Model Driven Paediatric European Digital Repository
Call identifier: FP7-ICT-2011-9 - Grant agreement no: 600932
Thematic Priority: ICT - ICT-2011.5.2: Virtual Physiological Human

Deliverable 14.4
Infrastructure Final Release Report

Due date of delivery: 31-05-2017
Actual submission date: 22-06-2017

Start of the project: 1st March 2013
Ending Date: 28th February 2017

Partner responsible for this deliverable: MAAT
Version: 1.0

SEVENTH FRAMEWORK PROGRAMME

Dissemination Level: Public

Document Classification

Title	MD-Paedigree Infrastructure, Final Release Report
Deliverable	14.4
Reporting Period	4
Authors	Sebastien Gaspard
Work Package	14
Security	PU
Nature	RE
Keyword(s)	Final release, Infrastructure Report

Document History

Name	Remark	Version	Date
Deliverable 14.4		0.1	19/06/2017
	Reviewed version	0.2	21/06/2017
Deliverable 14.4		1.0	22/06/2017

List of Contributors

Name	Affiliation
Sebastien Gaspard	MAAT
Jérome Revillard	MAAT
David Manset	MAAT
Harry Dimitropoulos	ATHENA
Emilie Pasche	HES-SO
Steven Wood	USFD

List of reviewers

Name	Affiliation
Harry Dimitropoulos	ATHENA
Bruno Dallapiccola	OPBG

Abbreviations

API	Application Programming Interface
CSV	Comma-Separated Values (file format)
DCV	Data Curation and Validation
DPS	Data Publication Suite
eCRF	Electronic Case Report
ETL	Extract Transform Load
KDD	Knowledge Discovery and Data Mining / Knowledge Discovery in Databases
MDP	MD-Paedigree (project)
MG	Multigrid Method
NND	Neurological and Neuromuscular Diseases
GPUs	Graphics Processing Units
GUI	Graphical User Interface
One–Versus–All (OVA)	OVA
PCG	Preconditioned Conjugate Gradient method
RBAC	Role-Based Access Control
SVM	Support Vector Machine
UDF	User-Defined Function
CDR	Clinical Data Repository
CBR	Case-Based Retrieval

Table of Contents

1 Project summary

MD-Paedigree is a clinically-led VPH project that addresses both the first and the second actions of part B of Objective ICT-2011.5.2:
1. it enhances existing disease models stemming from former EC-funded research (Health-e-Child and Sim-e-Child) and from industry and academia, by developing robust and reusable multi-scale models for more predictive, individualised, effective and safer healthcare in several disease areas;
2. it builds on the eHealth platform already developed for Health-e-Child and Sim-e-Child to establish a worldwide advanced paediatric digital repository.

Integrating the point of care through state-of-the-art and fast response interfaces, MD-Paedigree services a broad range of off-the-shelf models and simulations to support physicians and clinical researchers in their daily work. MD-Paedigree vertically integrates data, information and knowledge of incoming patients, in participating hospitals from across Europe and the USA, and provides innovative tools to define new workflows of models towards personalised predictive medicine. Conceived of as a part of the "VPH Infostructure" described in the ARGOS, MD-Paedigree encompasses a set of services for storage, sharing, similarity search, outcome analysis, risk stratification, and personalised decision support in paediatrics within its innovative model-driven data and workflow-based digital repository. As a specific implementation of the VPH-Share project, MD-Paedigree fully interoperates with it. It has the ambition to be the dominant tool within its purview. MD-Paedigree integrates methodological approaches from the targeted specialties and consequently analyses biomedical data derived from a multiplicity of heterogeneous sources (from clinical, genetic and metagenomic analysis, to MRI and US image analytics, to haemodynamic, to real-time processing of musculoskeletal parameters and fibres biomechanical data, and others), as well as specialised biomechanical and imaging VPH simulation models.

2 Executive summary

As an update of D14.3, this document will just present the novelties of the infostructure. To have an exhaustive view of the platform, please refer to D14.3 and D14.2.

This document will first present the current state of FedEHR based repository, followed by the importation tools (importers and anonymizer). Then, the GUIs offering new functionalities and the integration of HES-SO Case-Based Retrieval (CBR) will be described, followed by the Athena tools and the Sheffield university VPH-Share DPS tool.

Following the usual order since D14.1, work on GPUs will end this presentation just after a presentation of the anonymiser tools that gnúbila has provided to the project.

We remind the reader that all these components are coming in addition to the tools presented in deliverables D14.1, D14.2 and D14.3 (we chose not to present them another time in this document) and that this document does not by itself transcribe the entirety of the amount of work provided during the project.

3 FedEHR a Clinical Data Repository (CDR)

3.1 Hardware Architecture

Current architecture is composed of:

- 1 Node at OPBG (Rome) providing gateways for
  - OPBG
  - IGG
  - UMCU
  - VUMC
- 1 Node at DHZB (Berlin)
- 1 Node at UCL (London)
- 1 Node at KUL (Leuven)
- 1 Portal
- 1 Central Server

The nodes are currently installed and connected together through a FastWeb secured connection. This allows all sites to share information with ease.

3.1.1 D14.1 probable implementation

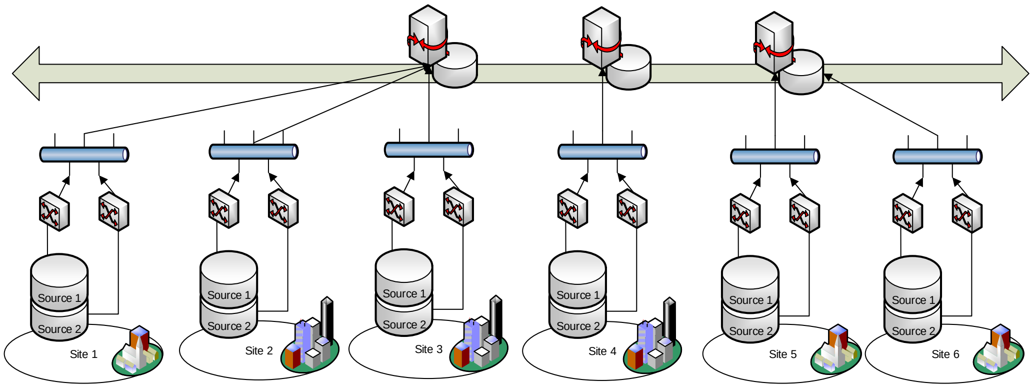


Figure 1: Initially proposed architecture implementation

Early in the project it has been identified that the architecture would probably be a hybridization of centralised and distributed nodes with probably 3 sites hosting hardware and the others partners using the nearest hardware to instantiate their gateway.

3.1.2 Final Installation

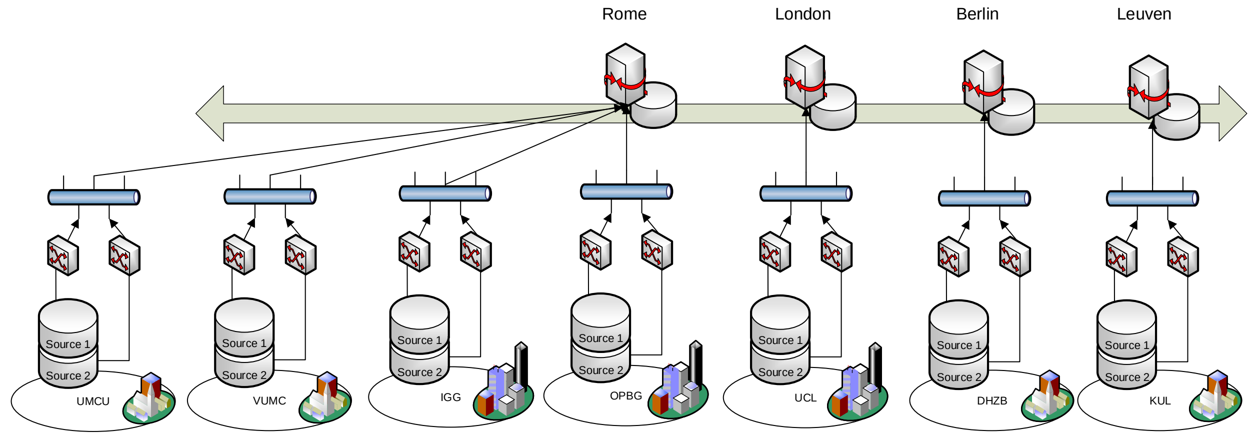


Figure 2: Final installation architecture

Thanks to the efforts of every centre, the final system supporting the repository corresponds to what was expected at the beginning of the project but with a little more site providing their own nodes.

