Project Number:	604102	Project Title:	Human Brain Project

Document Title:	Medical Informatics Platform v1 – Specification Document
Document Filename[(1)]:	SP8_D8.6.1_Resubmission FINAL.docx
Deliverable Number:	D8.6.1
Deliverable Type:	Specification document
Work Package(s):	8.1 - 8.2 - 8.3 - 8.4 - 8.5 - 8.6
Planned Delivery Date:	M 6/31 March 2014
Actual Delivery Date:	M 7/25 April 2014 RESUBMITTED M28/07 Jan 2016

Author:	Ferath KHERIF, P23	*Data Interactive Analysis, Overall MIP specifications, Biological signatures of disease*
Contributor:	Anastasia AILAMAKI, P01	*Local Data layer, Query Engine specifications*
Contributor:	Thomas HEINIS, P01	*Local Data layer, Overall MIP specifications*
Contributor:	Manos KARPATHIOTAKIS, P01	*Local Data layer, Query Engine specifications*
Contributor:	Xuesong LU, P01	*Local Data layer specifications*
Contributor:	Vasilis VASSALOS, P03	*Local Data layer, Data Integration*
Contributor:	Tassos VENETIS, P03	*Local Data layer, Data Integration*
Contributor:	Jing CUI, P23	*Biological signatures of disease*
Contributor:	Mihaela DAMIAN, P23	*Timeline specifications, Use cases*
Contributor:	Bogdan DRAGANSKI, P23	*Web Portal, Biological signatures of disease*
Contributor:	Richard FRACKOWIAK, P23	*Biological signatures of disease*
Contributor:	Harry DIMITROPOULOS, P37	*Data Federation specifications*
Contributor:	Herald KLLAPI, P37	*Data Federation specifications*
Contributor:	Omiros METAXAS, P37	*Data Federation specifications*
Contributor:	Lefteris STAMATOGIANNAKIS, P37	*Data Federation specifications*

Contributor:	Eleni ZACHARIA, P37	*Data Federation specifications*
Contributor:	Paolo BOSCO, P40	*MIP Training and Support Centre*
Contributor:	Alberto REDOLFI, P40	*MIP Training and Support Centre*
Contributor:	Tal GALILI, P55	*Web Portal, Data Mining*
Contributor:	Mira MARCUS-KALISH, P55	*Web Portal, Data Mining*
Contributor:	John ASHBURNER, P70	*Image Feature Extraction, Data Mining*
Editor:	Tea DANELUTTI, P23	*Overall MIP specifications, Editing*

Abstract:	This document describes the functionalities and implementation strategy of the Medical Informatics Platform (MIP). The MIP is one of six ICT platforms that will be delivered by the Human Brain Project. MIP users will be able to query large volumes of clinical data from hospitals and research databases without moving them or compromising patient privacy. The tools SP8 provides will allow discovery of meaningful biological interactions that explain etiological, diagnostic, and pathogenic observations and treatment effects. The identification of biological signatures of diseases will enable the development of new biologically grounded classifications of brain diseases, leading to a new systematic understanding of their causes, and new diagnostic tools. These will not only facilitate early diagnosis – a precondition for effective treatment – but will also help to optimise the selection of patients for clinical trials.

Keywords:	Medical informatics, brain disease, data federation, data integration, query engine, biological signature of disease, data mining.

Document Status

Version	Date	Status	Comments
0	19.02.2014	Draft	Inclusion of contributions (until 19.02.2014) from CHUV, UCL, EPFL, AUEB, UoA, FBF
1	24.02.2014	Draft	Inclusion of additional contributions (until 23.02.2014) from TV + TH/XL
2	26.02.2014	Draft	Changes in chapter 3 and contributors
3	26.02.2014	Draft	Various modifications by TV, EZ, TH
4	27.02.2014	Draft	Contributions accepted
5 to 13	from 28.02.2014 to 05.03.2014	Draft	Iterative changes from SP8 team members
14	06.03.2014	Draft	Document formatted, ready to be circulated to the HBP Scientific Writing Team and HBP S&T Office
15	19.03.2014	Draft	Included modifications by UoA
16	20.03.2014	Draft	Included modifications by AUEB and Tel Aviv University
17	24.03.2014	Draft	Integration of 1st-round comments from the Editorial office and the S&T office
18	01.04.2014	Draft	Integration of 2nd-round comments from the Editorial office and the S&T office
19	01.04.2014	Draft	Check and copy-editing (RF)
20	03.04.2014	Final	Final formatting and proof-reading (TD)
21	04.04.2014	Final edit	CL
22	07.04.2014	Final edit	RW

23	08.11.2015	Update according to the EC reviewers recommendations	FK Inclusion of contributions from CHUV, EPFL and AUEB. The following section have been added or updated. 3.4.2 Reproducibility/Provenance 3.8 The MIP: Relations to Other HBP Platforms 3.9.1 Data Providers recruitment 3.9.2 Deployment 3.9.3 Standardization of data and common ontology. 3.10 Risk analysis and management 7. Functions 7.2 WP 8.2 - Clinical/Research data analysis and preparation – on representative/ extended datasets - Clinical/Research data analysis and preparation – automation of processes 7.4 WP8.4 prototype/internal release prototype public release Acquisition of standardised federation software Detailed technical MIP architecture and component integration plan MIP Operational processes Project management of the integration
24	22.12.2015	Update according to STO recommendations	FK Update 3.8 The MIP: Relations to Other HBP Platforms (SP8/SP6/SP13) 3.9.3 Standardization of data and common ontology. 3.10 Risk analysis and management

Table of Contents

List of Figures & Tables

1. Executive Summary

Subproject 8 of the Human Brain Project (HBP) will deliver the Medical Informatics Platform (MIP) – one of six ICT platforms dedicated to brain research. The MIP will be accessible via the HBP Collaboratory by the end of the 30-month Ramp-Up Phase. The present document describes the functionalities and implementation strategy of the MIP.

The MIP will transform medical records into research data, extract knowledge and build models of brain diseases. Inference from these models will advance the field of medicine from descriptive symptomatology toward diagnostic, predictive and prescriptive personalised medicine.

The system will provide dedicated services for researchers to carry out neuroepidemiological and biological investigations on clinical data. It will enable researchers to query data without moving them from their storage sites and without compromising patient privacy. The information made available will include brain scans of various types; data from electrophysiology, electroencephalography and genotyping; data from validated clinical instruments; and metabolic, biochemical and haematological profiles.

The MIP functionality will be implemented on a 3-layer infrastructure, which will deliver a highly flexible and scalable solution for near real-time processing and dynamic analyses of data.

The MIP will federate data from multiple sources to allow researchers to interrogate them in a unified, transparent and efficient way. The privacy of data will be strictly maintained through rigorous anonymisation and access rules.

Web-based technologies adapted for neuroscientific, clinical, genetic, epidemiological and pharmacological users will deliver rich research tools and data queries. The Subproject will also implement high-performance computing tools from descriptive statistics to automated data mining.

Machine learning and data mining tools will provide a comprehensive classification of brain diseases based on biological signatures, i.e. parameterised combinations of biological features and markers. The biological signature of brain diseases will form the basis for a new multi-dimensional brain disease space, facilitating scientific investigation and permitting personalised medicine.

2. Introduction

The Human Brain Project (HBP) is a response to the fragmentation of brain research and the data it produces. In order to accelerate the pace of brain research, the HBP will provide an integrated system of ICT platforms offering services to neuroscientists, clinical researchers and technology developers. The Medical Informatics Platform (MIP), which is being developed in Subproject 8 (SP8), is one of five ICT platforms that will be set up in the Ramp-Up Phase of the HBP.

Today, medical researchers lack data and tools to understand the causes of brain diseases and to develop new treatments. To address these issues, the MIP will provide the research community with tools for epidemiological exploration, numerical and statistical analysis, data visualisation and data mining. The goal of the MIP is to federate imaging, genetic and other clinical data currently locked in hospital/research archives and databases while

guaranteeing protection for sensitive patient information. These resources will permit the identification and constant updating of unique biological signatures of brain diseases. These will be used for diagnosis, more accurate prognosis and new types of drug discovery for the development of new medicines. The services of the MIP will be accessible through the HBP Collaboratory.

This document describes the architectural specifications of the MIP. It is a collaborative effort of the partners involved in SP8.

Section 3 presents a general description of the MIP. Through use cases it also gives an overview of the research services that will be provided at the end of the Ramp-Up Phase: a) Exploration, b) Interactive Analysis and c) Biological Signatures of Diseases. The following sections explain in detail the three component layers that will support the research services and their interaction: the Web Portal layer (Section 4), the Data Federation layer (Section 5) and the Local Data layer (Section 6). Finally, Section 7 and Section 8 summarise the overall implementation plan and Key Performance Indicators.

3. The Medical Informatics Platform (MIP)

3.1 The MIP: Overall Goals

The goal of the MIP is to allow researchers to identify biological mechanisms that explain the complex nature of brain disease. The MIP will provide end-to-end solutions ranging from data to advanced analytical tools. Researchers will be able to investigate questions requiring data correlations, distributions and interactions in the context of disease processes and epidemiological factors. Using analytical tools provided (and eventually included as personal APIs) researchers will be able to decipher the relationship between biological variables and clinical phenotypes. Simultaneously, as data accrue and new hospitals and data generators are recruited, data mining tools will explore all the data to detect recurrent patterns and identify biological signatures of disease. The biological signatures of disease will form the basis for a new disease space that neuroscientists and clinicians can explore.

The MIP will build on public and research databases and hospital data federated by novel data management and query techniques. This federation software and hardware will allow researchers to query and analyse a very large volume of data without moving them from local servers and without compromising data privacy.

3.2 The MIP: Services and Users

3.2.1 Services

The MIP will provide the following services:

- **Epidemiological Exploration** (in the broadest sense) of clinical data available at hospitals and disease mapping of biological variables.
- **Interactive Analysis** of new models/methods, and objective diagnoses and treatment of brain diseases.
- **Identification of Biological Signatures of Diseases** using data mining algorithms to facilitate drug target discovery and diagnostics.

The user will need to work across services to extract the information required. For example s/he might start with the Epidemiological Exploration or Biological Signatures of Diseases services, and following the initial results returned by the system, s/he might decide to view further specific details at local hospitals and therefore enter the Interactive Analysis service. 9The use cases in Section 3.3 will explain in more detail this interaction between services.

