even non-grid middleware or infrastructure.
- **Pandora core utilities (inherited from MD-Paedigree):** Pandora management utilities: data management-related Java archives
- **FedEHR Service (extended from MD-Paedigree):** As this is one of the most important services in MD-Peadigree and CARDIOPROOF, FedEHR is defined in section 4.3 ~~below~~.

[6] See http://www.gnubila.fr/

These components are installed using the Debian packages described in the table 1. Required dependencies are installed automatically by the software package manager. All of these components are continuously updated, either for bug fixes or new features implementation.

| Package Name | Version |
|---|---|
| pandora-gateway-desktopfusion | 2.2.0 |
| pandora-gateway-desktopfusion-splash-translational-medecine | 1.0-1 |
| pandora-wbar | 1.1-1 |
| pandora-gateway-idal-amga-node-configuration | 2.2.0 |
| pandora-gateway-sal-pipeline | 2.2.3 |
| pandora-gateway-sal-saga | 2.2.1 |
| pandora-gateway-idal-fedehr | 2.3.3.1 |
| pandora-gateway-sal-gateone | 2.2.0 |
| pandora-gateway-gateone-management | 2.2.0 |
| pandora-gateway-sl-core | 2.2.0 |
| pandora-gateway-sl-utils-management | 2.2.1 |
| pandora-gateway-sl-utils-misc | 2.2.7 |
| pandora-gateway-sl-utils-misc-jsaga | 2.2.2 |

### 4.1.2 The System Frontend (extended from MD-Paedigree)

The main entry point to the system is the frontend, which exposes all the CARDIOPROOF applications and tools. It is provided by the backend through a common and simple web-based interface.

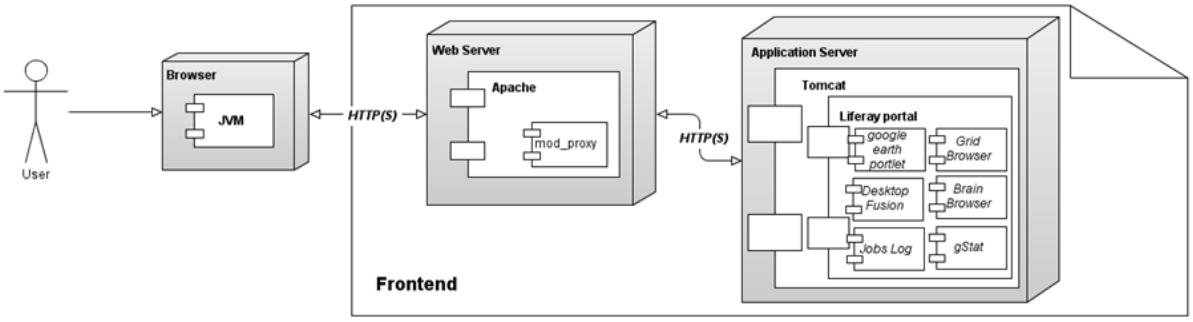



Figure 3: frontend deployment diagram

The system front end is now powered by Liferay 6.1.2 CE running using a SUN JDK 1.7. The authentication to the system is done using Security Assertion Markup Language (SAML) standard to provide an efficient and interoperable browser-based Single Sign On (SSO) service.

The SAML 2.0 Web Browser SSO Profile is used to allow the federation of identity providers. The MD-Paedigree Identity Provider (IdP) was setup using Shibboleth, but any external SAML-compliant IdP can be easily integrated with the Discovery Service. Shibboleth is the reference implementation of the SAML 2.0 specification and is widely used in the scientific community all over the world making it the perfect tool for making the system accessible to a broader audience.

A number of the CARDIOPROOF applications - exposed using portlets - are still being developed and will be delivered during the remaining project timeframe, but a number of applications are already available and accessible using the Liferay portal.

As shown in Figure 4, the portal has been extended to manage a CARDIOPROOF context and a theme has been created to render project identity. Whilst sharing lots of elements with MD-Paedigree, CARDIOPROOF has its own visual identity and users can utilise both portal instances or be dedicated to one.