3.2 Software Repository elements

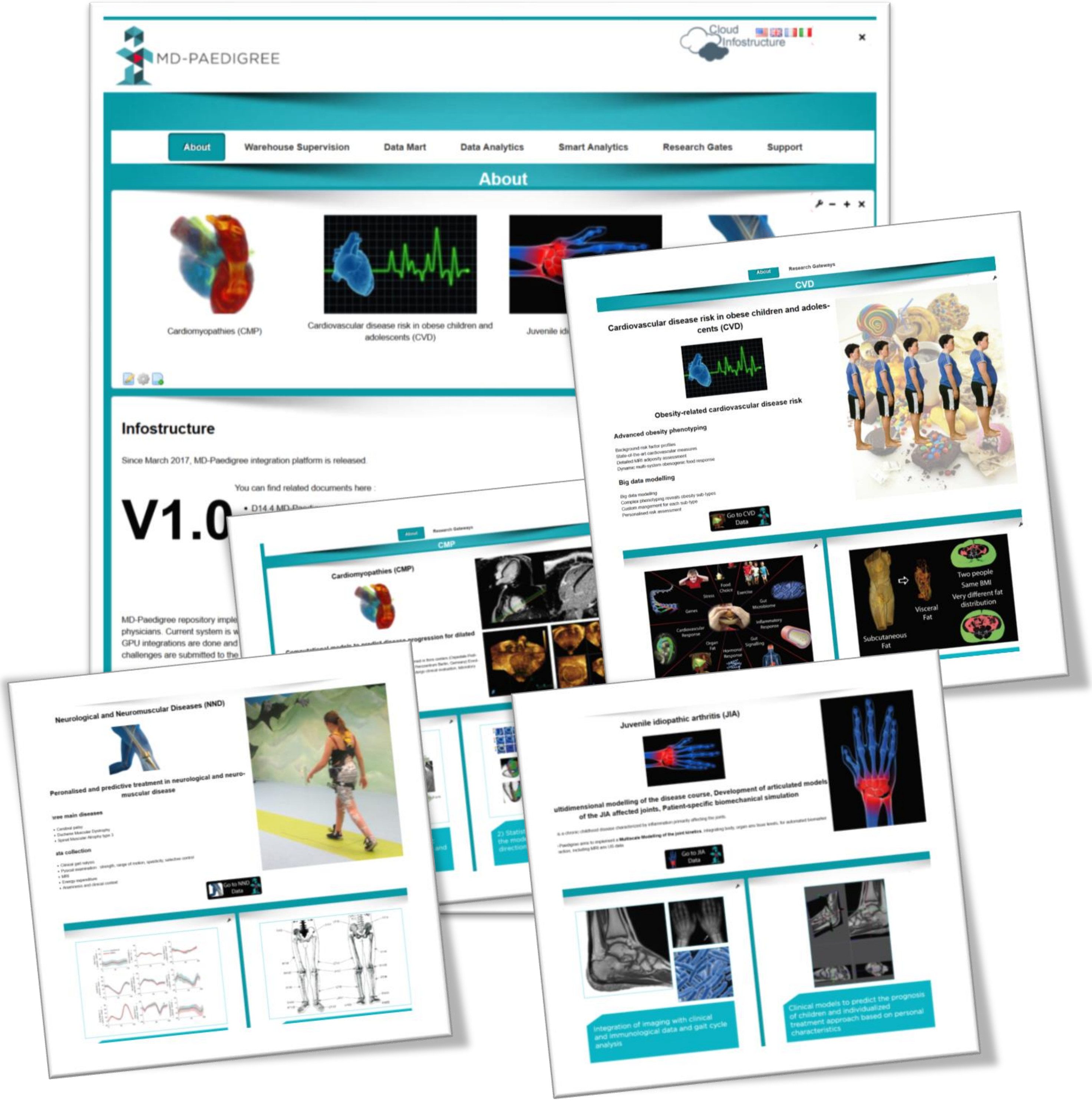
All functionalities have been developed and installed since the Beta version, only testing and improvement have taken place during this last year.

System provides the following elements:

pandora-gateway-desktopfusion-management	2.1.0
pandora-gateway-desktopfusion-splash-translational-medecine	1.1-1
pandora-gateway-gateone-management	2.2.1
pandora-gateway-idal-amga-node-configuration	2.2.4
pandora-gateway-idal-fedehr	2.4.3
pandora-gateway-sal-desktopfusion	2.2.0
pandora-gateway-sal-gateone	2.2.0
pandora-gateway-sal-pipeline	2.2.4
pandora-gateway-sal-saga	2.2.3
pandora-gateway-sl-core	2.2.5
pandora-gateway-sl-utils-management	2.2.5
pandora-gateway-sl-utils-misc	2.2.14
pandora-gateway-sl-utils-misc-jsaga	2.2.4

3.3 Portal

Liferay portal has been installed in the new version generated for the MD-Paedigree Release, substantial amount of work has been carried out on stability and optimisation, and a huge refactoring of dependencies have been realised to simplify maintenance and evolutions.
The portal is composed of several pages and some portlets designed and configured for the project.



Infostructure

Since March 2017, MD-Paedigree integration platform is released.

V1.0

You can find related documents here :

- D14.4 MD-Paed

MD-Paedigree repository imple
physicians. Current system is w
GPU integrations are done and
challenges are submitted to the

Cardiovascular disease risk in obese children and adolescents (CVD)

Obesity-related cardiovascular disease risk

Advanced obesity phenotyping

Background risk factor profiles
State-of-the-art cardiovascular measures
Detailed MRI adiposity assessment
Dynamic multi-system obesogenic food response

Big data modelling

Big data modelling
Complex phenotyping reveals obesity sub-types
Custom mangement for each sub-type
Personalised risk assessment

Cardiomyopathies (CMP)

Computational models to predict disease progression for dilated

Neurological and Neuromuscular Diseases (NND)

Personalised and predictive treatment in neurological and neuromuscular disease

Juvenile idiopathic arthritis (JIA)

ultidimensional modelling of the disease course, Development of articulated models of the JIA affected joints, Patient-specific biomechanical simulation

Integration of imaging with clinical and immunological data and gait cycle analysis

Clinical models to predict the prognosis of children and individualized treatment approach based on personal characteristics

Figure 3: Portlet screenshot examples

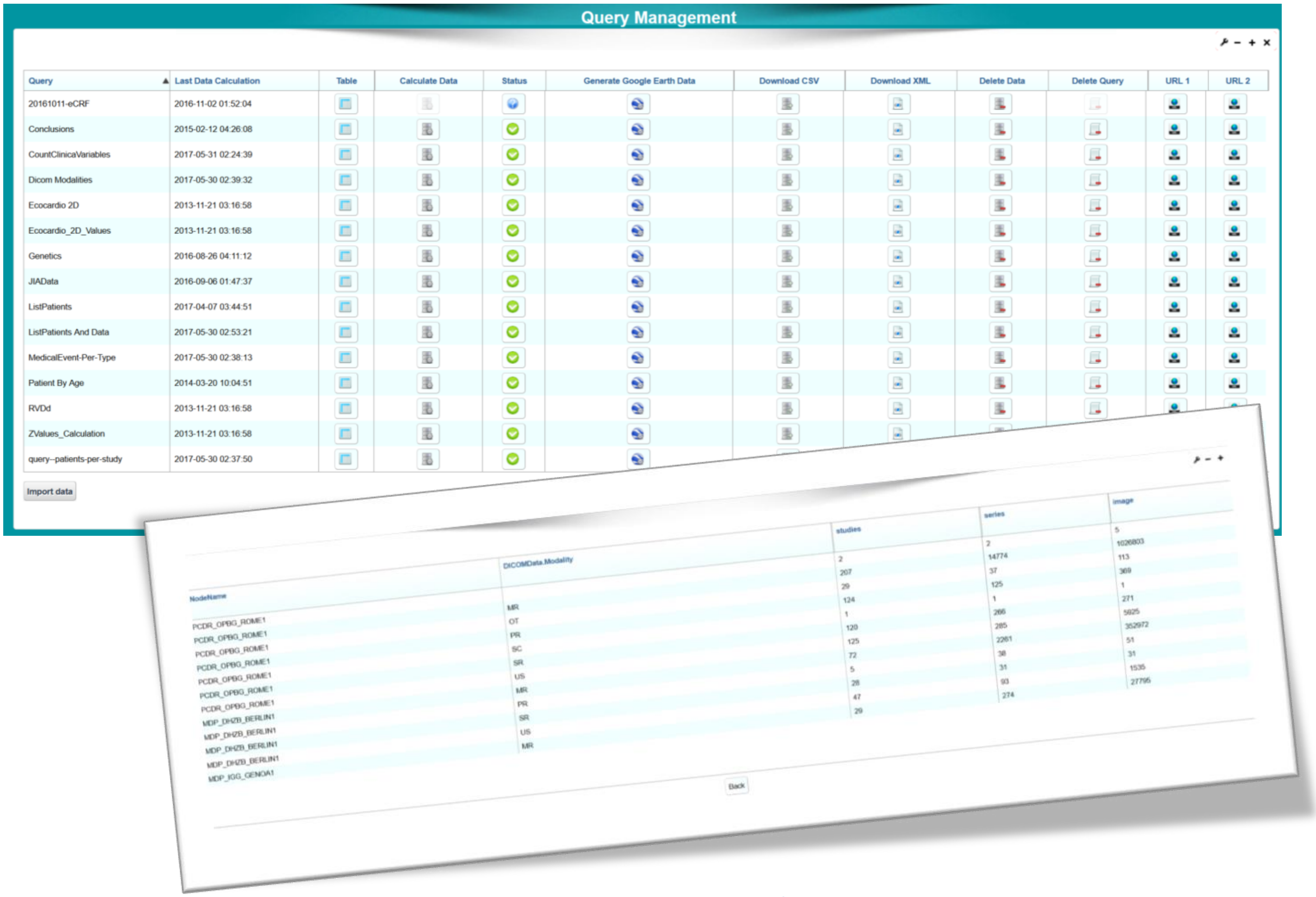
3.3.1 Technical hidden portlets :

pandora-ext	1.6.1
pandora-global-lib	1.6.1
pandora-services-portlet	1.6.1

3.3.2 Graphical interfaces portlets :

3.3.2.1 data-management-portlet 1.6.1

Stored data is not useful without a query system. 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 resultset that can be downloaded in a variety of formats and a graph visualisation tool.