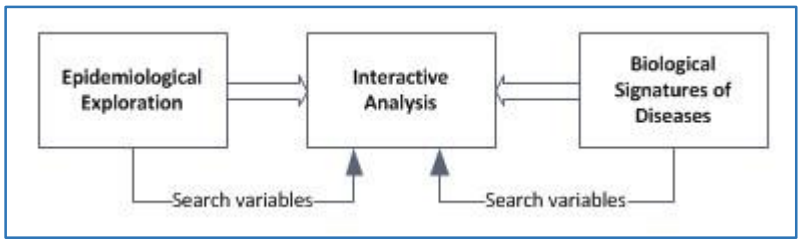

Figure 1: Services interaction at the MIP

3.2.2 User Roles

There are three user profiles in the MIP. Each type of user will have access to all three services described above.

3.2.2.1 End User

General User (GU) – User with scientific expertise in medical/neuroscience domains and low technical expertise. The category includes: clinicians, neuroscientists, epidemiologists and pharmaceutical R&D researchers.

Developer User (DU) – User with development skills in scientific and statistical software. The category includes: scientific/statistical developers, computational neuroscience developers.

3.2.2.2 Admin User

Admin User (AU) – User with extended permissions. The category includes: platform administrators, platform developers and database managers.

3.2.2.3 Data Provider

Data Provider (DP) – User with access and permission rights to the hospital databases. The category includes: hospital management, hospital IT management, hospital database managers and research database owners.

3.3 The MIP: Use Cases

The Use Cases below are Use Cases for the MIP, which SP8 will address.

3.3.1 Use Case 1 – Epidemiological Exploration

Use case number:

SP8-UC-001

Primary actor:

Alan is a GU. He is an epidemiologist. **Preconditions:**

- The clinical disease classification Variables (e.g. ICD-10) have been released and are available at local data sources.
- Metadata describing the local data-source Variables are available at the MIP Web Portal.

Success scenario:

1) Alan wants to write a report for the World Health Organisation on Alzheimer's disease. He is interested in estimating the incidence rate (i.e. the number of cases) of the disease in a sampled population in Europe for demographic factors.
2) Alan logs into the Collaboratory.
3) He selects the Epidemiological Exploration service and uses the interface.
4) He uses the multiple-choice drop-down menu to select "Alzheimer's disease" from the ICD-10 classification.
5) Alan narrows his research to the specific geographical region and to demographic factors.
6) The interface displays a graph/table output with the number of cases of Alzheimer's disease patients in Europe for the demographic factors.

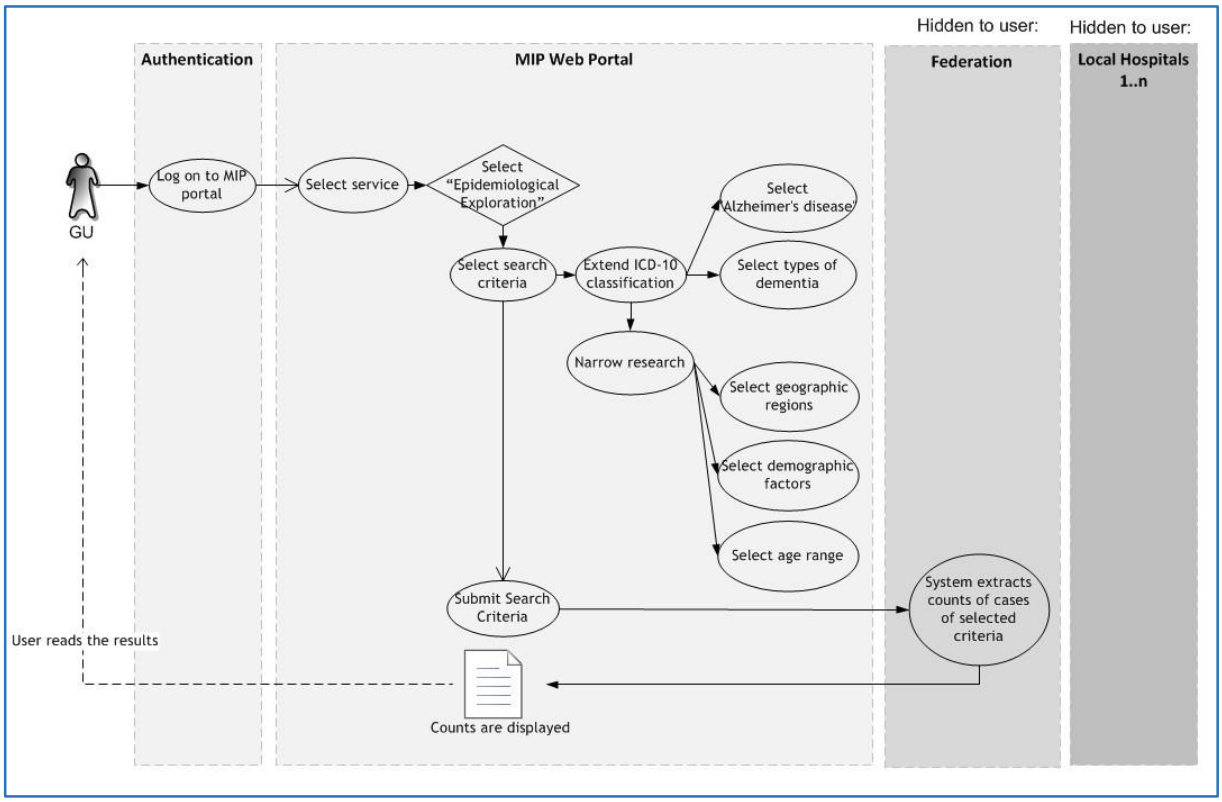

Figure 2: Use Case 1 & Use Case 2

3.3.2 Use Case 2 – Epidemiological Exploration

Use case number:

SP8-UC-002

Primary actor:

Beth is a GU. She is a clinician in neurology. **Preconditions**:

- The clinical disease classification Variables (e.g. ICD-10 World Health Organisation International Statistical Classification of Diseases and Related Health Problems 10th Revision; code adopted by hospitals) have been released and are available at local data sources.
- Metadata describing the local data source Variables are available at the MIP Web Portal.

Success scenario:

1) Beth is interested in carrying out a study on dementia. First, she wants to know the number of cases for different types of dementia in Europe: dementia in Alzheimer's disease with early and late onset, vascular dementia and dementia in Pick's disease.
2) Beth logs into the Collaboratory.
3) She selects the Epidemiological Exploration service and uses the interface.
4) She uses the multiple-choice drop-down menu to select the different types of dementia from the ICD-10 classification.
5) Beth narrows her research to the geographical region and age ranges.
6) The interface displays a graph/table output with the number of cases for each type of dementia and age range.

3.3.3 Use Case 3 – Interactive Analysis

Use case number:

SP8-UC-003

Primary actor:

Charlotte is a GU. She is a neuroscientist. **Preconditions**:

- The brain "grey matter volume." Variables have been produced at the local level, released and are available through the MIP Web Portal.
- Metadata describing the local data source Variables are available at the MIP Web Portal.

Success scenario:

1) Charlotte wants to study the relation between the mean volume of grey matter of the hippocampi and the types of dementia.
2) Charlotte logs into the Collaboratory.
3) She accesses the Epidemiological Exploration service (as per Use Cases 1&2) where she selects the Variable "Volume of grey matter of hippocampi" in the cases of dementia.
4) The system retrieves the summary counts, i.e. the number of records for the selected Variables, and displays it.
5) Charlotte views the results, and she decides that the sample size is sufficient to examine further the values for a particular type of population.
6) She selects the Interactive Analysis service and chooses the statistical analysis for comparing differences in grey matter volume for the sub-types of dementia.
7) She retrieves the mean volume of grey matter for populations of interest and their standard deviations.

8) Charlotte can then use these measures and compare them to the values that she observed in her clinic, and evaluate whether they are typical or unusually high/low.

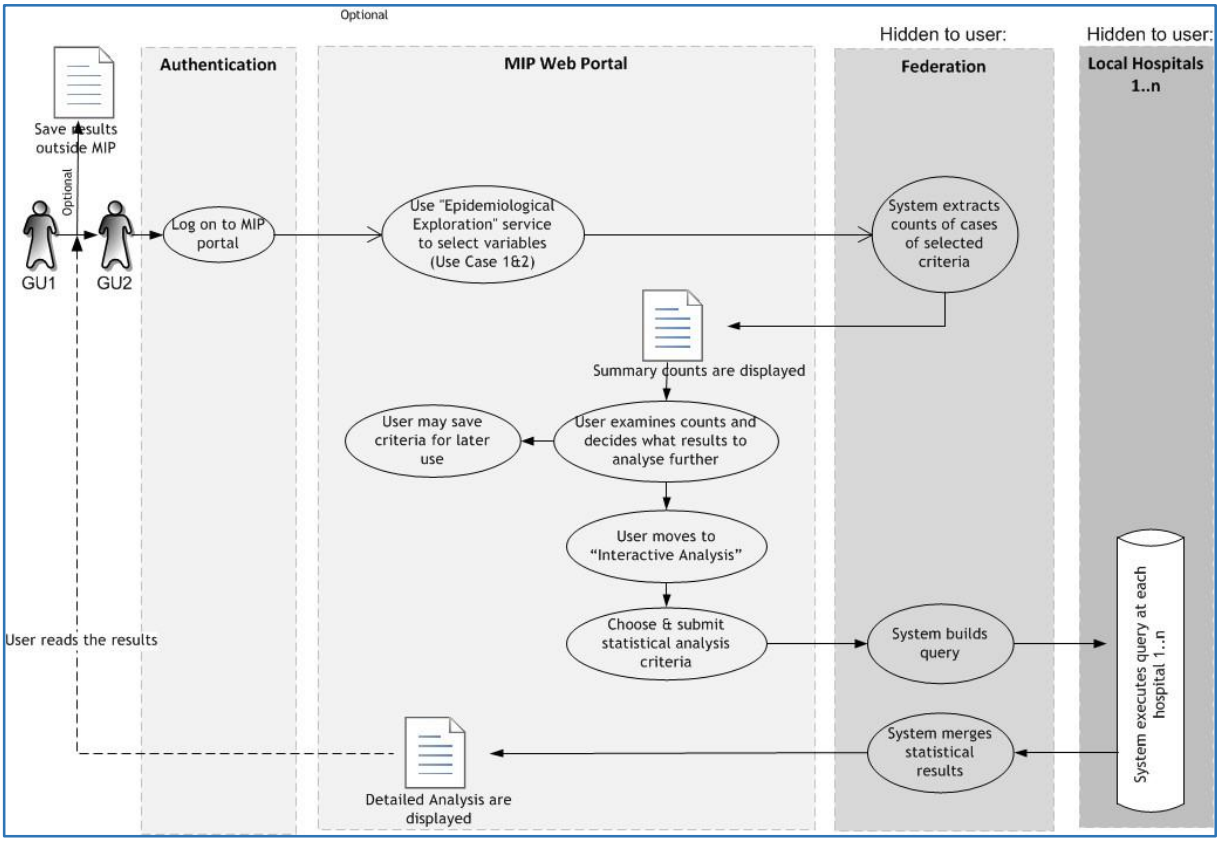

Figure 3: Use Case 3 & Use Case 4

3.3.4 Use Case 4 – Interactive Analysis

Use case number:

SP8-UC-004

Primary actors:

- Charlotte is a GU. She is a neuroscientist expert in neuroimaging.
- Mike is a GU. He is a neuroscientist expert in genomics. Preconditions:
- The brain "grey matter volume" Variables have been released and are available at the local data source.
- The "ApoE gene" Variables have been released and are available from the local data source.
- Metadata describing the local data source Variables are available at the MIP Web Portal.

Success scenario:

1) Charlotte and Mike want to collaborate on a project for understanding the link between genetic and brain feature in dementia.

2) Charlotte and Mike log into the Collaboratory.

3) Charlotte accesses the Epidemiological Exploration service (as per Use Cases 1&2) where she selects the Variable "Volume of grey matter of hippocampi" in the cases of dementia. With Mike's expertise in genetics, she is now able to also select the "ApoE genetic" phenotype.

4) The system retrieves the summary counts, i.e. the number of records for the selected Variables, and displays it.

5) Charlotte and Mike view the results and they decide to examine further the values for a particular type of population.

6) They select the Interactive Analysis service and choose the statistical analysis for comparing differences in grey matter volume for the sub-types of dementia and the "ApoE genetic" phenotype. They can then assess the replicability of the results on another subset.

7) Charlotte and Mike save their results in a common project folder in the Collaboratory.

3.3.5 Use Case 5 – Interactive Analysis

Use case number:

SP8-UC-005

Primary actor:

John is a DU. He is a scientific developer.

Success scenario:

1) John developed new tools for combining genetic information and brain volume.

2) John logs into the Collaboratory.

3) John uploads his script and tests the script in the sandbox.

4) The script is tested and validated by the Admin User's group. The script is then made available to the GU.

5) The script is made available to all users in the Interactive Analysis service.

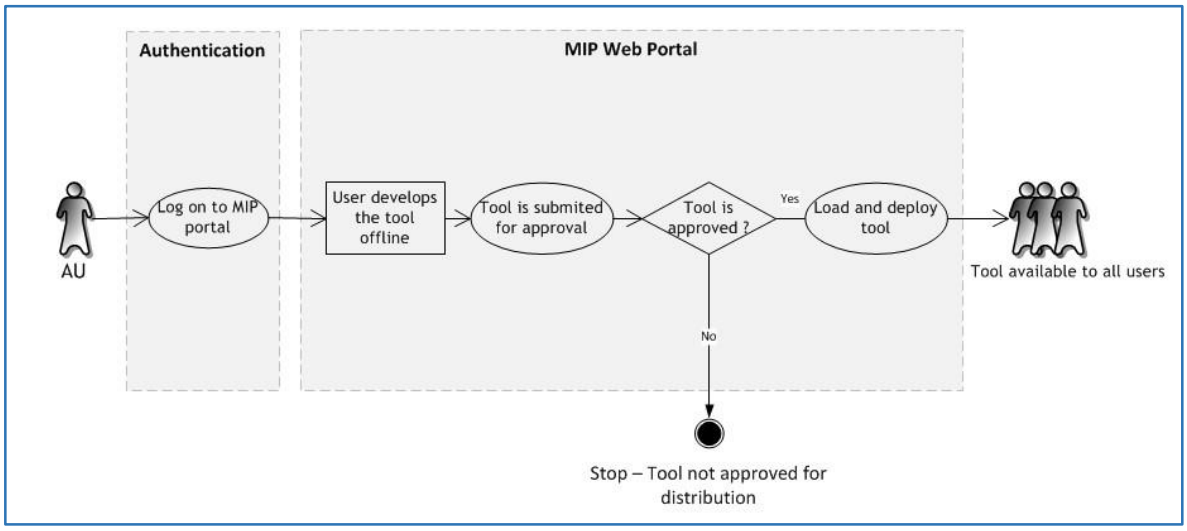

Figure 4: Use Case 5

3.3.6 Use Case 6 – Biological Signatures of Diseases

Use case number:

SP8-UC-006

Primary actor:

Beth is a GU. She is a clinician in neurology. **Preconditions**:

- The biological signatures of diseases produced by the data mining algorithms are available at the MIP Web Portal.
- The Variables that describe each disease signature cluster have been released and are available at the MIP Web Portal.

Success scenario:

1) Beth is interested in taking forward personalised diagnostics using the biological signatures of the disease.

2) Beth logs into the Collaboratory.

3) She selects the Biological Signatures of Diseases service and uses the interface to classify her own patient by comparing his clinical and biological characteristics with the whole range of provided biological signatures of diseases using an optimal matching algorithm.

4) She does this by selecting Variables of interest – e.g. demographic data, blood cholesterol, neuropsychological scores, genetic burden, etc.

5) She enters values for those Variables.

6) She retrieves a list of disease signatures ordered according to the best match. The distribution of values of the other unselected Variables is also displayed along with their uncertainty – e.g. genotype, clinical scores and cardiovascular risk factors.

7) She also retrieves a 3D brain map with highlighted anatomical regions affected by the particular disease corresponding to the optimally matched disease signature. She can compare the map with the anatomy pattern of her own patients.

8) Depending on how well the disease signature cluster matches her criteria, Beth can add new Variables to determine the stability of her classification in relation to the number of criteria or Variables used.

9) She can compare the derived disease signature cluster to conventional clinical classification - e.g. ICD-10, DSM V classification.

10) If needed, she can review her patients (data) to verify the derived disease signature cluster by similarity and by differences with other patients.

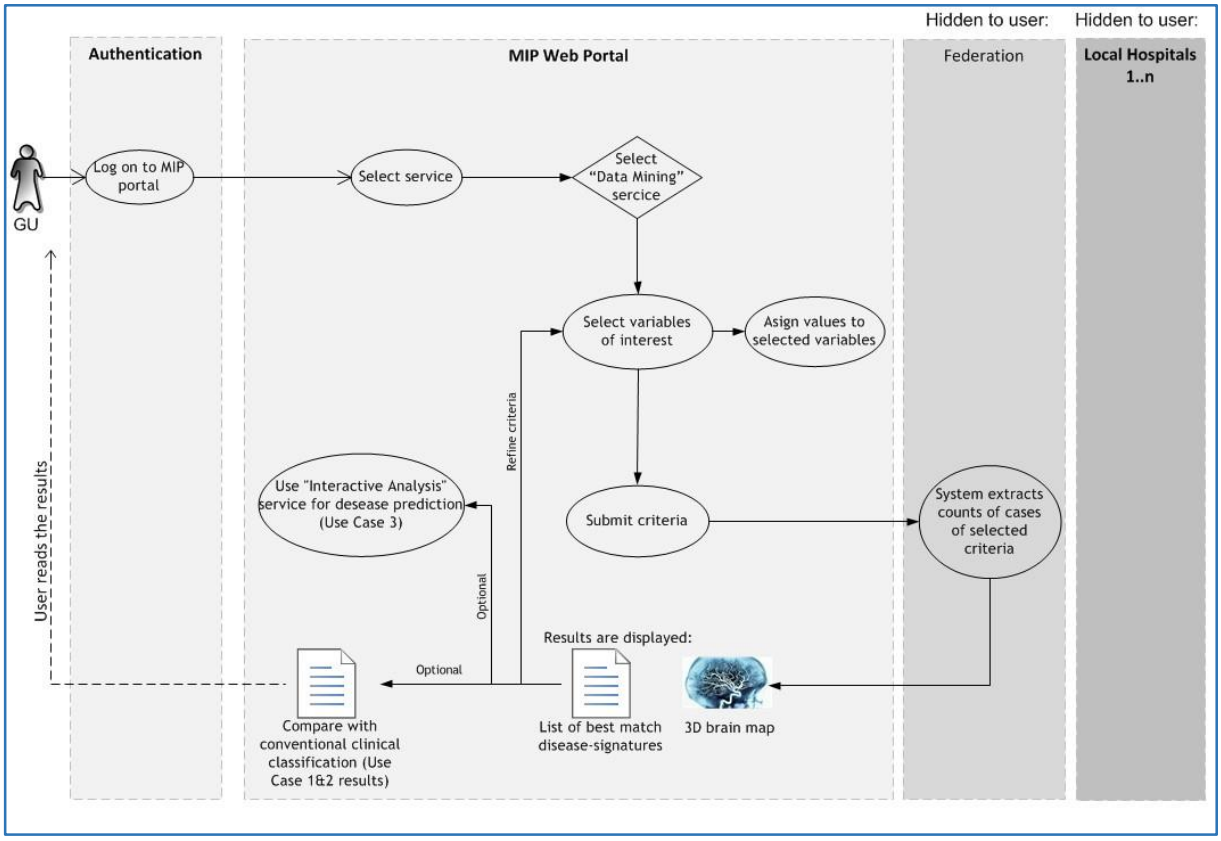

Figure 5: Use Case 6 & Use Case 7

3.3.7 Use Case 7 – Biological Signatures of Diseases

Use case number:

SP8-UC-007

Primary actor:

Nathalie is a GU. She is a researcher in pharmaceutical R&D. **Preconditions:**

- The biological signatures of diseases produced by the data mining algorithms are available at the MIP Web Portal.
- The Variables that describe each disease signature cluster have been released and are available at the MIP Web Portal.