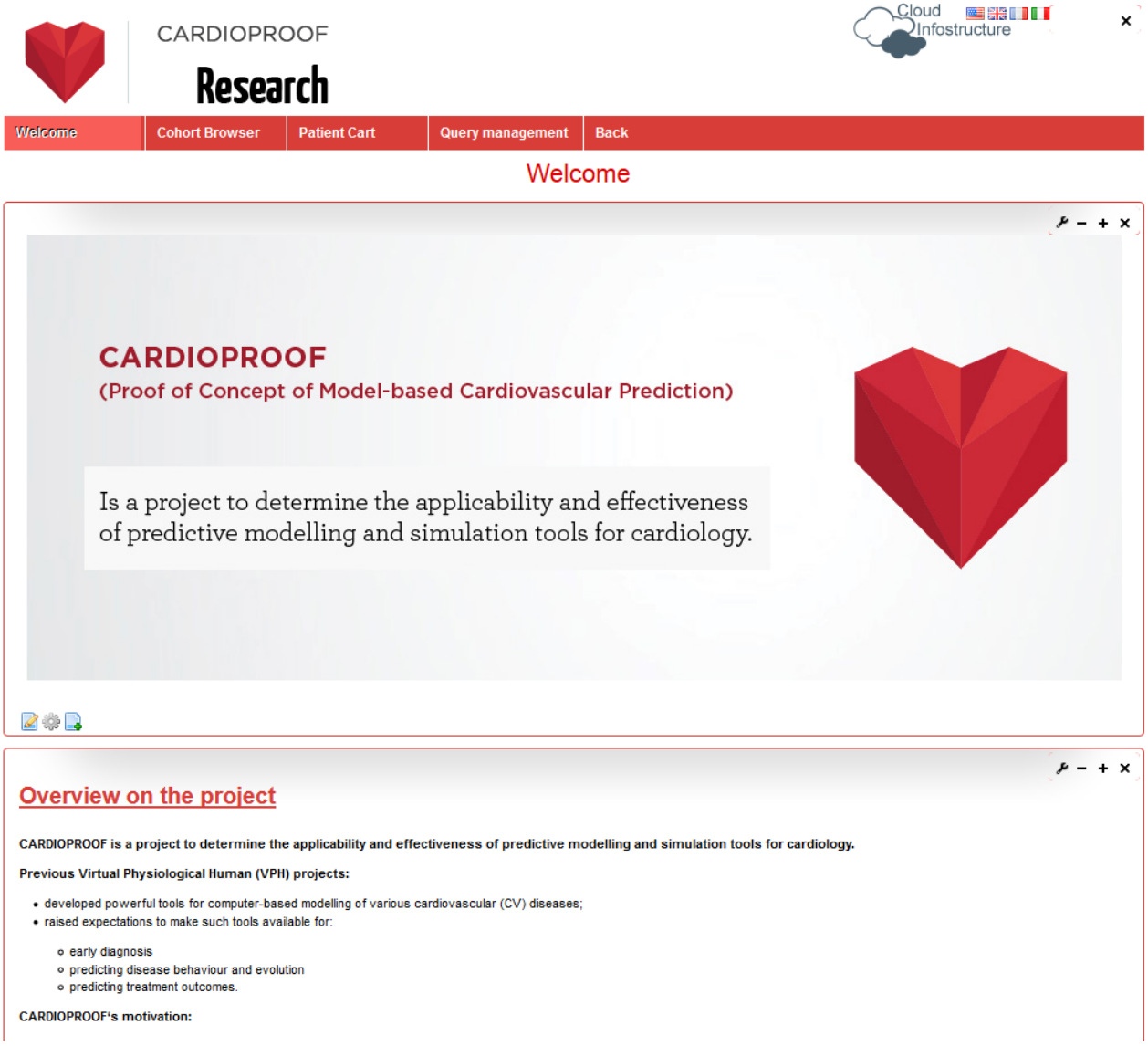



Figure 4: CARDIOPROOF Portal main page

## 4.2 FedEHR: Storage infrastructure

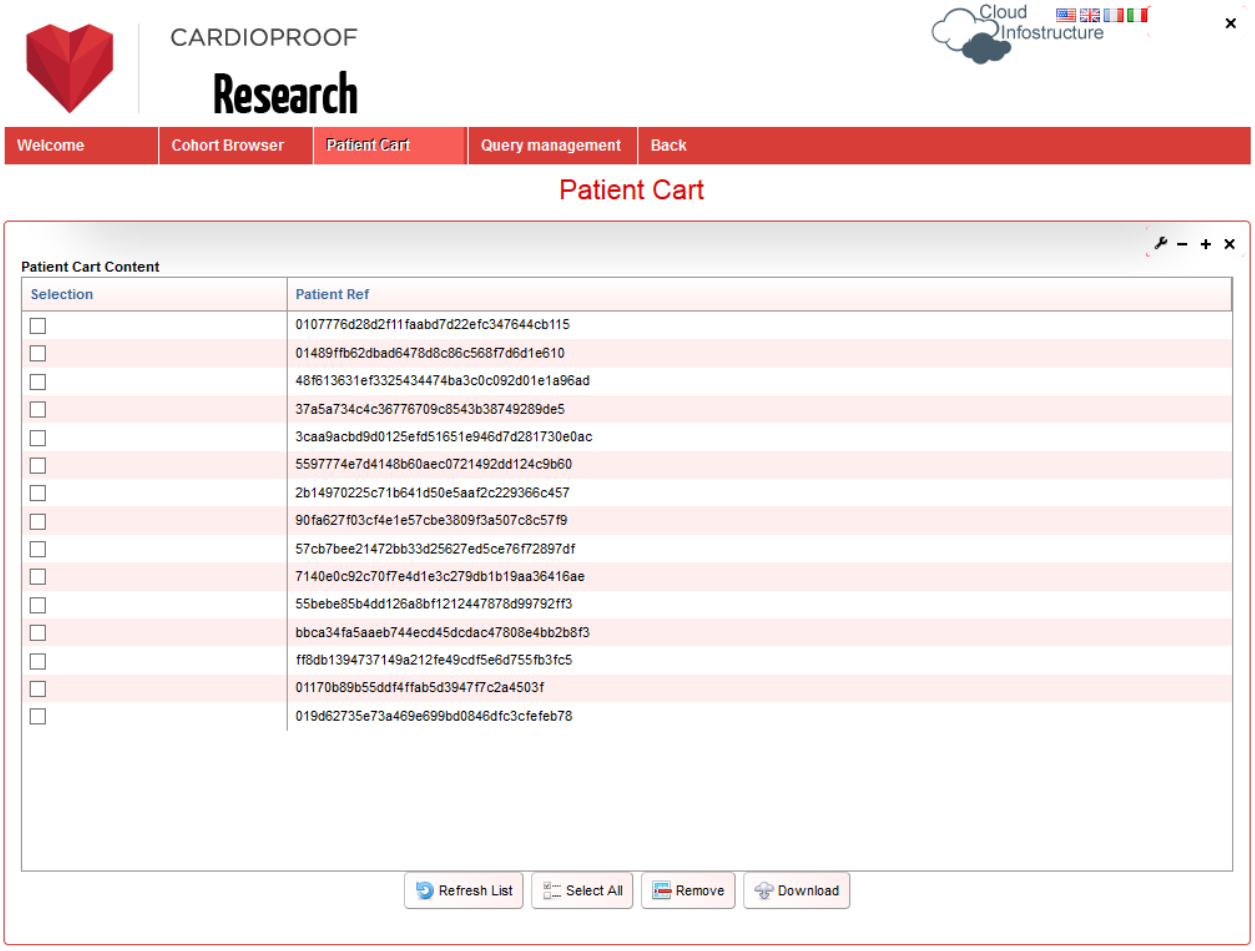
FedEHR is a service that provides storage abilities for medical events and patient information. The interfaces are exposed as web services consumable by other applications and user interfaces are integrated in the portal.