Query Management

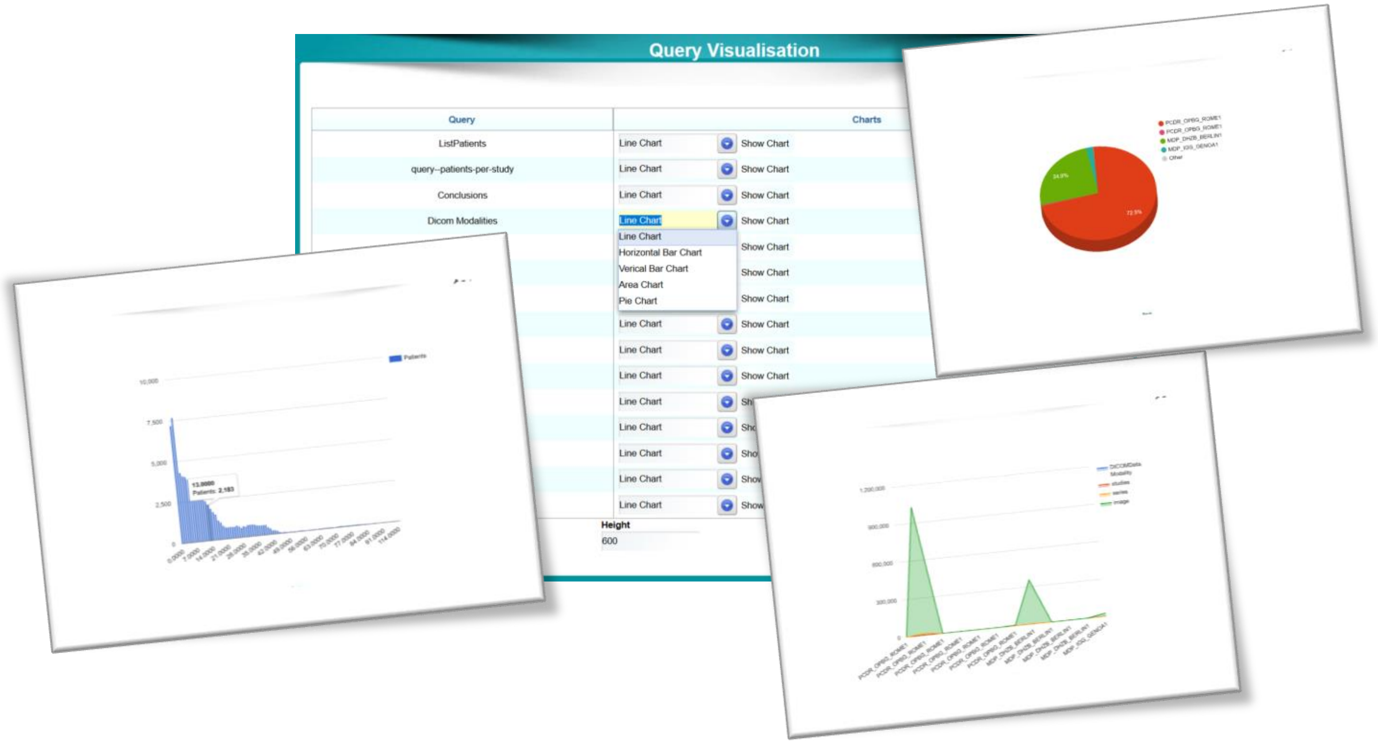
Query ▲	Last Data Calculation	Table	Calculate Data	Status	Generate Google Earth Data	Download CSV	Download XML	Delete Data	Delete Query	URL 1	URL 2
20161011-eCRF	2016-11-02 01:52:04										
Conclusions	2015-02-12 04:26:08			✓							
CountClinicaVariables	2017-05-31 02:24:39			✓							
Dicom Modalities	2017-05-30 02:39:32			✓							
Ecocardio 2D	2013-11-21 03:16:58			✓							
Ecocardio_2D_Values	2013-11-21 03:16:58			✓							
Genetics	2016-08-26 04:11:12			✓							
JIAData	2016-09-06 01:47:37			✓							
ListPatients	2017-04-07 03:44:51			✓							
ListPatients And Data	2017-05-30 02:53:21			✓							
MedicalEvent-Per-Type	2017-05-30 02:38:13			✓							
Patient By Age	2014-03-20 10:04:51			✓							
RVDd	2013-11-21 03:16:58			✓							
ZValues_Calculation	2013-11-21 03:16:58			✓							
query--patients-per-study	2017-05-30 02:37:50			✓							

NodeName	DICOMData.Modality	studies	series	image
PCDR_OPBG_ROME1		2	2	5
PCDR_OPBG_ROME1	MR	207	14774	1026803
PCDR_OPBG_ROME1	OT	29	37	113
PCDR_OPBG_ROME1	PR	124	125	369
PCDR_OPBG_ROME1	SC	1	1	1
PCDR_OPBG_ROME1	SR	120	266	271
PCDR_OPBG_ROME1	US	125	285	5925
MDP_DHZB_BERLIN1	MR	72	2261	352972
MDP_DHZB_BERLIN1	PR	5	38	51
MDP_DHZB_BERLIN1	SR	28	31	31
MDP_DHZB_BERLIN1	US	47	93	1535
MDP_IGG_GENOA1	MR	29	274	27795

Figure 4: Data management portlet

3.3.2.2 fedehr-chart-portlet 1.6.1

The portlet allows to plot data calculated by the data-management-portlet.



Query Visualisation

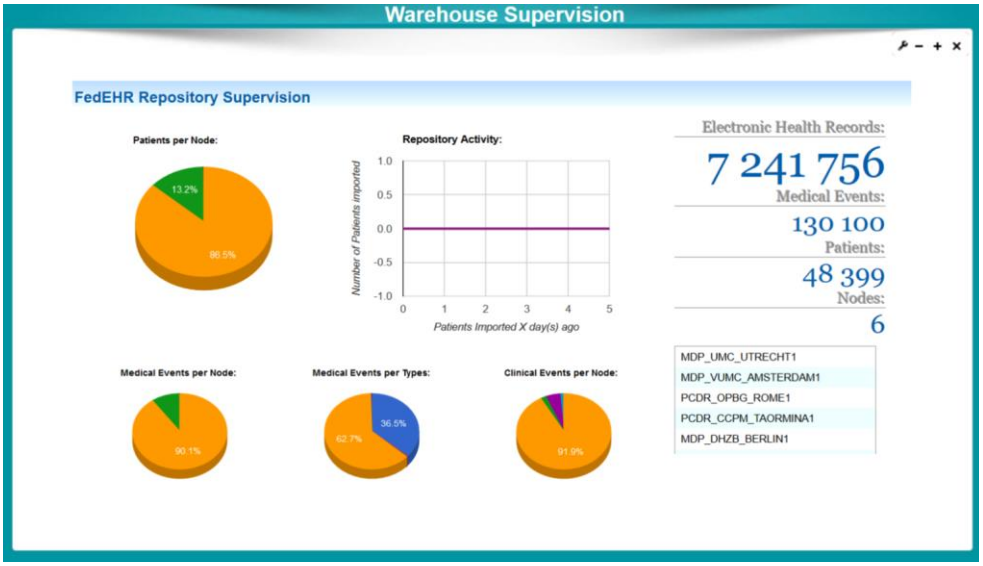
Query	Charts	
ListPatients	Line Chart	Show Chart
query--patients-per-study	Line Chart	Show Chart
Conclusions	Line Chart	Show Chart
Dicom Modalities	Line Chart	Show Chart
		Show Chart
		Show Chart
		Show Chart
	Line Chart	Show Chart
	Line Chart	Show Chart
	Line Chart	Show Chart
	Line Chart	Sh
	Line Chart	Sho
	Line Chart	Sho
	Line Chart	Show
	Line Chart	Show

Height: 600

Figure 5: FedEHR chart portlet

3.3.2.3 fedehr-repository-supervision-portlet 1.6.1

This portlet automatically reports the statistics and evolution on the repository data.