Success scenario:

1) Nathalie is interested in defining inclusion criteria and a set of non-invasive biomarkers for a clinical trial on a new drug for Alzheimer's disease.

2) Nathalie logs into the Collaboratory.

3) She selects the Biological Signatures of Diseases service and uses the interface to retrieve the set of features of interest according to the provided disease signatures for dementia of the Alzheimer type.

4) She identifies the features leading to the creation of homogeneously stratified set of rules she would apply to create the cohorts undergoing pharmacological intervention.

5) She uses the provided predictive tools to infer potential therapeutic targets and positive as well as adverse effects based on multi-scale information – e.g. molecular pathways, proteomics interactions, genetic profiles, etc. up to the system/behavioural level.

6) She is now in a position to specify trials using well defined homogeneous, and therefore small cohorts to test the effects of a drug or cocktail of drugs that modulates the targets suggested by the rules that define her disease signature of interest.

3.3.8 Use Case 8 – Data Mining

Description:

Configuration of data mining algorithms and release of a new Biological Signature of Disease.

Use case number:

SP8-UC-008

Primary actor:

Paul is an AU. He is a platform developer. **Preconditions:**

- Paul developed, tested and validated a new functionality of the rule-based clustering algorithm on a small dataset.
- The algorithm has been deployed onto the data mining servers.

Success scenario:

1) Paul wants to run the new functionality of the data mining algorithm on the data available through the MIP. The results – the Biological Signature of Disease – will then be available for other users.

2) Paul identifies the input data for the algorithm. He sets research criteria using the Ontology, Variables and Provenance descriptions.

3) Paul sets up the parameters of the data mining servers to execute the algorithm on the local data.

4) Paul runs the algorithm.

5) The results - the Biological Signature of Disease – are stored in the system.

6) Paul explores the rules on the variables that the algorithm has returned as biological signatures.

7) Paul registers the rules of the Biological Signature of Disease as new Variables in the Variables descriptions. He also describes the method used to obtain them in the Provenance database.

8) Paul validates the Biological Signature of Disease and releases it. The Biological Signature of Disease can now be used by other users.

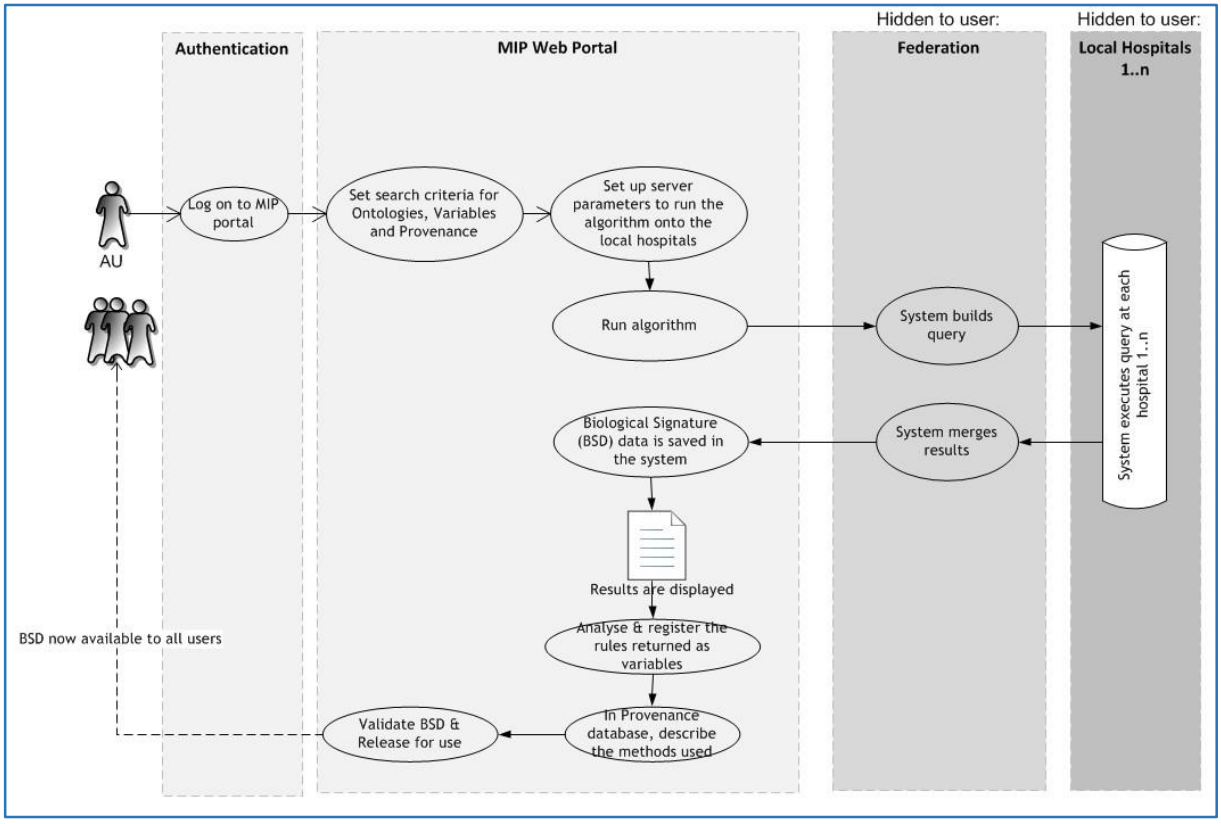

Figure 6: Use Case 8

3.4 The MIP: Functional Requirements

3.4.1 General Requirements

The MIP software shall allow users to access and interrogate data (web interface). It shall also provide processing engine to receive, transform and send on queries from the web interface (the federated infrastructure). Finally, it shall provide individual hospital data (Local Data Store Mirror) to which the federated infrastructure will send the transformed queries. All users will be able to access three different services within the system: Epidemiological Exploration, Interactive Analysis and Biological Signatures of Diseases.

Requirement ref.	Description	Applies to Users/Use Case
SP8-FR-001-G	The Web Portal shall provide access to the three types of services: Epidemiological Exploration, Interactive Analysis and Biological Signatures of Diseases.	End Users SP8-UC-001 to SP8-UC-008
SP8-FR-002-G	The MIP shall provide descriptions of data in a way that is both accurate and meaningful to users. This shall be done using Ontologies, Variables and Provenance.	End Users SP8-UC-001 to SP8-UC-008
SP8-FR-003-G	The MIP shall allow querying from heterogeneous and fragmented databases with no copy and no move of original data.	End Users SP8-UC-001 to SP8-UC-008
SP8-FR-004-G	Queries submitted at the Web Portal shall be split and resent to the appropriate local hospitals to gather the required information. This shall be done at the Data Federation layer.	End Users SP8-UC-001 to SP8-UC-008
SP8-FR-005-G	The system shall only query anonymised data. The anonymisation shall be made at each local hospital. Data records shall remain at local hospitals.	All users SP8-UC-001 to SP8-UC-008
SP8-FR-006-G	The MIP shall provide software solutions adapted for each local hospitals to capture data, extract features and their descriptions (Variables and Provenance)	All users SP8-UC-001 to SP8-UC-008
SP8-FR-007-G	The analysis sent by users via the web interface shall be executed at local hospitals, as a query on local data.	All users SP8-UC-001 to SP8-UC-008

Table 1: MIP general requirements

Coding format: [Subproject]_[Functional/Non-functional Requirement]_[Requirement number]_[Type/Area of Requirement], where: [Subproject] = SP8; [Functional or Nonfunctional Requirement] = FR or NFR; [Type of Requirement] = G for general MIP requirements; WeP for the Web Portal layer; DF for Data Federation layer; LD for the Local Data layer.

3.4.2 Reproducibility/Provenance

Provenance data is collected on all levels of the MIP, at the local data layer, the federation layer and the web portal layer.

- We use the PROV standard on the local and global level to store the provenance XML files. Doing so enables us to leverage available tools (e.g., visualization) developed for PROV.
- The MIP ensures reproducibility by keeping track of all anonymized input data (e.g., MRI images), intermediate data products, versions of software used (e.g., SPM) as well as what software uses and produces what intermediate or final data product.
- The MIP tracks precisely what (software and data) was used in producing a data item and, crucially, what data items were available at a given point in time.
- The MIP stores local provenance at each hospital, not all information is accessible by external users.
- At the federation and web portal level, the provenance tracks only the results on the aggregated data items. The provenance stores what data sources (LDSMs) were available when a particular query was executed and what software was used on the federation level. This provenance information is made available to a user executing queries on the MIP.

3.4.3 Privacy

All query results coming from a Local Data Store Mirror are filtered for any personal identifiers left (e.g. headers of imaging data) when they leave the Local Data Store Mirror and enter the Data Federation layer. Data security and privacy follow principles established by data governance bodies, which include SP8 team and local data providers.

Requirement ref.	Description	Applies to Users/Use Case
SP8-FR-008-G	**Role-based permissions:** Several level permissions shall be defined to access data, according to user types within the system (e.g. End Users will be able to only read information). The system shall allow specific permissions to be assigned to individual users. **Access types:** Users shall be able to read or modify tools used by the services (i.e. statistical functions and methods) according to their assigned permissions. **Audit trial:** The system shall have full audit trial capabilities to record each action performed by users (e.g. who accessed the system, on which date and time, and what actions they performed). The audit trial log can be used for debugging purposes and investigation of security breaches.	All users SP8-UC-001 to SP8-UC-008
SP8-FR-009-G	**Aggregation:** The platform shall respond to queries by only returning aggregated results and never individual patient details.	All users SP8-UC-001 to SP8-UC-008
SP8-FR-010-G	The system shall provide *k*-anonymity with groups containing less than *k* values.	All users SP8-UC-001 to SP8-UC-008

Table 2: Measures to anonymise results

3.4.4 Training and Support Centre

A dedicated support centre will be created for the MIP. Its purpose is to provide users with training and technical support in the form of: documentation, courses, web-based tutorials, demonstration sandboxes, and a dedicated web site.