### 4.2.1 Patient centric data structure (inherited from MD-Paedigree)

In accordance with the latest data modelling concepts in the literature, FedEHR proposes a storage model that is fully centred on patients. All data that is stored in the system is organised around a data structure representing a patient model. The current description of FedEHR architecture provides an evolutionary structure of data starting from the patient. Currently, the data structure is oriented around medical concepts of medical events and clinical variables. These abstract models can be refined and specialised using metadata definitions created from physicians' descriptions of diseases and exams or from normative works.

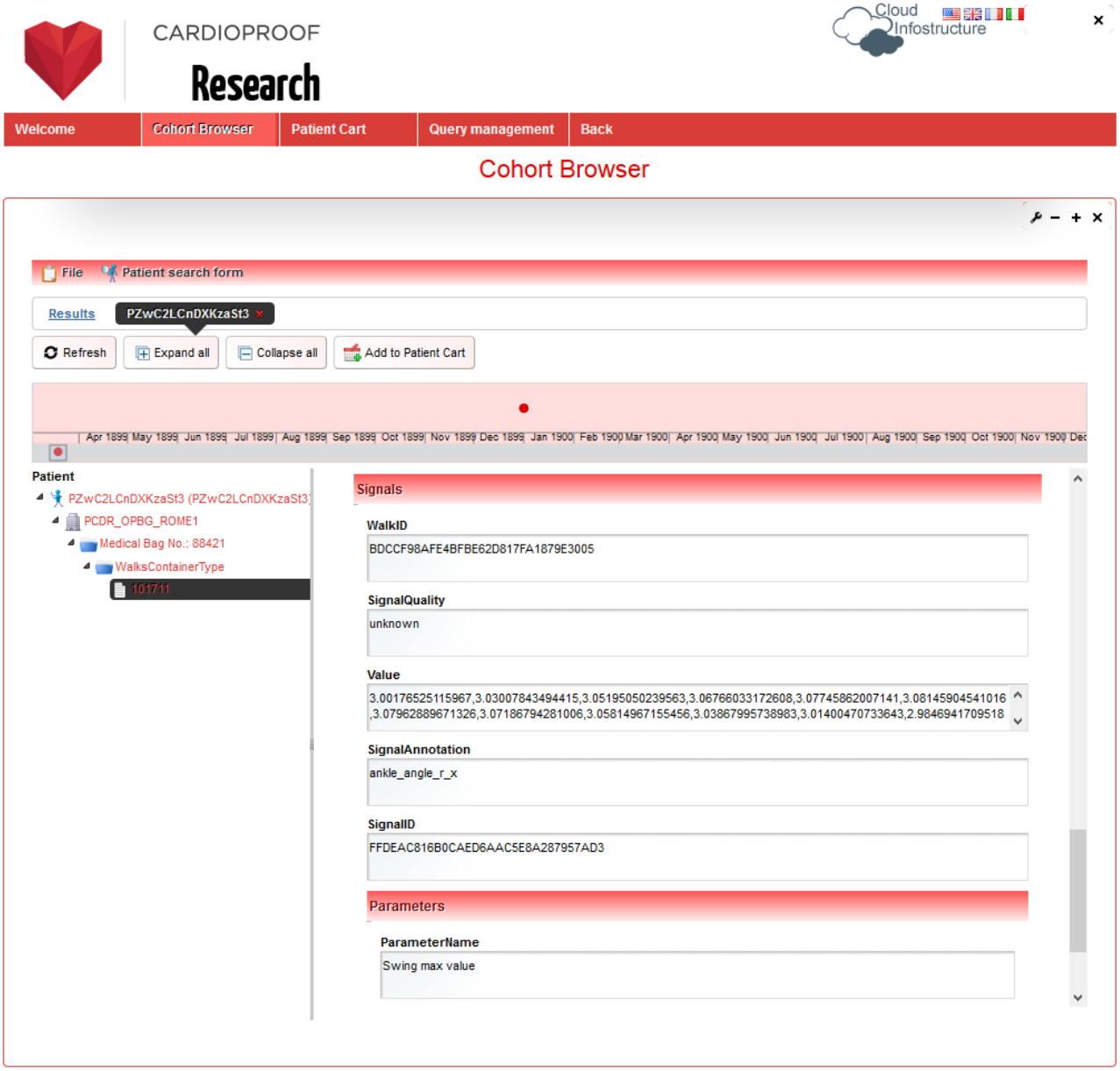
### 4.2.2 Patient Cart (new)