Warehouse Supervision

FedEHR Repository Supervision

Electronic Health Records: 7 241 756

Medical Events: 130 100

Patients: 48 399

Nodes: 6

- MDP_UMC_UTRECHT1
- MDP_VUMC_AMSTERDAM1
- PCDR_OPBG_ROME1
- PCDR_CCPM_TAORMINA1
- MDP_DHZB_BERLIN1

Figure 6: FedEHR repository supervision portlet

3.3.2.4 patient-browser-portlet 1.6.1

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.

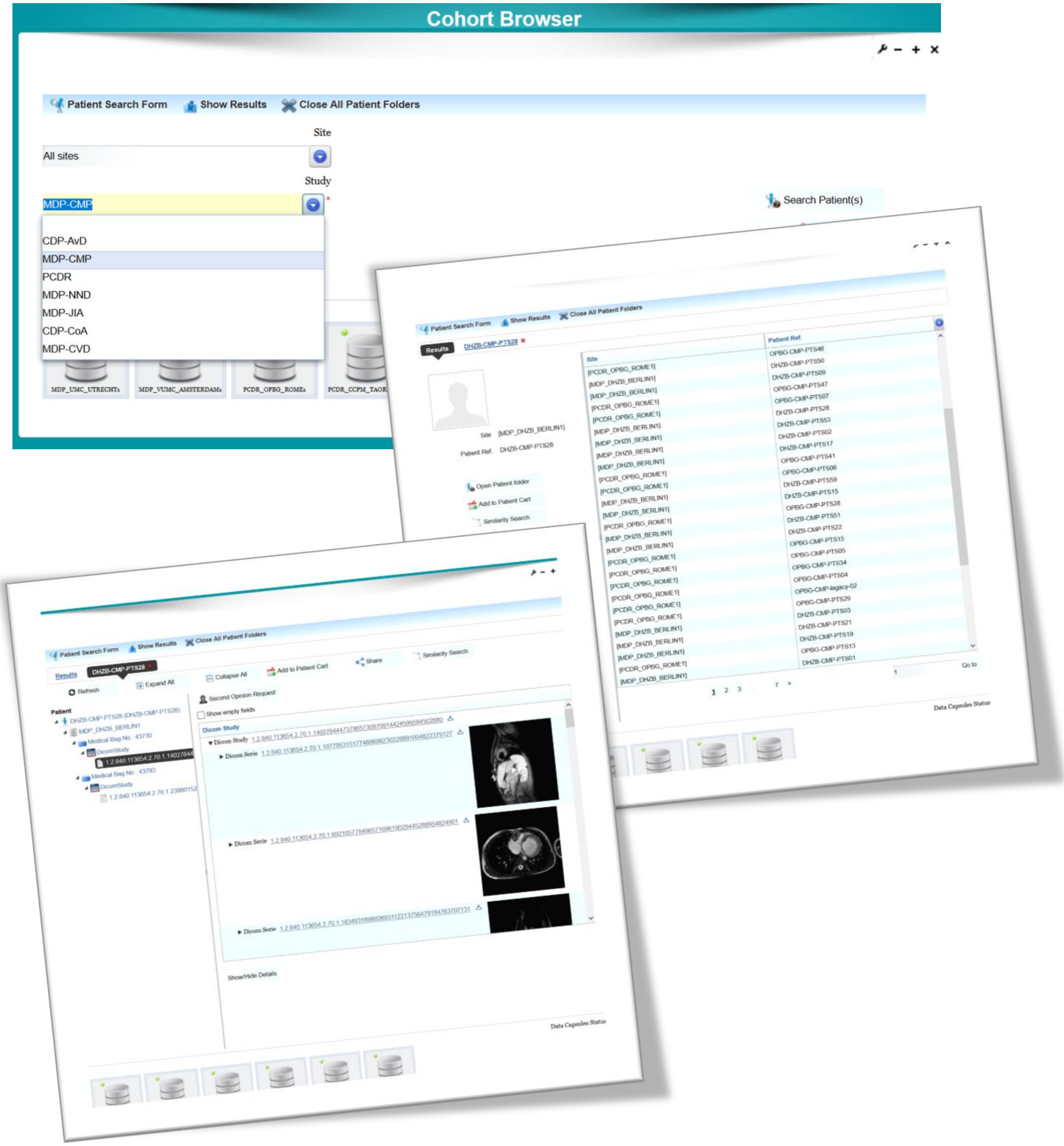
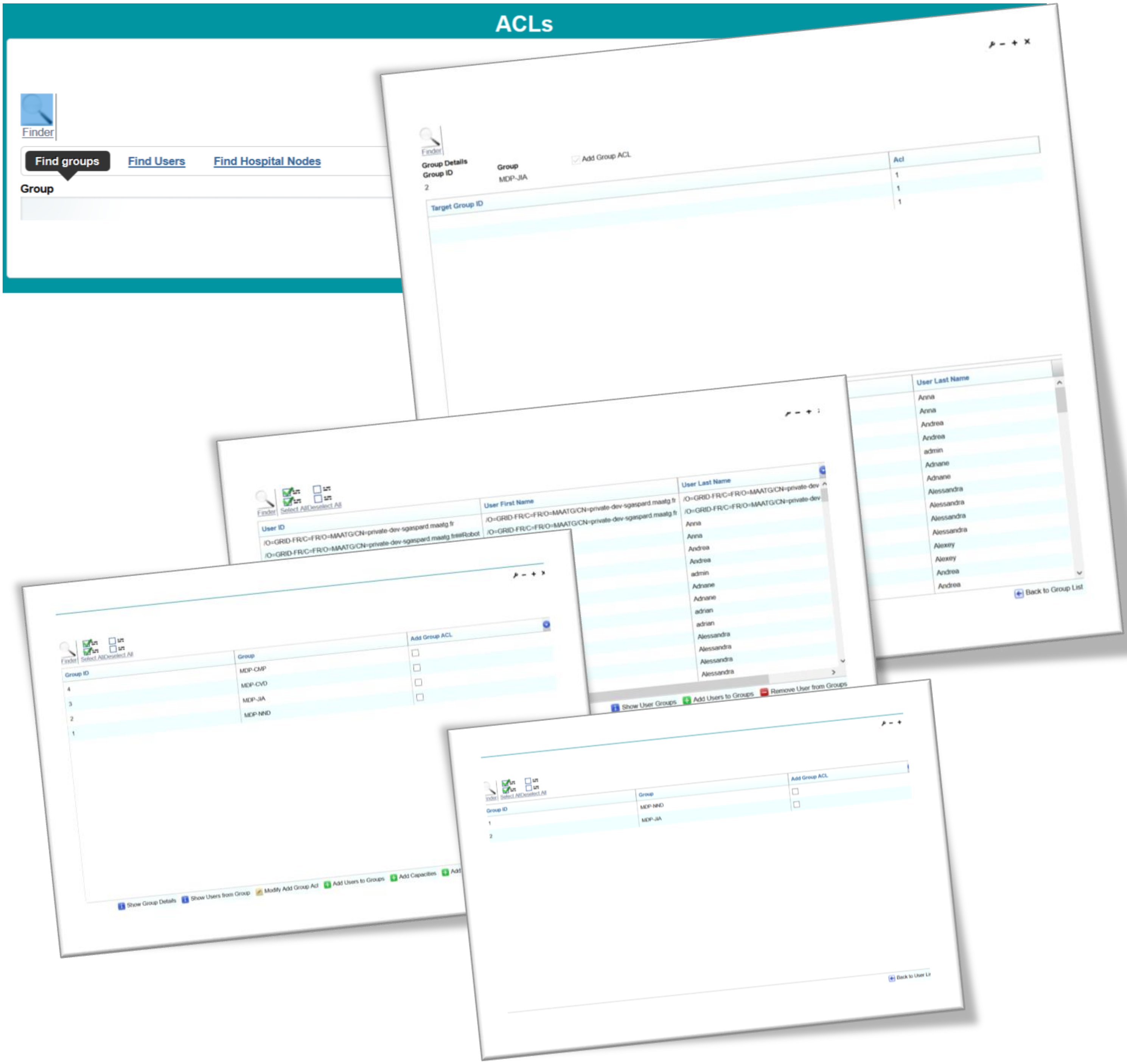


Figure 7: Patient browser portlet

3.3.2.5 fedehr-acls-management-portlet 1.6.1

This portlet has been developed to allow the management of access rights, managing Roles and users. As described in D14.3 the access rights is a RBAC (Role-Based Access Control).