Requirement ref.	Description	Applies to Users/Use Case
SP8-FR-011-G	MIP Training and Support Centre shall be used by the users to learn and ask for specific support to familiarise themselves with data analysis and data mining tools integrated into the MIP Web Portal. The MIP Training and Support Centre's strategy shall be based on a multilevel approach which relies on: 1. Workshops and tutorials for clinicians and researchers. 2. Remote demonstrations via web tools and web seminars. 3. Self-learning tools – e.g. SCORM [1] or Valamis[2]. The objective is to provide a flexible tool, fully integrated in the MIP, and tutoring users with platform usage, data analysis and data retrieval. These tasks will leverage the expertise, experience and technical solutions developed so far in the neuGRID project[3].	All users

Table 3: Training and support centre

3.5 The MIP: Non-Functional Requirements

Requirement ref.	Description
SP8-NFR-001-G	**Hospitals:** Within the first 30 months the MIP shall integrate data from five hospitals in Europe. We expect each hospital to make available data ranging from 150 to 1 000 patients.
SP8-NFR-002-G	**Data:** Resulting in a maximum of 10-100 TB (if data in all modalities are available for all patients).
SP8-NFR-003-G	**Users:** The platform shall allow hundreds of users to concurrently explore the clinical data, tens to hundreds of users to interactively apply and test new models of brain disease, and tens of users to perform data mining tasks at the same time.
SP8-NFR-004-G	**Scalability:** Hardware and software allowing high scalability shall be employed to enable concurrent and fast access to the MIP on the targeted scale (hospital data of up to 100 TB, min 100 users).
SP8-NFR-005-G	**Flexibility:** The platform shall allow easy integration of new hospitals and their data, at minimal risk and using minimal project implementation resources. All software needed by hospitals to set up the LDSM (data processing pipeline that extracts features, anonymization software, local federation software etc.) will be packaged and made available in a Virtual machine.
SP8-NFR-006-G	**Deployment/upgrage:** The platform shall allow development and easy integration of new analysis tools. **All software** will be packaged and made available in a Virtual Machine (VM) or Containers. Along with the packaged software we will made available to data providers.

Table 4: MIP non-functional requirements

3.6 The MIP: Software

The functionality of the MIP is to support the three different types of services: a) Exploration, b) Interactive Analysis and c) Biological Signatures of Diseases. Each serving the different types of users, as defined in Section 3.2.2.

The software of the MIP is functionally divided into three horizontal layers: the web interface, the federated infrastructure and the hospital data infrastructure.

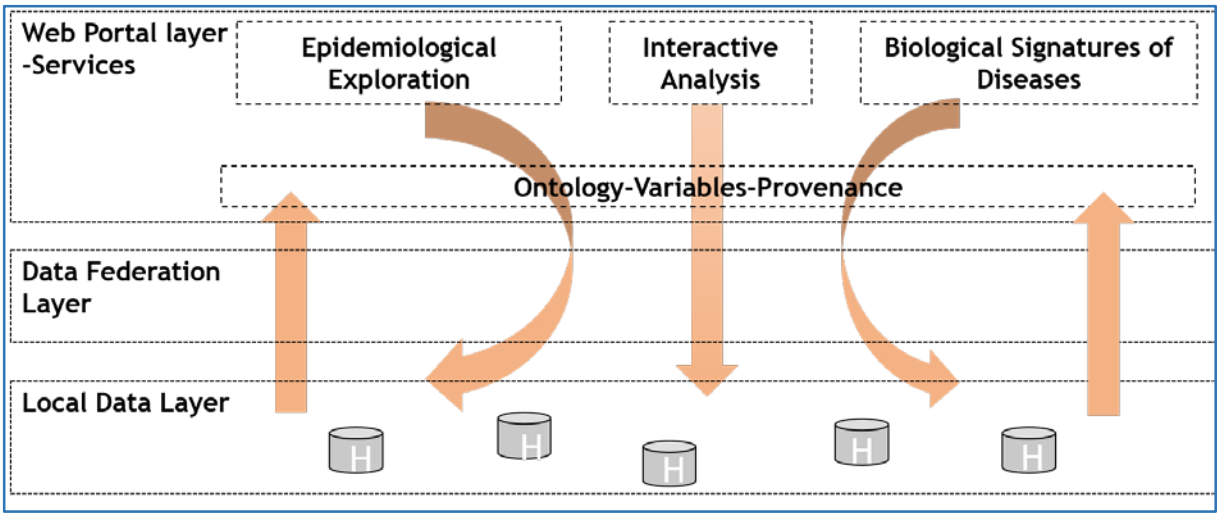

Figure 7: The three-layered MIP infrastructure

3.6.1 Web Portal Layer

Users will access the MIP via the Collaboratory shared by all the HBP platforms. User authorisations and user interface will be based on the specifications described by SP6. The MIP services will be hosted on the UP.

Users interact with services by forming queries and launching analyses, as described in the Use Cases. Forming a query is based on Ontologies, Variables and Provenance, which are stored in the Web Portal and extracted from Local Data Store Mirrors. The Web Portal also returns the provenance of Variables to users with their results.

3.6.2 Data Federation Layer

The Data Federation software hides from the users of the platform the distributed architecture, i.e. the fact that all hospital data remains at the local hospital.

The Data Federation layer receives queries from the Web Portal and forwards them to local hospitals. Queries can be either related to simple data retrieval or more complex data mining flows (see Use Cases). Based on local schemas as well as summary information, the federated infrastructure forwards queries to relevant hospitals and collates their results, returning a complete result to the web interface or to relevant portal located applications.

3.6.3 Local Data Layer

The Local Data layer is responsible for retrieving, anonymising, pre-processing and extracting features and metadata from local hospital databases at each hospital.

The anonymised data are stored at the hospital (Local Data Store Mirror), ready to be queried. It will never leave the hospital. The Local Data layer answers queries received by the Data Federation layer, but only for a subset of the data stored locally at the hospital.

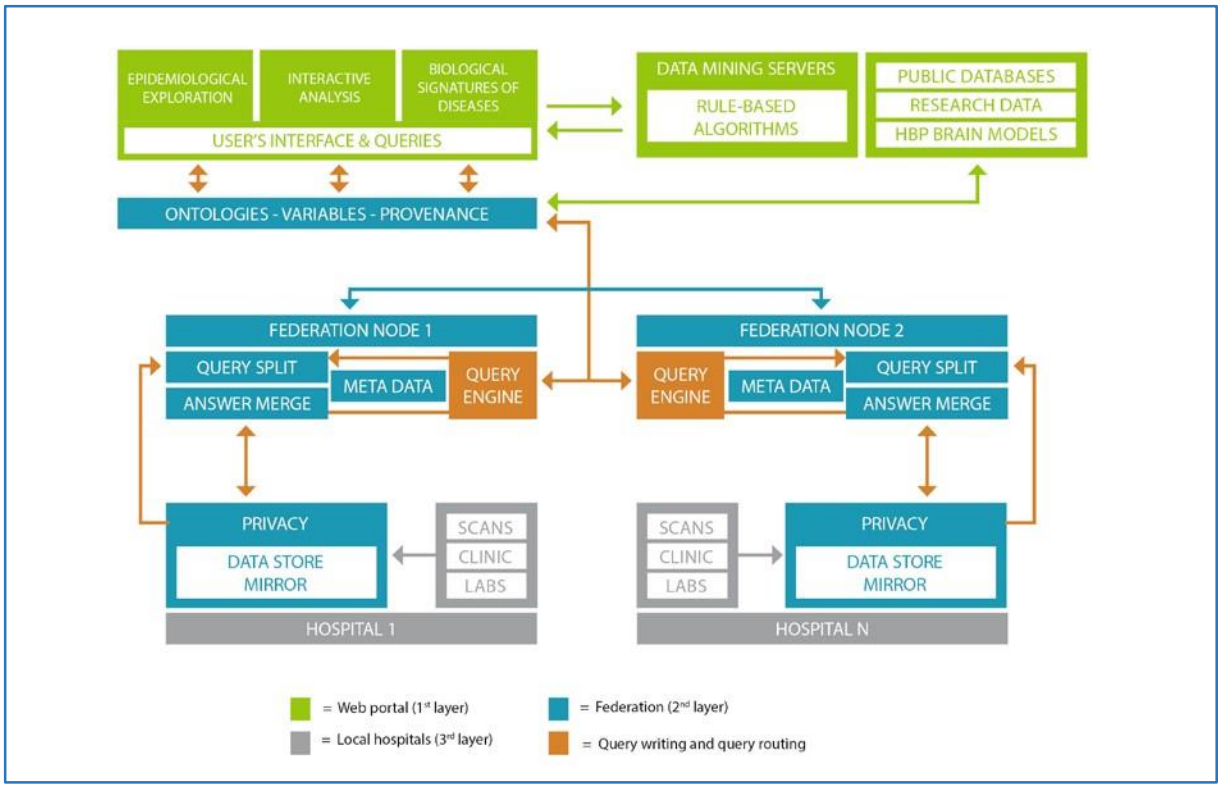

Figure 8: The detailed layers of the MIP

3.7 The MIP: Physical Architecture

The MIP specific hardware will be based on the following requirements.

- A central server with 128 GB main memory and 10 TB RAID-1 and 4 TB RAID-0 for the disk, for running services and storing the federated Ontologies, Variables and Provenance databases. The server will be connected to the UP web servers.
- Dedicated servers with a minimum of 16 cores will be used for data mining and analyses.
- A server will be hosted at each data provider location. The server will contain the Federation Node and the Local Data Store Mirror. It will be connected to the central server (see above) via internet by a T1 line with end-to-end encryption.