The first need expressed during the project was to be able to easily choose a patient to add to a cohort of study. To respond to this issue, a first version of Patient Cart portlet has been designed and developed. This portlet is composed of 2 elements, a service that is harvesting the patient to add to the cart patent list from some other portlets, and a graphical user interface that manages the list. The aim of this development is to add ability to select patient(s) in all the existing interfaces that show, sort, or list patients on criteria and to send these patients to the patient cart. The current patient cart portlet only permits showing and downloading the ID list of patients as a csv file. The evolution of this portlet will allow users to generate, share, and save patient cohorts into the system. This cohort will allow running platform connected tools on its element. A download button may be added to download the full list of patients with their data (remaining anonymised). Particular attention will be paid to determine at any moment who downloaded data in order to ask him or her to remove the data in case of legal need (such as revoke of consent form a patient or change of employer for a physician).



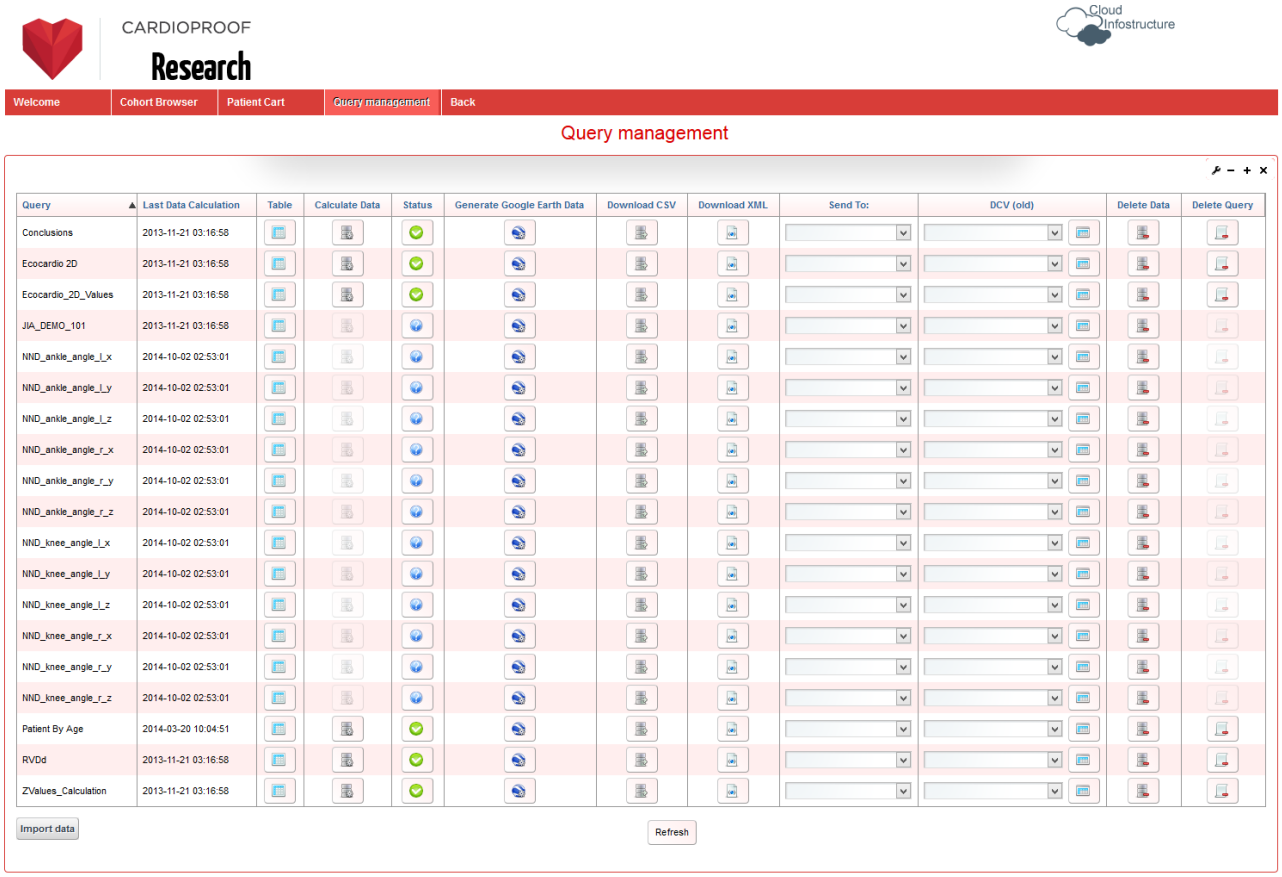
### 4.2.3 Cohort Browser (extended from MD-Paedigree)