Group Details

Group ID	Group	☑ Add Group ACL
2	MDP-JIA	

Target Group ID	Acl
	1
	1
	1

User ID	User First Name	User Last Name
/O=GRID-FR/C=FR/O=MAATG/CN=private-dev-sgaspard.maatg.fr	/O=GRID-FR/C=FR/O=MAATG/CN=private-dev-sgaspard.maatg.fr	/O=GRID-FR/C=FR/O=MAATG/CN=private-dev
/O=GRID-FR/C=FR/O=MAATG/CN=private-dev-sgaspard.maatg.frisRobot	/O=GRID-FR/C=FR/O=MAATG/CN=private-dev-sgaspard.maatg.fr	/O=GRID-FR/C=FR/O=MAATG/CN=private-dev
/O=GRID-FR/C=FR/O=MAATG/CN=		Anna
		Anna
		Andrea
		Andrea
		admin
		Adnane
		Adnane
		adrian
		adrian
		Alessandra
		Alessandra
		Alessandra
		Alessandra

User Last Name
Anna
Anna
Andrea
Andrea
admin
Adnane
Adnane
Alessandra
Alessandra
Alessandra
Alessandra
Alexey
Alexey
Andrea
Andrea

Group ID	Group	Add Group ACL
4	MDP-CMP	□
3	MDP-CVD	□
2	MDP-JIA	□
1	MDP-NND	□

Group ID	Group	Add Group ACL
1	MDP-NND	□
2	MDP-JIA	□

Figure 8: FedEHR ACLS management portlet

3.4 Repository Importers

3.4.1 DICOM

This importer is a key feature for the projects, independent of the provider the machine used for acquisition or the nature of the image, this importer is able to insert any DICOM data into the repository referencing some selected metadata and anonymizing metadata and pixels on the fly.
This importer has been used to add 901 studies containing 24.432 series with 1.754.387 images.

3.4.2 DPS

VPH-Share DPS has been used to import complex data. Using the VPH-Share existing tool, a standard connector has been designed to enable DPS generated XML to be imported directly in the repository. The connector analyses the structure of the XML and automatically creates the corresponding types in the repository before storing data.

3.4.3 CSV

Generic CSV to repository connector has been finalised and tested. This connector can automatically load CSV data to a corresponding structure in the repository. Coupled with the anonymizer, it becomes easy to load any CSV data to the repository.

3.4.4 eCRF

Specific eCRF offered by gnùbila has provided the results and saw some adjustments. An export to CSV has been developed and the tool automatically generates the repository type corresponding to its structure. Doing so, the CSV importer could be reused to speed up the developments.

4 Gnùbila Anonymizer

Gnùbila anonymizer is utilised for DICOM images at two different moments in the project:

- when data acquisition partners encounter difficulties in anonymising their data and when regulations make it necessary to anonymise before doing anything with the data
- within data importers to ensure the anonymity of provided data.

Anonymization is based on the HIPAA validated norm DICOM PS3-15 modified for the specific needs of the project. New Anonymization rules have been added to deal with specific images containing text provided by the different data providing partners. All non-matching images have been identified, analysed and a profile has been created.

5 Case-based retrieval

5.1 Text-based

The Case-Based Retrieval (CBR) service, developed by HES-SO, aims to help physicians to find similar patients based on the clinical reports of a given patient. In the system, the basic search item is the episode of care. The service proposes several functionalities, such as automatic MeSH normalization, Rocchio-based query expansion, automatic feature extraction.

Figure 9A illustrates the pre-processing workflow of the CBR. On the HES-SO server, an index is created:

1. Data are extracted from the PCDR through the secure PCDR API developed by Gnùbila. HES-SO obtained a GRID certificate delivered by SwiNG (i.e. one of the certificate authorities delivering GRID certificates in Switzerland) in order to be trusted by the PCDR server. Thus, the global workflow includes a secured synchronization between HES-SO and the MD-Paedigree Portal. The following demographic information are extracted: gender, birth date, date of the episode of care, conclusion content (i.e. clinical synthesis).
2. Data are then automatically processed: a set of up to 10 MeSH concepts are attributed to each discharge summary and specific data type extraction is performed (e.g. ejection fraction, SaO_2, etc.).
3. Data are then accessed using a local index, which stored stems and the aforementioned data. The index is currently composed of 47,433 episodes of care corresponding to 33,674 patients.

Figure 10B illustrates the workflow at query time. The graphical user interface, together will all dependent services, are located on the MD-Paedigree Portal. The services communicate with the index to obtain the similar cases list using Json exchange messages:

1. The user query is automatically reformulated: up to 10 MeSH concepts are automatically suggested to the user and specific data type extraction is performed, thus enabling the user to do range search on these fields (e.g. to find patients with a similar SaO_2 value).
2. Based on the reformulated query, similar cases are retrieved from the Solr index.
3. Finally, query refinement can be performed, based on a Rocchio algorithm suggesting keywords to add to the query.

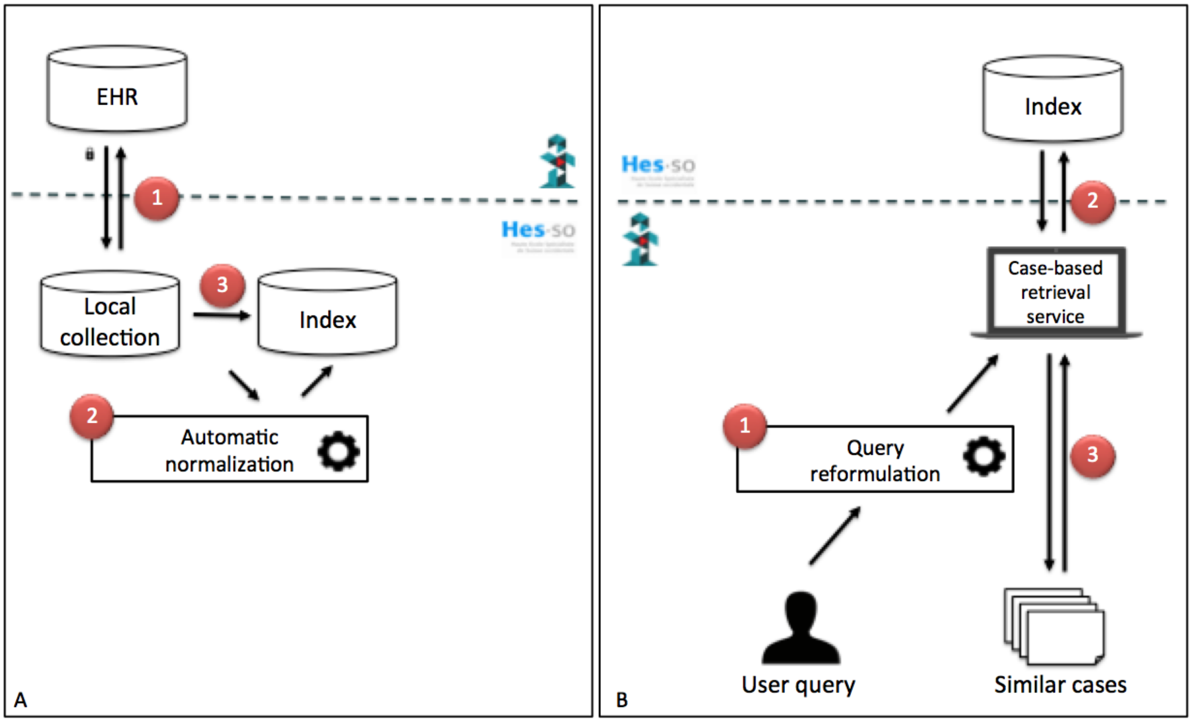
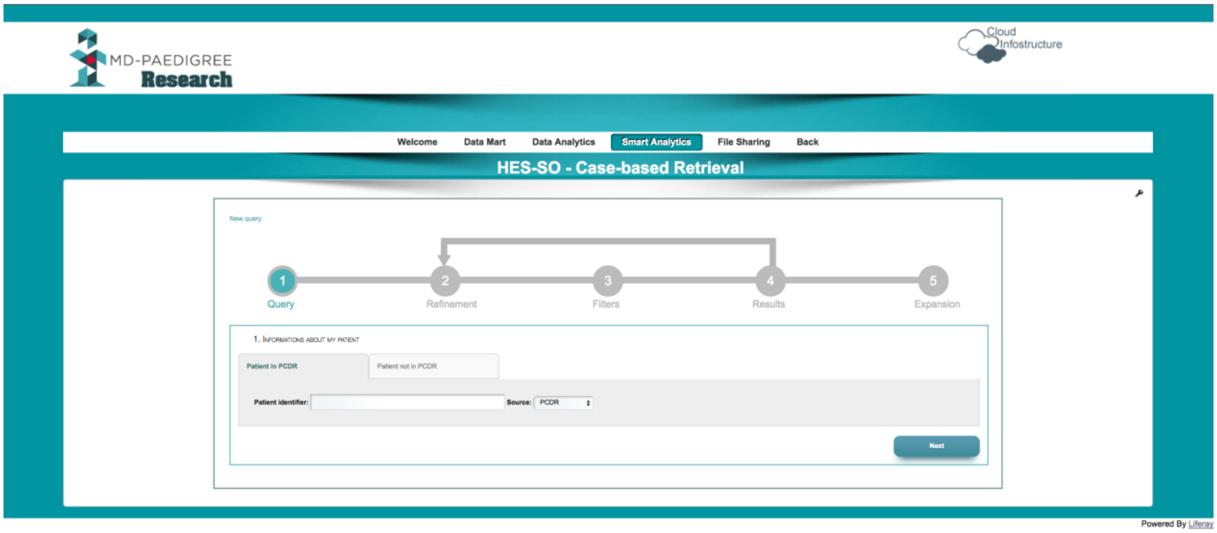


Figure 9: Case-based retrieval: A) pre-processing of the data. B) workflow at query time.