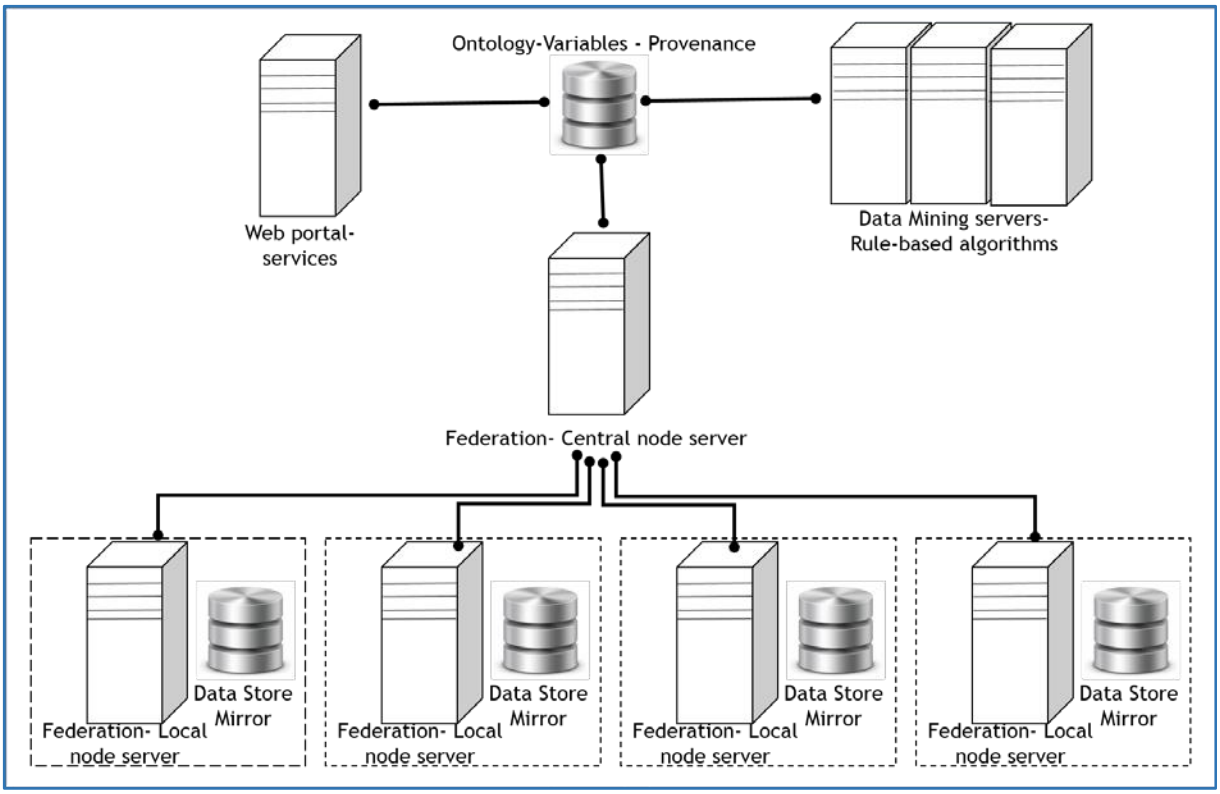

Figure 9: Physical architecture of the MIP

- We understand very well that there are clear limitations with respect to what computations can be run distributed and thus require local resources. We will make available to the hospital recommendations (primarily computational power and storage) for the hardware needed depending on the amount of data stored locally. The recommendations are driven and defined by what data analytics algorithms are expected to be run distributed in the MIP in the foreseeable future (at the end of the ramp up phase). We will strongly encourage hospitals to comply with the recommendations. In case hospitals use hardware that does not provide enough computational capacity for a specific complex analyses, they will be excluded for these queries (exclusion will be accounted for in the provenance information to ensure reproducibility). Exclusion will ensure to reduce the load on these hospitals and will also ensure timely execution of the queries.

3.8 The MIP: Relations to Other HBP Platforms

The MIP collaborates with:

- **The Neuroinformatics Platform (SP5):**
  - Use, share and develop the brain atlas Ontologies.
  - Integrate the MIP disease-related terms from Ontologies, Variables and Provenance into the knowledge graph, the ontology and provenance services created by SP5.

  - Integrate the aggregate phenotype descriptions and the biological signatures of diseases into the knowledge graph.

- **The HBP platform (SP6/SP13):**
  - Integration of the MIP within the Collaboratory (user authentification, authorisations, logins, user interfaces, collaborative projects).

3.9 The MIP: Necessary Parallel Activities

3.9.1 Data Providers recruitment

The work will focus on Increasing the number of data providers (activity led by WP 8.2 and WP 8.5) and the collection of data on patients with neurological and psychiatric diseases. Up to now, three sources of information have been identified:

- **Hospital data warehouses** from a list of 20 potential candidates. The final selection of hospitals will be based on the type, the amount of data and the existence of electronic medical record (EMR) systems. Hospitals will be included as R&D in the main EU countries (France, Germany, UK)
- **Registers of drug trials** conducted by major drug companies ("big pharmas"), mainly in Europe, over the past 15 years and link to biobanks (Department of Neuroimaging of the Politecnica Madrid)
- **Large-scale research databases** (neuroimaging and genetics data on healthy controls and patients) from international initiatives (ADNI, CBRAIN, …) or specific programs (3Cities Studies and Memento in France, Life in Germany, DPUK in United-Kingdom)

3.9.2 Deployment

The work will focus on specifying and defining the deployment strategy of MIP infrastructure with the selected data providers.

- **Establishing principles for data governance** (data ownership, privacy and usage) with hospital management, database management, security officers and local ethics committees.
- **Development of a scalable, large scale, data mining approach** - The process of choosing and/or building and/or linking efficient fit-for-purpose data mining algorithms to carry out the task optimally is a crucial aspect of the MIP. We expect that expertise from areas outside medicine, where data mining is currently used, will interact with the MIP to facilitate this functionality.
- **Using Regional and National Hubs.** If national or local legislation permits moving the data within the country (or region) local hubs can be set up. The architecture of the MIP accommodates this type of set up: several hospitals together act as a virtual hospital, i.e., they set up an LDSM together. The federation architecture will not be affected by this as such an LDSM is integrated as any other LDSM. In fact, from the perspective of the MIP we encourage local hubs as they remove points of failure and reduce overall latency. Of course, combining several hospitals into a local hub also requires more computational power as the amount of data increases.

- **Subcontracting the Data Federation** (activity led by WP 8.4) / Subcontracting Privacy protocols (activity led by WP 8.1):
  - Preparing the tendering process (legal procedures with local authorities).
  - Building a high-level product requirements specification for potential subcontractors.
  - Organising workshops with data providers where potential subcontractors present their proposed solutions.
  - Short-listing solutions, based on functionality and long-term reliability of the proposers.
  - Selecting a subcontractor on the basis of cost effectiveness and the result of proof-of-concept solutions.

3.9.3 Standardization of data and common ontology.

The work will focus on adapting existing data standards for potential data providers to transform their data into a form acceptable by the MIP.

- **Data:** The data provided by hospitals include raw data, such as imaging data (MRI, PET), CSV/XLS files, raw text, output of proprietary medical systems, as well as relational databases. The data traverse a series of sub-processes. They are anonymised, normalised and integrated. Finally, the data and the information extracted from them are released (i.e. made available for in-situ query) as Variables associated with Provenance descriptors.
- **Variables:** Variables are metadata that provide scientific descriptions of data. Variables are one component of the Federated Query Language. Variables are extracted by a mixture of number crunching, image processing and text mining, mapped to a common standard. Variable extraction data processing happens at the Data Store Mirror level in each data provider site, locally in most cases.
- Below are the Variable descriptions, along five dimensions:
  - Scale – genetic, molecular, cellular, circuits, systems.
  - Time – data acquisition at single point/multiple points per subject.
  - Space – centred on the brain (based on brain regions).
  - Pathology – clinical labels of brain pathology.
  - Demographics – age, gender.
- **Provenance:** For each of the Variables the Provenance describes precisely the materials and methods (exact algorithms and their versions) used to create them:
  - Materials: brain imaging – magnetic resonance imaging/computer tomography/PET, neurophysiology, etc.
  - Methods: voxel-based morphometry – volume, surface-based morphometry – cortical folding, power spectra, SI units for biochemical assays, etc.
- **Ontologies:** Ontologies are controlled vocabularies describing the Variables and relations among them with a meaningful grammar in the specific domain of interest (clinical phenotype, medical terms, brain regions, genes and proteins). In the first instance, we will use established ontologies for each domain, which will be updated as required.

- Examples of Ontologies: ICD-10/9 disease classification from the World Health Organisation adopted by hospitals, OMIM, DiseasesDB, eMedicine.

The approach, regarding (data and query) standardization, concerns describing the variables from the following standards

- **LOINC** provides a universal code system for reporting laboratory and other clinical observations and contains more than 30,000 different observations. For each observation the LOINC database includes a code, a long formal name, a “short" name as well as synonyms.

- **Cortical Labelling Protocol is used to identify brain regions.** Region examples are “Right Cerebellum Exterior", “Left Hippocampus" and others and corresponding values are estimated volumes in cc produced by a feature extraction process.

- **Single Nucleotide Polymorphisms (SNPs)** are probably the most important category of genetic changes influencing common diseases, since they have been shown to influence disease risk, drug efficacy and side-effects, provide information about ancestry, and predict aspects of how humans look and even act. We follow rs names provided by dbSNP. For example, rs2075650 is an example of an SNP found on gene TOMM40 of chromosome 19 associated with high risk of Alzheimer disease.

- **International Classification of Diseases (ICD-10),** diagnostics that refer to classification of mental and behavioral disorders. For example, Alzheimer disease with early onset, represented by code G30.1, is a subcategory of Alzheimer disease, represented by code G30.

3.10 Risk analysis and management

The following Risk Assessment methodology is used by the HBP in order to identify threats, and assess and mitigate the risks. The HBP Risk Assessment methodology covers the following five phases, which are explained in detail below:

1) Identifying the threats,

2) Evaluating the risk and vulnerability by assessing the impact and probability of an identified threat occurring,

3) Determining the degree of Control and Comprehension on or over the event,

4) Selecting the mitigating strategy, and

5) Identifying methods to fulfil the selected strategy and mitigate the risks.

The practical implementation of the HBP risk monitoring system requires a flow of information from the Subproject via the Management Subproject to the BoD and to the EC.

The identified risks are as follows:

- **Implementation**
  - **Impact:** Downgrading of the functionality it provides, but will not compromise the release or performance of the other modules.
  - **Contingency plan:** Adoption of a modular, incremental development process in which no module depends on a specific version of another module.
  - **End of risk:** Completion of releases and documentation.
- **Adoption, data provision**
  - **Impact:** If recruitment of data providers does not reach the targets set by the work plan, the amount of data will not allow statistically relevant analyses
  - **Contingency plan:** The Medical Informatics Platform will extend its recruitment effort to organizations and countries not included in its original plan. Mitigation of these risks will require larger strategic collaborative action with existing initiatives (Discussions with possible Partners have begun).
  - **End of risk:** When a quantitative and qualitative data pool is generated.
- **Successful hospital server deployment**
  - **Impact:** Essential to collect hospital data, but the HBP research groups cannot provide this service
  - **Contingency plan:** Mitigation of risk by subcontracting a commercial company with adequate expertise
  - **End of risk:** When first five hospitals are included in federated infrastructure – proof of concept.