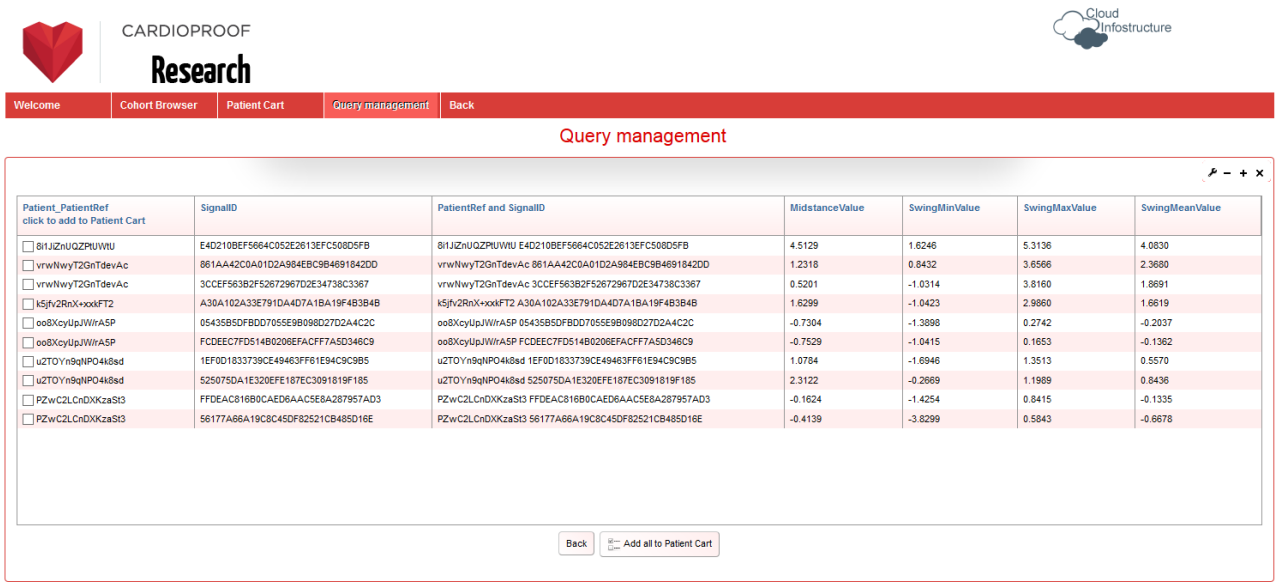
The Patient Browser, as its names indicates, is a portal integrated feature that allows a physician to access the full information about a patient from all the nodes of the system. It provides a complete medical history of the patient regardless of the physical location of data. Inherited from MD-Paedigree, the Patient Browser has been updated for CARDIOPROOF in order to be integrated with the Patient Cart. Patients can be added to the Patient Cart from the patient list or the patient view.



### 4.2.4 Query System (extended from MD-Paedigree)

FedEHR provides an inter-site query system presented as an SQL query for end users. These queries are managed by a query management system which generates a result set that can be downloaded in a variety of formats and also includes a graph visualisation tool. Stored data is not useful without a query system. Inherited from MD-Paedigree, the Query System has been updated to be able to enrich the CARDIOPROOF Patient Cart by adding some checkboxes to add or remove patients from the list.