Regarding the graphical user interface, a full integration of the CBR has been performed in January 2016. The HES-SO team has been supported by the Gnùbila team, in order to be able to integrate the current version of the CBR and its future update. An instance of the Liferay Portal has been locally installed (version 6.1.2) at HES-SO for sake of development. The CBR servlet has been transformed to a portlet. Once ready, the HES-SO team pushed the final version on a web-based Git repository manager and the Gnùbila team deployed it on the MD-Paedigree portal. In April 2017, the same procedure was applied and an updated version of the CBR has been deployed on the portal (Figure 10).



HES-SO - Case-based Retrieval

1. Informations about my patient

Patient identifier: Source: PCDR

Figure 10: Fully integrated GUI of the CBR

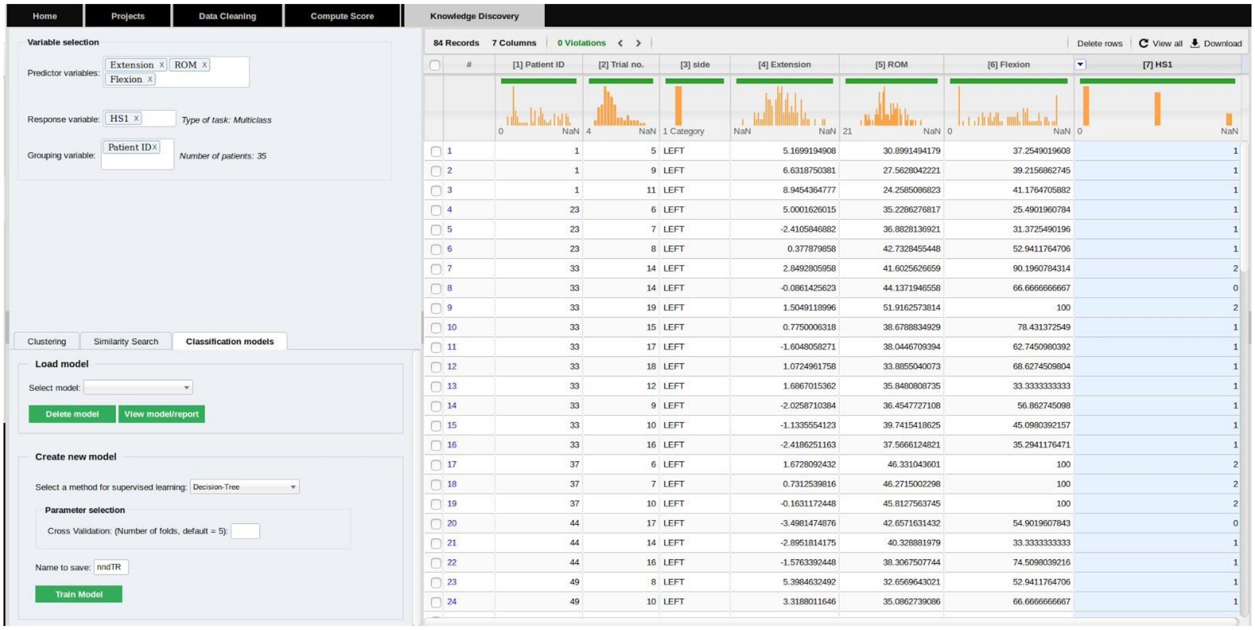
6 Biomedical Knowledge Discovery (KDD) and simulation platform.

During the last eighteen months of the project, ATHENA's web-based biomedical knowledge discovery platform [D16.2, "Beta Prototype of KDD & Simulation platform"] came to its final release by integrating with the Data Curation and Validation (DCV) tool [D15.2, "DCV curation tools and services to automatically and manually acquire high-quality curated data"].

The platform provides a web-based, end-to-end data profiling, curation/cleaning, pre-processing, analytics and knowledge discovery platform for big data healthcare. The platform runs on top of ATHENA's EXAREME[1] dataflow processing system (described later), which provides distributed processing and parallelization of resource/time-consuming algorithms related to knowledge discovery, simulation and data mining. The entire platform has only one point of integration with the MD-Paedigree platform.

Since the beta release of the platform [reported in D14.3, "Beta version Infrastructure Deployment Report"], the final release includes a completely redesigned user interface to allow for the integration of the DCV component with the KDD modules of the platform. In addition, a number of new pre-processing features were added. Each user can now export data regions of interest through interactive visualisations, or can save a subset of the dataset by deleting selected rows and columns. Missing values can be imputed to handle the missing values in the datasets during the training phase of models. Similarity search was also enriched with new parameter selection options and posterior probability tables were added after a classification analysis. Once a computational model has been developed, a comprehensive report is provided to assist the user with the model's assessment.

Since the platform was developed under WP16, complete details of these platform developments are reported in D16.3, "Final Release of KDD & Simulation Platform".



Variable selection

Predictor variables: Extension x ROM x Flexion x

Response variable: HS1 x *Type of task: Multiclass*

Grouping variable: Patient IDx *Number of patients: 35*

Load model

Select model:

Create new model

Select a method for supervised learning: Decision-Tree

Parameter selection

Cross Validation: (Number of folds, default = 5):

Name to save: nndTR

84 Records 7 Columns 0 Violations

#	[1] Patient ID	[2] Trial no.	[3] side	[4] Extension	[5] ROM	[6] Flexion	[7] HS1
	0 NaN	4 NaN	1 Category	NaN NaN	21 NaN	0 NaN	0 NaN
1	1	5	LEFT	5.1699194908	30.8991494179	37.2549019608	1
2	1	9	LEFT	6.6318750381	27.5628042221	39.2156862745	1
3	1	11	LEFT	8.9454364777	24.2585086823	41.1764705882	1
4	23	6	LEFT	5.0001626015	35.2286276817	25.4901960784	1
5	23	7	LEFT	-2.4105846882	36.8828136921	31.3725490196	1
6	23	8	LEFT	0.377879858	42.7328455448	52.9411764706	1
7	33	14	LEFT	2.8492805958	41.6025626659	90.1960784314	2
8	33	14	LEFT	-0.0861425623	44.1371946558	66.6666666667	0
9	33	19	LEFT	1.5049118996	51.9162573814	100	2
10	33	15	LEFT	0.7750006318	38.6788834929	78.431372549	1
11	33	17	LEFT	-1.6048058271	38.0446709394	62.7450980392	1
12	33	18	LEFT	1.0724961758	33.8855040073	68.6274509804	1
13	33	12	LEFT	1.6867015362	35.8480808735	33.3333333333	1
14	33	9	LEFT	-2.0258710384	36.4547727108	56.862745098	1
15	33	10	LEFT	-1.1335554123	39.7415418625	45.0980392157	1
16	33	16	LEFT	-2.4186251163	37.5666124821	35.2941176471	1
17	37	6	LEFT	1.6728092432	46.331043601	100	2
18	37	7	LEFT	0.7312539816	46.2715002298	100	2
19	37	10	LEFT	-0.1631172448	45.8127563745	100	2
20	44	17	LEFT	-3.4981474876	42.6571631432	54.9019607843	0
21	44	14	LEFT	-2.8951814175	40.328881979	33.3333333333	1
22	44	16	LEFT	-1.5763392448	38.3067507744	74.5098039216	1
23	49	8	LEFT	5.3984632492	32.6569643021	52.9411764706	1
24	49	10	LEFT	3.3188011646	35.0862739086	66.6666666667	1

Figure 11: Knowledge discovery extension of the DCV user interface

[1] Chronis, Y. et al. A Relational Approach to Complex Dataflows. in Proceedings of the Workshops of the EDBT/ICDT 2016 Joint Conference (2016).

6.1 Back-end functionalities and architecture

The main architecture of MD-Paedigree's infrastructure extended DCV module is illustrated in the figure below. The user interface mentioned previously refers to the *front-end* section while the *back-end* consists of EXAREME's worker, named *madIS*, enhanced with extra python modules for the knowledge discovery extension. The open source python library, *scikit-learn*[2] as been incorporated which contains a lot of well-established machine learning algorithms and techniques implemented in python libraries. These libraries can be easily imported on top of EXAREME as specific User-Defined Functions (UDFs) of the madIS system[3]. Such UDFs, construct predictive or clustering models, estimate new values for unlabelled data, and categorize new data samples [D16.1, "First report on Biomedical knowledge discovery and simulation for model-guided personalized medicine"].