Further risks:

- **Insufficient financial resources for effective operation of the platform**
  - **Impact:** Financial resources are crucial to run the platforms effectively due to high maintenance costs.
  - **Contingency Plan:** If necessary, the Project will seek funds from outside the Project for running and operating the ICT Platforms.
  - **End of risk:** Robust sustainable infrastructure established
- **Lack of users interests and community uptake**
  - **Impact:** If uptake is low, it will most probably be because the tools are not tailored to Use Cases that the users care most about. This questions the usefulness of the platform.
  - **Contingency Plan:** The Project is already investing significant resources to recruit and engage potential users. This will be addressed by iteratively refining ICT Platform development goals to focus on the most valuable Use Cases. The ICT Platform teams are following an iterative release strategy that expects and encourages regular refinement of development goals.
  - **End of risk:** Steadily increasing user numbers from different communities.

4. Web Portal (1st Layer)

4.1 Web Portal: Functional Requirements

4.1.1 Management

The Web Portal provides the Admin Users with access to the management interface.

Requirement ref.	Description	Applies to Users/Use Case
SP8-FR-012-WeP (WP 8.2, WP 8.1)	The platform shall allow an Admin User to modify metadata stores: Ontology, Variables and Provenance, via the Web Portal.	Admin User
SP8-FR-013-WeP (WP 8.3)	The platform shall allow an Admin User to add and modify statistical tools used by the services and Data Mining algorithms, via the Web Portal.	Admin User
SP8-FR-014-WeP (WP 8.1)	The platform shall allow an Admin User to register new data sources (i.e. new hospitals) by setting up mappings via the Web Portal and across the MIP ontology/schema.	Admin User

Table 5: Requirements for web portal management interface

4.1.2 Epidemiological Exploration

Requirement ref.	Description	Applies to Users/Use Case
SP8-FR-015-WeP (WP 8.2, WP 8.3)	The web portal shall allow all its users to explore Ontologies and Variables through the web interface, and to view the retrieved results.	End Users SP8-UC-001 & SP8-UC-002

Table 6: Web portal requirements for epidemiological exploration

4.1.3 Interactive Analysis

Requirement ref.	Description	Applies to Users/Use Case
SP8-FR-016-WeP (WP 8.1)	The web portal shall provide users with a way of defining queries and aggregation functions.	End Users (GU & DU) SP8-UC-003 & SP8-UC-004
SP8-FR-017-WeP (WP 8.3, WP 8.1)	The web portal shall forward queries to the federated infrastructure and present all received results back to the user for analysis.	End Users (GU & DU) SP8-UC-003 & SP8-UC-004
SP8-FR-018-WeP (WP 8.3)	The system shall have a rich library of functions with established statistical and analytical tools, to enable calculations and presentation of results, preserving data correctness, as well as its variability at different spatial and temporal granularities.	End Users (GU & DU) SP8-UC-003 & SP8-UC-004
SP8-FR-019-WeP (WP 8.3)	More than 30 generic functions and advanced mathematical tools shall be defined, implemented, tested and added to the library of functions. Examples: summaries (numerical summaries, frequency distributions, count missing observations, correlation matrix), contingency tables, multi-way table; statistical parametric and non-parametric tests (single sample *t*-test, ANOVA); machine learning techniques for data classification; etc.	End Users (GU & DU) SP8-UC-003 & SP8-UC-004
SP8-FR-020-WeP (WP 8.3)	The system shall be configured to allow modifications of existing tools to increase the replicability of results by means of multiple tests corrections (e.g. controlling the false discovery rate).	End Users (GU & DU) SP8-UC-003 & SP8-UC-004
SP8-FR-021-WeP (WP 8.3)	The system shall include specially tailored analysis tools to increase interpretability beyond that gained by using traditional learning tools.	End Users (GU & DU) SP8-UC-003 & SP8-UC-004

Table 7: Web portal requirements for interactive analysis

4.1.4 Biological Signatures of Diseases

Requirement ref.	Description	Applies to Users/Use Case
SP8-FR-022-WeP **(WP 8.2, WP 8.3)**	Biological Signatures of Diseases shall be produced by rulebased algorithms running on dedicated data mining servers, and possibly on supercomputers in due course.	End Users (GU & DU) SP8-UC-006, SP8-UC-007 & SP8-UC-008

Table 8: Web portal requirements for biological signatures of diseases

4.1.5 Data Mining

Requirement ref.	Description	Applies to Users/Use Case
SP8-FR-023-WeP **(WP 8.3)**	Data mining servers shall contain both statistical, data mining and rule discovery tools. The statistical tools library shall include state-of-the-art supervised methods and purposely built methods able to run in a distributed infrastructure.	End Users (GU & DU) SP8-UC-006, SP8-UC-007 & SP8-UC-008
SP8-FR-024-WeP **(WP 8.3)**	The data mining process shall be dynamic and near real time. As new data arrive (by patient accrual in participating hospitals and by new hospital or research database recruitment), they shall be incorporated into this dynamic, continuous and background process.	End Users (GU & DU) SP8-UC-006, SP8-UC-007 & SP8-UC-008

Table 9: Web portal requirements for data mining

4.1.6 User Interface

Requirement ref.	Description	Applies to Users/Use Case
SP8-FR-025-WeP (WP 8.1,WP 8.2, WP 8.3)	The web interface shall allow users to browse and build queries for Ontology, Variables and Provenance.	All users SP8-UC-001 to SP8-UC-008
SP8-FR-026-WeP (WP 8.1,WP 8.2, WP 8.3)	The web interface shall allow users to view results as counts, plots or histograms.	All users SP8-UC-001 to SP8-UC-008
SP8-FR-027-WeP (WP 8.1,WP 8.2, WP 8.3)	The system shall allow users to visualise 3D brain results.	All users SP8-UC-001 to SP8-UC-008
SP8-FR-028-WeP (WP 8.1,WP 8.2, WP 8.3)	The system shall allow users to save queries for further analysis.	All users SP8-UC-001 to SP8-UC-008

Table 10: Web portal requirements for user interface

4.2 Web Portal: Functional Implementation of Use Cases and Services

Use Case	System functions of the layer	Functional requirements
SP8-UC-001 **Epidemiological Exploration**	The system forms a query and sends it to the common metadata space. The system produces a chart plot.	SP8-FR-015-WeP
SP8-UC-002 **Epidemiological Exploration**	The system forms a query and sends it to the common metadata space. The system produces a chart plot.	SP8-FR-015-WeP
SP8-UC-003 **Interactive Analysis**	The system forms a query using the common schemata. It sends it to the Data Federation layer. The system produces a chart plot.	SP8-FR-016-WeP to SP8-FR-021-WeP
SP8-UC-004 **Interactive Analysis**	The system forms a query using the common schemata. Then it sends the query and the UDF statistical function to the Data Federation layer.	SP8-FR-016-WeP to SP8-FR-021-WeP
SP8-UC-005 **Interactive Analysis**	The script is added to the statistical tools library.	SP8-FR-016-WeP to SP8-FR-021-WeP
SP8-UC-006 **Biological Signatures of Diseases**	The system displays the biological signatures of diseases and 3D visualisation.	SP8-FR-022-WeP
SP8-UC-007 **Biological Signatures of Diseases**	The Web Portal loads the biological signatures of diseases and runs the best match algorithm. It also displays the Variables.	SP8-FR-022-WeP
SP8-UC-008 **Data Mining**	The system forms a query using the common schemata. Then it sends the query and the UDF statistical function to the Data Federation layer.	SP8-FR-023-WeP & SP8-FR-024-WeP

Table 11: Web portal functional implementation of use cases and services

4.3 Web Portal: Non-Functional Requirements

Requirement ref.	Description
SP8-NFR-007-WeP	Scalability
SP8-NFR-008-WeP	Flexibility
SP8-NFR-009-WeP	Privacy

Table 12: Web portal non-functional requirements

4.4 Web Portal: Software

4.4.1 Operations Software

Ontologies are used in the Web Portal to define a standard workflow for users to explore Variables, and to provide the necessary vocabulary for issuing queries to the MIP.

- MIP ontology driven data mapping will be designed with the use of a sophisticated visual mapping tool such as ++Spicy[4], an open source information integration tool. The suite is written in Java and includes an extensible plug-and-play platform for information integration tools.
- For the representation of Ontologies W3C[5] has standardised the OWL Web Ontology Language[6] that has been recently extended to OWL 2.[7] Additionally, many tools have been developed for editing Ontologies. The most prominent is Protégé[8], an open source tool written in Java that also supports a web-based interface (Web Protégé[9]).

4.4.2 User-Facing Software

Application Programming Interfaces (APIs) will be used for data visualisation (brain imaging), data summary. The APIs will allow graphical/numerical output, interactive zoom-in functions and visual mapping creation.