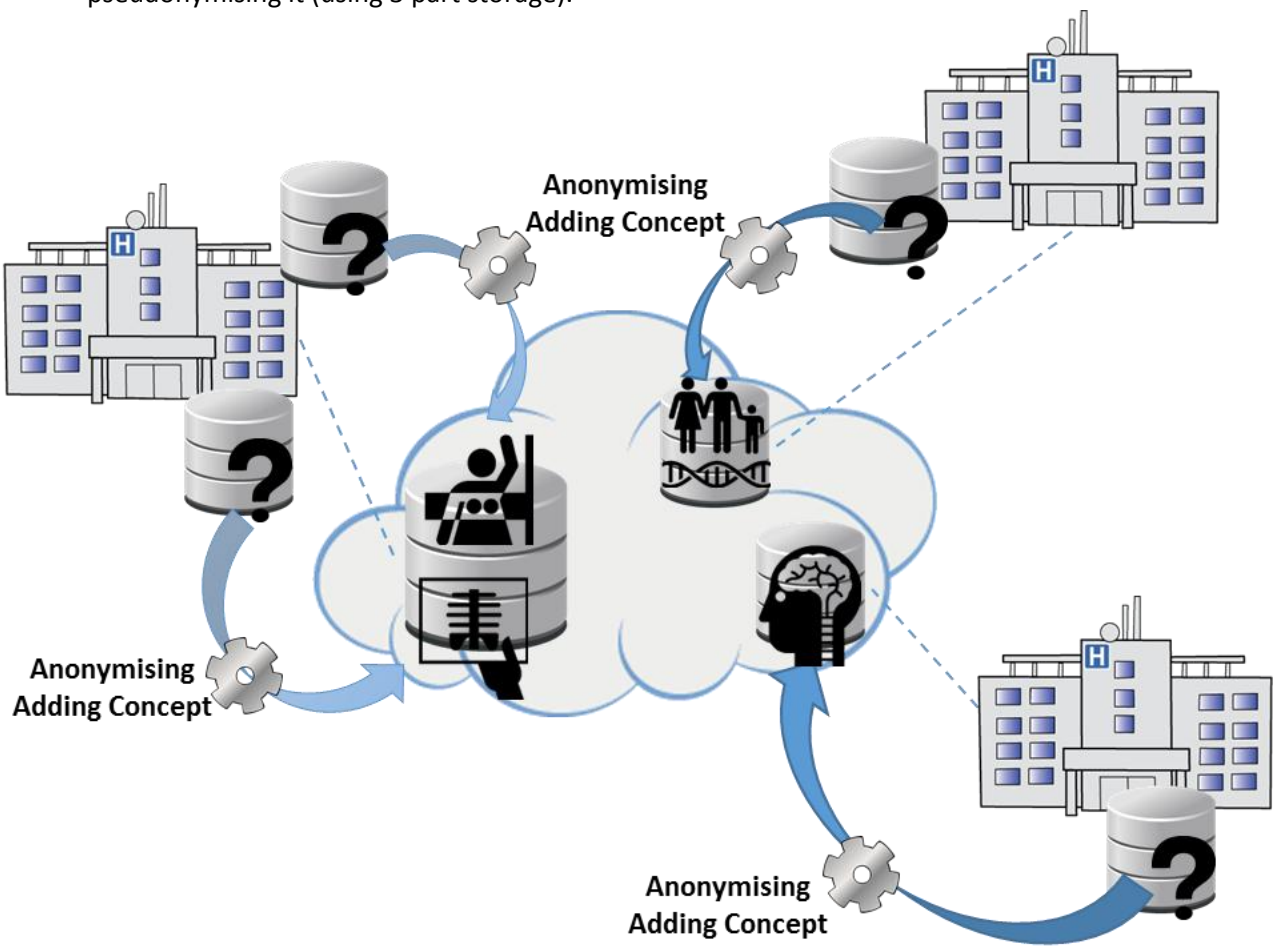


### 4.2.5 Importers (new)

Importers are usually written in Java and are used in the export of data from routine systems to provide metadata information from images and measurements and textual information from annotations.
The importers are:

- taking data from the routine system,
- normalising the data (using a hospital specific structure), and
- pseudonymising it (using 3 part storage).