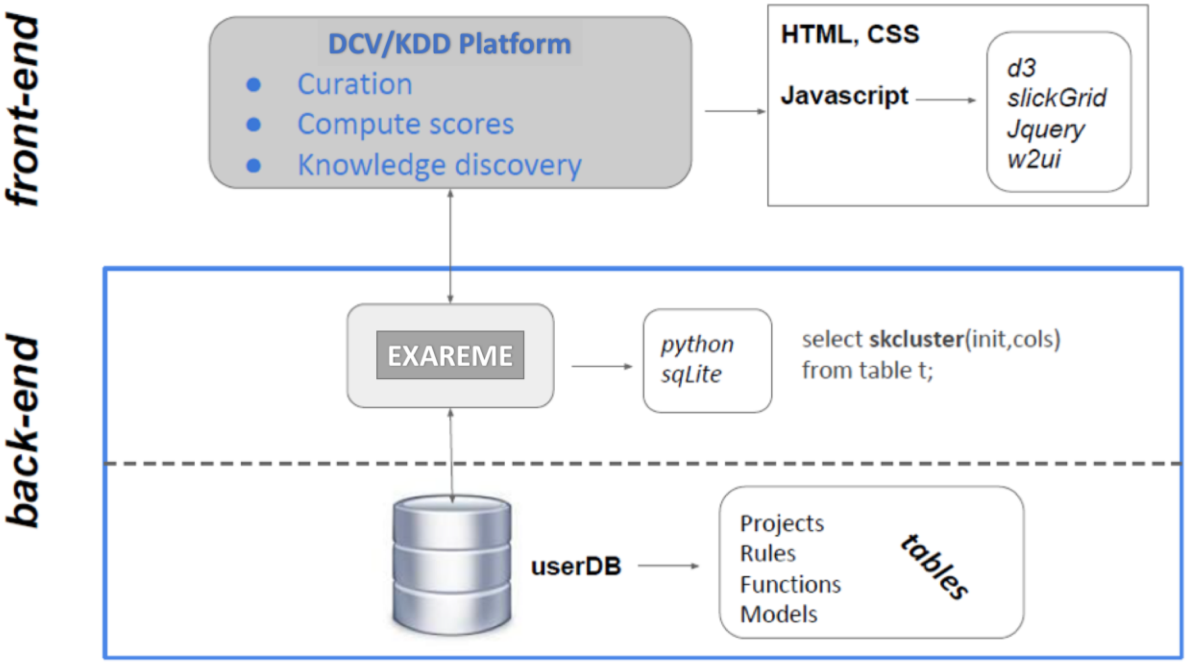


Figure 12: Comprehensive description of extended DCV architecture

6.2 EXAREME

EXAREME[4] which has been developed integrating Athena Distributed Processing Engine (ADP) with madIS extensible relational data analysis system is an open source project supported by the MaDgIK group at ATHENA.

EXAREME offers a declarative language which is based on SQL with user-defined functions (UDFs) extended with parallelism and data pipeline primitives. It is separated into the following components: The *Master* is elected from the worker pool and is the main entry point, through the gateway, to the system. The Master is responsible for the orchestration of all the components. The *Execution Engine* communicates with the resource manager and schedules the operators of the query respecting their dependencies in the dataflow graph and the available resources. It also monitors the dataflow execution and handles failures. All the information related to the data and the allocated resources is stored in the Registry. The *Resource Manager*

[2] http://scikit-learn.org/stable/

[3] https://github.com/madgik/madis

[4] http://www.exareme.org/

is responsible for the allocation and deallocation of resources on each node. The *Optimizer/Scheduler* engine translates a high-level query into the distributed machine code of the system and creates the final execution plan by assigning operators to workers. Finally, the *Worker* executes operators (relational operators and UDFs) and transfers intermediate results to the master. MadIS is the core engine of the Worker. MadIS is a wrapper of SQLite based on the python APSW. It processes the data in a streaming fashion and performs pipelining when possible, even for UDFs. The UDFs are executed inside the database along with the relational operators to push them as close to the data as possible.

EXAREME offers a relational processing engine able to support scalable distributed execution of complex, resource, and time-consuming data processing flows mainly related to data mining and decision support. In addition (and if this is ever required), data mining algorithms can be implemented with EXAREME in a *privacy-preserving* way, transmitting only aggregated hospital data (sufficient statistics). It currently supports the following functionalities for supporting distributed data mining algorithms in a privacy-preserving way, i.e. transmitting only aggregated hospital data (sufficient statistics):

1. Get list of the available algorithms such as ID3 Decision Trees, K-Means, Linear Regression, Covariance Matrix, PCA, Standard Deviation, Summary Statistics.
2. Submit any of the available algorithms for execution.
3. Get the execution status of a submitted algorithm.
4. Get the execution results of a completed algorithm

In the final year of the project, an updated version of EXAREME was released, as described in our paper: I.Chronis et al., "A Relational Approach to Complex Dataflows", MEDAL 2016 (EDBT Workshop), 2016.

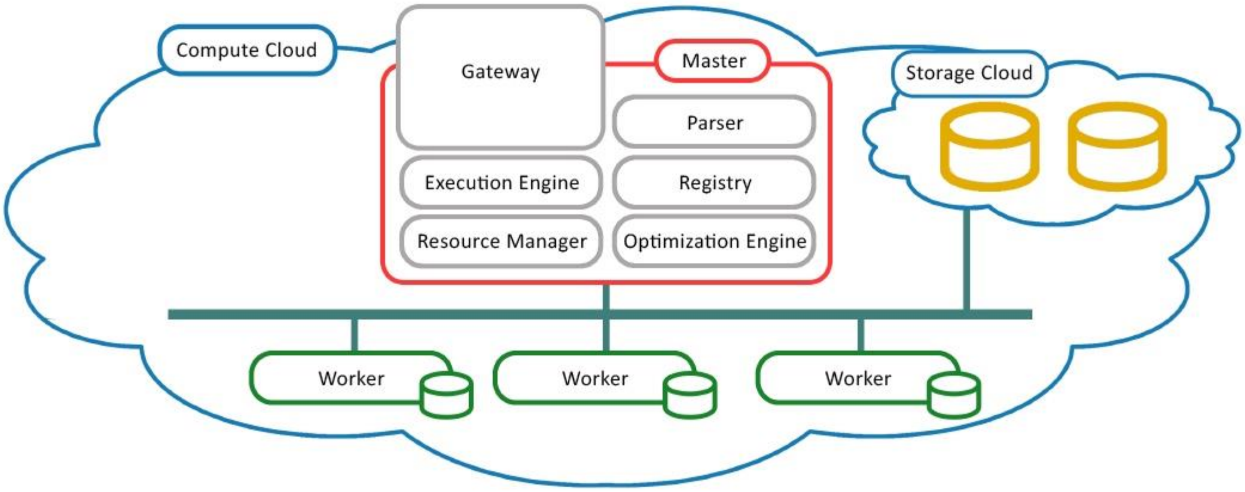


Figure 13: EXAREME's Architecture.

7 DPS

The NND work stream had some distinct challenges from the other clinical specialties. First the strategy was to use existing clinical systems where possible, through structured extracts. Centres which did not have existing clinical sources, or where they were not comprehensive enough to capture the full range of parameters, a Microsoft Excel tool was used to capture the data. In addition to the more standard clinical data this group also generated a large amount of processed outputs form their Gait analysis investigations. This took the form of both numeric values and signal segments.

Given the large heterogeneity within the intended dataset it was decided that using the toolset developed in the VPH-Share project, the DPS, might be a more practical approach. This has the advantage of only *requiring* a single data connector developed within the core infrastructure and affords more flexibility when dealing with Microsoft Excel, since the DPS is a windows and Microsoft based tool.

In addition to this more technical advantage, the semantic capabilities of the DPS also help in the data processing pipeline which is used to identify key data items, in particular signal based data, which is required for the meaningful interpretation of the data within the platform.

In the final analysis of the collected data artefacts for this clinical domain, ~2000 individual data files were produced and ingested for 497 subjects within the study and over 10,000 signals. As an additional benefit of this process, each site was left with a simple Microsoft access database containing all of their integrated data which was also expressed as a desirable requirement at the beginning of the process.

8 Data collection

8.1 Data collection issue

Data collection has encounter a huge bias in this project for many reasons. The lack of budget for buying hardware has had a negative effect, avoiding the purchasing of "on site servers" to serve as anonymization gateway and automatically push data from routine system to the repository. It has been a very long and difficult process to acquire hardware dedicated to the project in the centres.

But even when servers have been provided, the data access has generated problems. Security policies and legal aspects have broken the original process aiming to get non-anonymised data from routine systems, anonymise on site then automatically export to repository.