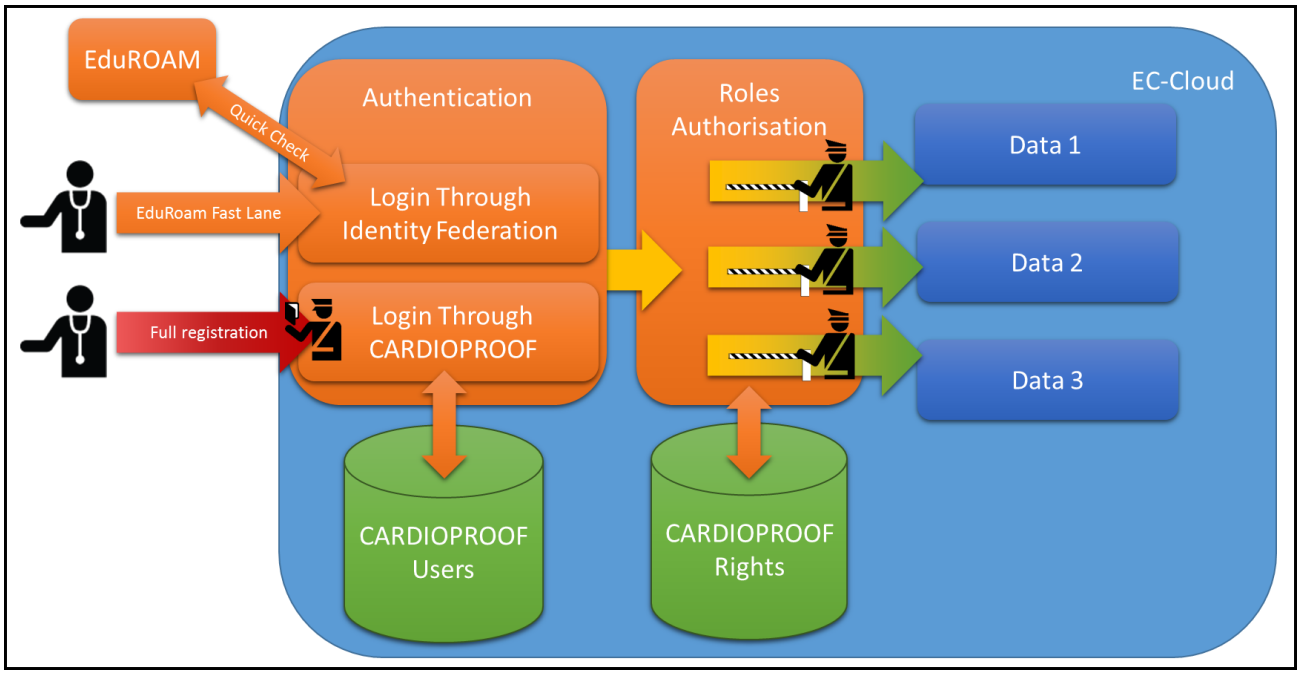

During this process, some ontological information such as ICD or SNOMED could be added in order to provide a uniform concept-based query capacity. This would maintain the ability to have a personal, easy, and understandable data structure at each node.
The first importer from CARDIOPROOF will integrate DHZB data from the TrialConnect system.

### 4.2.6 Access Rights

FedEHR system includes a RBAC (Role Based Access Control) mechanism for data security. For each patient medical event, the access can be defined based on roles given or not given to the users.

The provider of the data decides what access is given to whom and is able at any moment to see the list of access provided to a particular user. The data owner can revise these access rights at any moment.
Roles can be given to groups that can be multiplied as needed. A group can be assigned to member of a whole hospital, a special service, or represent inter-hospital group. Group mangers can manage the groups. Groups can be defined for the whole system or gateway by gateway.

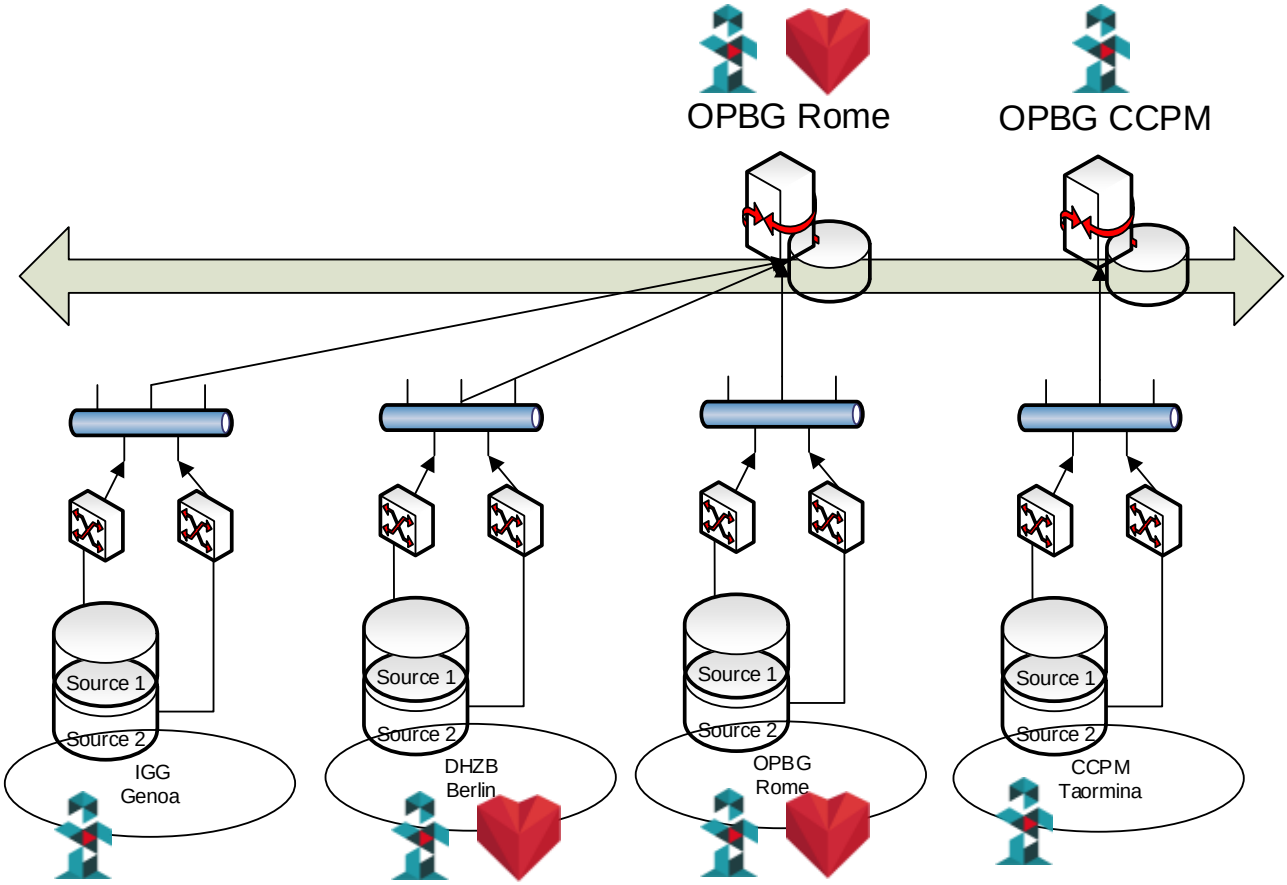
This access controller is coupled by the capacity to authenticate using an identity provider (such as EduRoam), which can also provide institution and status dependent group information.



# 5 Overview of infrastructure node installation

## 5.1 Current infrastructure

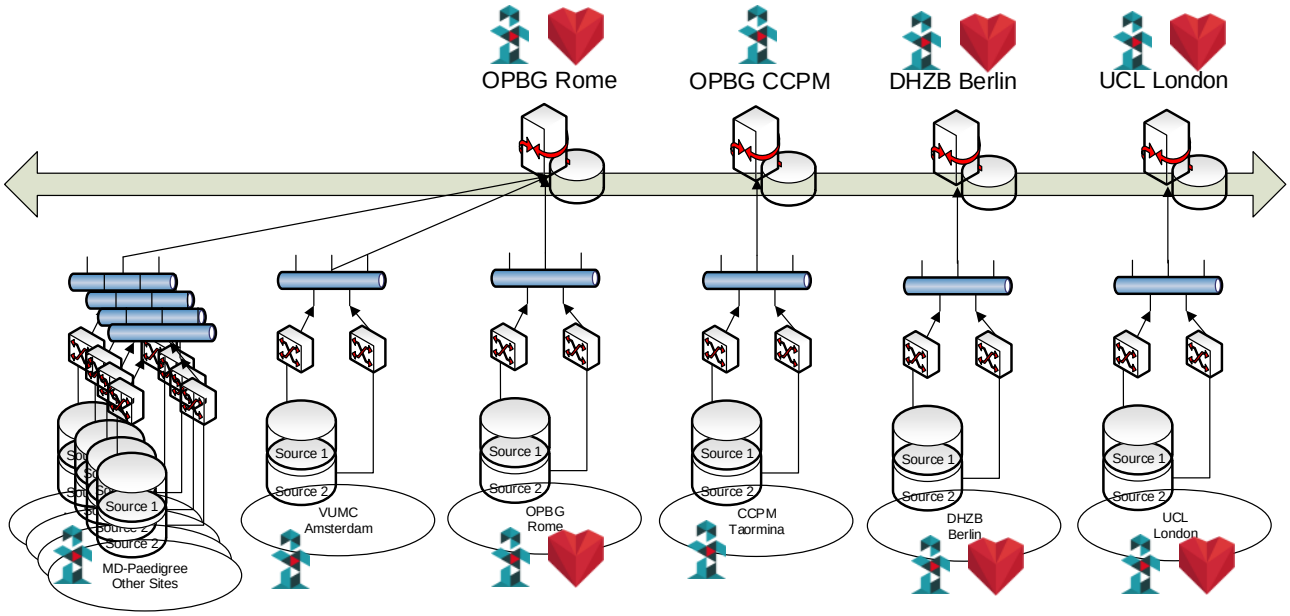
The original infrastructure is based on MD-Paedigree one. This begins with data collection without waiting for the additional hardware to be installed.




The Rome Gateway has been updated to allow CARDIOPROOF data to be stored. The DHZB Gateway has been installed in Rome and is waiting for migration to system in Berlin when available. This gateway was needed for MD-Paedigree, but was created in advance for the needs of CARDIOPROOF.

## 5.2 Next step infrastructure

It is planned to add at least two more nodes for CARDIOPROOF, one in London at UCL, and one in Berlin at DHZB. Once the hardware is installed, the prepared or existing virtual machines will be moved to their final destination. Lastly, the system should provide multiple physical nodes in different countries all collaborating for data sharing. Access rights will define if data is available for CARDIOPROOF users, MD-Paedigree Users, or both.




# 6 Conclusion

CARDIOPROOF's Data Infrastructure and Information System is based on gnúbila's FedEHR product, which is already the core of the MD-Paedigree infrastructure.

FedEHR has been presented and divided into a backend and a frontend ~~part~~. The backend provides a set of secure services which provides all the needed functionalities to the platform whereas the frontend is mainly composed of graphical user interfaces embedded into a Liferay portal.

Some of the infrastructure components were directly inherited from MD-Paedigree but others have been created in order to respond to the CARDIOPROOF project needs, in particular with the creation of the first version of the Patient Cart interfaces.

Finally, the current status of the infrastructure and the proposed future update were presented and show that everything is already well advanced.