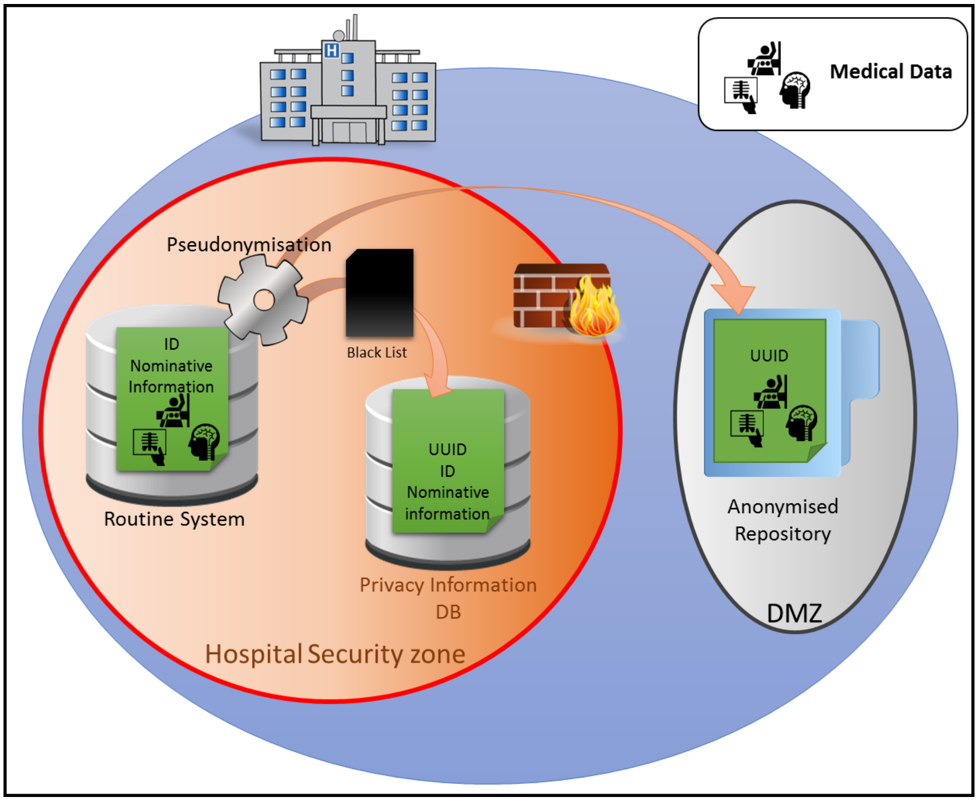


Figure 14: Data collection challenge

8.2 Data collection solution

To address the problem and make it possible to get data, a “File sharing system”, a type of file repository where any data provider could put his already anonymised data, has been put in place. The data is then taken back to the hospital on a node composing the repository and the originally planned process can take place.

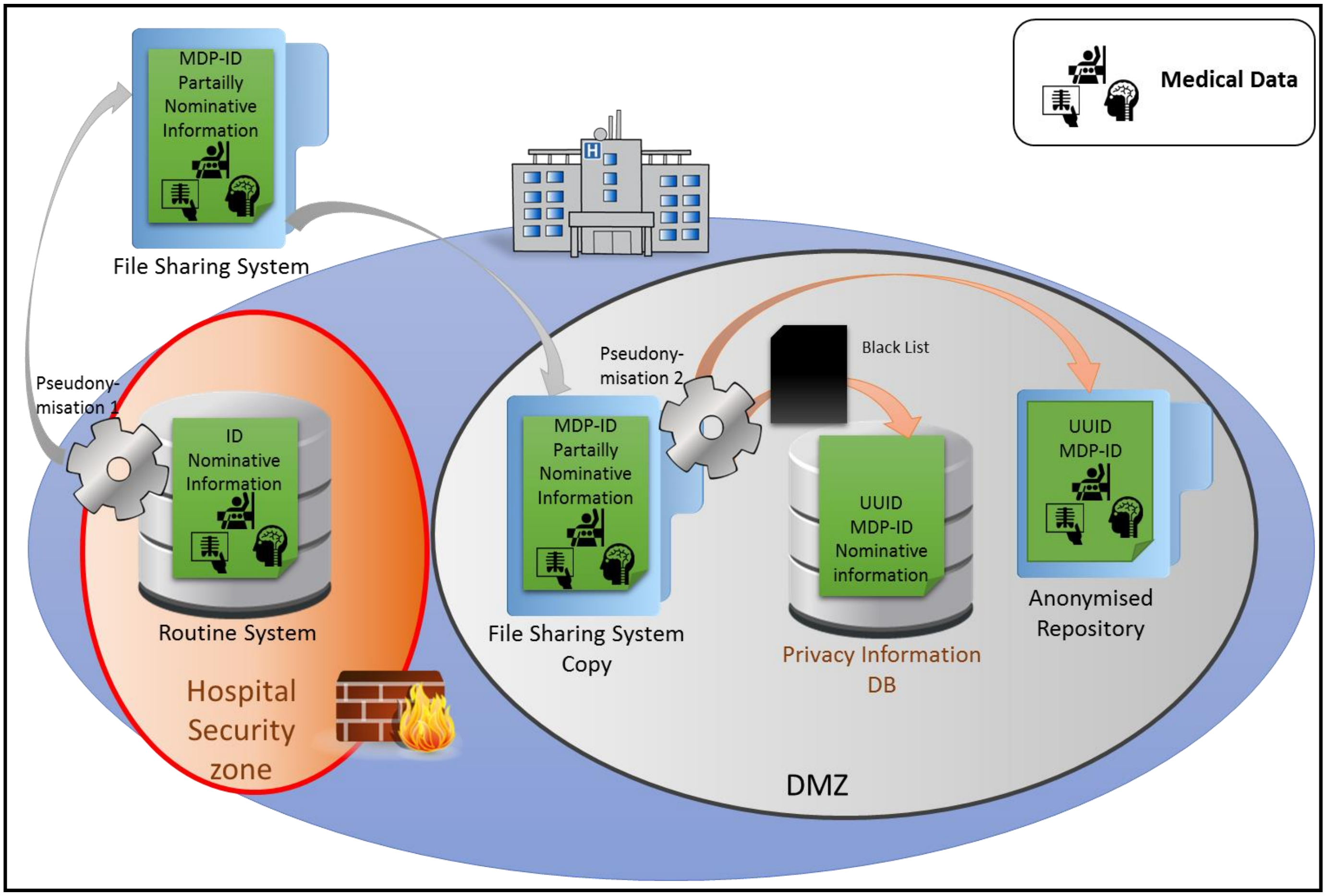


Figure 15: Data collection solution

8.3 Encountered issues

This solution has unlocked the data sharing problem but is far less optimal than what was aimed for. This is a list of encountered problems:

- Some chosen MDP-IDs were not anonymous (Patient Initials + Birthdate). The IDs had to be replaced with anonymous IDs.
- Some chosen MDP-IDs were entered by humans and were sometimes mistaken (OPBG-003 ~ OBG03 ~ OPBG 03…) resulting in additional overhead to double check and create a specific manual process to ensure the consistency of the patient data.
- The File Sharing system was physically over the internet and introduced a security issue to be addressed.
- File transfer had to happen twice through an internet connection instead of once into a local network.
- Even if the on site pseudonymisation was respecting the provider national rules, it was not enough secure to respect the global project anonymisation guidelines, with the implication that we had to re-anonymise data with more data loss as far as the patient information was not available to identify the data to be removed finely.
- The pseudonymisation has been processed by a physician from each group, the ITs have not been implied into the process and could not ensure the persistency of the back link in time.

8.4 Lessons

In future projects, particular attention will be paid to the feasibility of the data collection. Most of partners considered that once the data is acquired, the process is over, but acquisition is only the first step of the global long process.

The solutions that have been provided for this project are fine for today but since 05/2018, with the introduction of the GDPR, all countries will have to follow some more constrained rules, and this implies that prior to doing anything we will need to ensure that the aimed process will be followed.

Most of the deadlocks have come from the administrative rules of each hospital; this element should be taken into account as far as it could totally block a whole project. In case of MD-Paedigree, it has resulted in a huge delay at the beginning of the project and a large amount of unplanned work for WP14.1.

8.5 Imported data for MD-Paedigree

As of today, imported data represent:

- More than 1000 Medical Events
- More than 700 DICOM series
- More than 22 000 DICOM Studies
- More than 1 400 000 images
- More than 500 Patients
- More than 600 GBytes

9 Conclusion

The Infostructure is released and up and running. All functionalities are in place. Tests and fixes of all these applications have been achieved (see D17.3). This makes the infostructure status “corresponding to the planning” despite the data collection issues encountered during the whole project that have consequently slowed down its advancement. From a hardware standpoint, the current installation matches the original target and the distribution of the repository over 5 physical sites provides a good demonstration of the feasibility of data distribution using current network limitations.