4.5 Web Portal: Physical Architecture

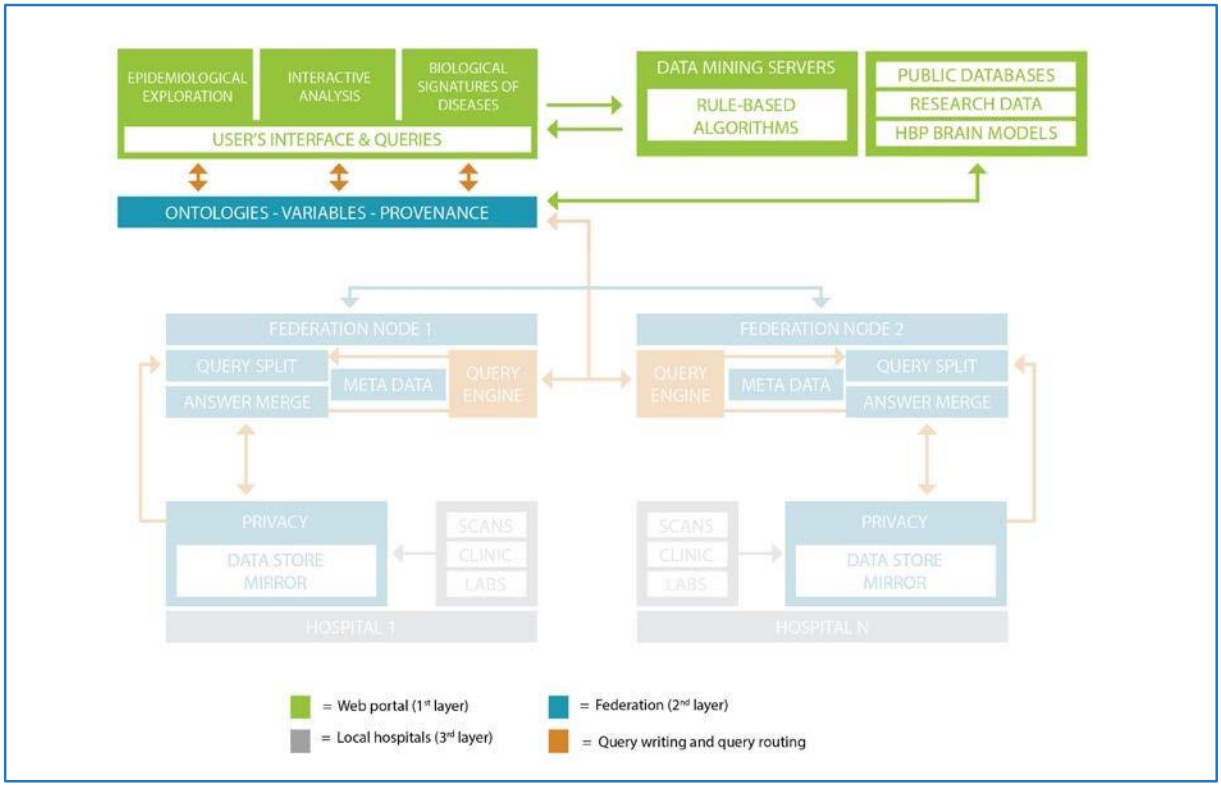

Figure 10: Web Portal layer

4.6 Web Portal: Interfaces to Other Layers

The Web Portal sends queries to the Data Federation layer and returns results to the user.

4.7 Web Portal: Prerequisites

The Web Portal will be based on the tools provided by the Collaboratory.

5. Data Federation (2nd Layer)

5.1 Data Federation: Functional Requirements

5.1.1 Federated Query Processing

Queries posed on the Web Portal are forwarded to local hospital Data Store Mirrors through the Data Federation layer.

Requirement ref.	Description	Applies to Users/Use Case
SP8-FR-029-DF (WP 8.1, WP 8.4)	The Data Federation layer shall receive queries built by users at the Web Portal level.	All users SP8-UC-001 to SP8-UC-008
SP8-FR-030-DF (WP 8.1, WP 8.4)	The Data Federation layer shall receive queries of two types (abbreviated QT below): QT1 - simple statistical queries, if the Epidemiological Exploration service was selected at the Web Portal level. QT2 - queries requiring more complex data mining, if Interactive Analysis or Biological Signatures of Diseases services were selected at the Web Portal level.	All users SP8-UC-001 to SP8-UC-008
SP8-FR-031-DF (WP 8.4)	If QT1: the Data Federation shall run and extract the summary of counts available, without accessing local hospital data.	All users SP8-UC-001 to SP8-UC-008
SP8-FR-032-DF (WP 8.4)	If QT2: the Data Federation shall re-engineer the QT2 queries and fire them as QT2.1, QT2.2,…, QT2.n to the appropriate "n" local hospitals to collect the information required by QT2 criteria.	All users SP8-UC-001 to SP8-UC-008
SP8-FR-033-DF (WP 8.1, WP 8.4)	The re-engineered QT2.1, …, QT2.n queries shall collect the anonymised information from local hospitals and send it back to the Data Federation layer.	All users SP8-UC-001 to SP8-UC-008
SP8-FR-034-DF (WP 8.1, WP 8.4)	The Data Federation layer shall merge the results received from QT2.1, …, QT2.n.	All users SP8-UC-001 to SP8-UC-008
SP8-FR-035-DF (WP 8.1, WP 8.4)	The Data Federation layer shall transfer the merged results of QT2.1, …, QT2.n or the results of QT1 back to the Web Portal, and display them to the user.	All users SP8-UC-001 to SP8-UC-008
SP8-FR-036-DF (WP 8.1, WP 8.4)	The Data Federation layer shall schedule, monitor and execute the set of QT2 queries.	All users SP8-UC-001 to SP8-UC-008

Table 13: Data federation functional requirements

5.1.2 Data Integration

The vocabulary of queries will be the one provided by the MIP schema. It might be different from the vocabulary used at the Local Data layer, i.e. the vocabulary used in each hospital Data Store Mirror schema. This is the reason why a data integration approach shall be used at the Data Federation layer.

Requirement ref.	Description	Applies to Users/Use Case
SP8-FR-037-DF **(WP 8.1, WP 8.4)**	Virtual data integration shall be employed, where the MIP ontology/schema acts as a mediated schema. The schema of the virtual database shall relate to the Data Store Mirror schemata (sources) through schema mappings.	All users SP8-UC-001 to SP8-UC-008
SP8-FR-038-DF **(WP 8.1, WP 8.4)**	Schema mappings shall be defined declaratively. They specify how attributes in the sources correspond to attributes in the MIP Ontology, how different groupings of attributes are resolved into tables, and how to resolve differences regarding the specification of data values from different sources. Hence, the Data Federation layer uses these mappings to transform the query to a set of queries that match the schemata of individual hospital Data Store Mirrors through a process called Query Rewriting.	All users SP8-UC-001 to SP8-UC-008

Table 14: Data integration requirements

5.1.3 Workflow Management

Requirement ref.	Description	Applies to Users/Use Case
SP8-FR-039-DF **(WP 8.1, WP 8.4)**	The workflow engine shall receive, schedule, execute and monitor queries at the Data Federation layer.	All users SP8-UC-001 to SP8-UC-008
SP8-FR-040-DF **(WP 8.1, WP 8.4)**	The workflow engine shall support both simple queries and complex queries for complex data processing, analyses and mining (defined on SQL extended with User Defined Functions (UDFs)).	All users SP8-UC-001 to SP8-UC-008

Table 15: Workflow management requirements

5.2 Data Federation: Use Cases

Use Case	System functions of the layer	Functional requirements
SP8-UC-001 **Epidemiological Exploration**	The query is executed in the common metadata space. The system sends back results to the Web Portal layer.	SP8-FR-029-DF to SP8-FR-031-DF and SP8-FR-039-DF
SP8-UC-002 **Epidemiological Exploration**	The query is executed in the common metadata space. The system sends back results to the Web Portal layer.	SP8-FR-029-DF to SP8-FR-031-DF and SP8-FR-039-DF
SP8-UC-003 **Interactive Analysis**	The system routes the query using the local schemata to the relevant local hospitals. The results are merged using the common schemata.	SP8-FR-029-DF & SP8-FR-031-DF and SP8-FR-033-DF & SP8-FR-036-DF and SP8-FR-037-DF & SP8-FR-038-DF and SP8-FR-039-DF & SP8-FR-040-DF
SP8-UC-004 **Interactive Analysis**	The system routes the query using the local schemata to the relevant local hospitals. The results are merged using the common schemata.	SP8-FR-029-DF & SP8-FR-031-DF and SP8-FR-033-DF & SP8-FR-036-DF and SP8-FR-037-DF & SP8-FR-038-DF and SP8-FR-039-DF & SP8-FR-040-DF

SP8-UC-005 **Interactive Analysis**	N/A	SP8-FR-029-DF to SP8-FR-036-DF and SP8-FR-040-DF
SP8-UC-006 **Biological Signatures of Diseases**	The system retrieves the biological signatures of diseases from the common metadata space. It continuously propagates the data mining to the local data sources.	SP8-FR-029-DF to SP8-FR-036-DF and SP8-FR-040-DF
SP8-UC-007 **Biological Signatures of Diseases**	The system retrieves the biological signatures of diseases from the common meta-data space. Continuous data mining queries.	SP8-FR-029-DF to SP8-FR-036-DF and SP8-FR-040-DF
SP8-UC-008 **Data Mining**	The system routes the query using the local schemata to the relevant local hospitals. The results are merged using the common schemata.	SP8-FR-029-DF & SP8-FR-031-DF and SP8-FR-033-DF & SP8-FR-036-DF and SP8-FR-037-DF & SP8-FR-038-DF and SP8-FR-039-DF & SP8-FR-040-DF

Table 16: Data federation use cases

5.3 Data Federation: Non-Functional Requirements

Requirement ref.	Description
SP8-NFR-010-DF	Efficiency – scalability, optimise resource and time management
SP8-NFR-011_DF	Reliability - fault tolerance, recoverability
SP8-NFR-012_DF	Interoperability with other systems
SP8-NFR-013_DF	Compliance
SP8-NFR-014_DF	Security
SP8-NFR-015_DF	Maintainability
SP8-NFR-016_DF	Extensibility

Table 17: Data federation non-functional requirements

5.4 Data Federation: Software

5.4.1 Operations Software

The software to be used in the Data Federation layer includes:

- The software anonymisation filter, developed by a subcontractor (Task 8.1.4).
- The software supporting the federation infrastructure – i.e. the software module that maintains a metadata directory, dispatches/routes queries and merges and returns result. This software will also be developed by a subcontractor (Task 8.4.2).

In addition, there will be a next generation data-flow language engine on top of the Data Federation layer. This engine will be capable of supporting distributed execution of complex, resource, and time-consuming data processing flows.

In order to create mappings, sophisticated software needs to be employed. Candidates for the final choice of such software include ++Spicy and similar tools (AlignmentAPI[10], Clio[11]).

Schema mappings can be expressed through three mapping formalisms:

- The first – called Global As View (GAV) – expresses how entities in the mediated schemata are related to entities in the Data Store Mirrors.
- The second formalism – called Local As View (LAV) – uses the opposite approach to GAV. It specifies how entities in Data Store Mirrors are expressed with the aid of entities in mediated schemata.
- Finally, a combination of the previous two formalisms - the Global and Local As View (GLAV) – carries both their advantages and disadvantages. This formalism is equivalent to expressiveness in Tuple-Generating Dependencies also known as TGDs.

5.5 Data Federation: Physical Architecture

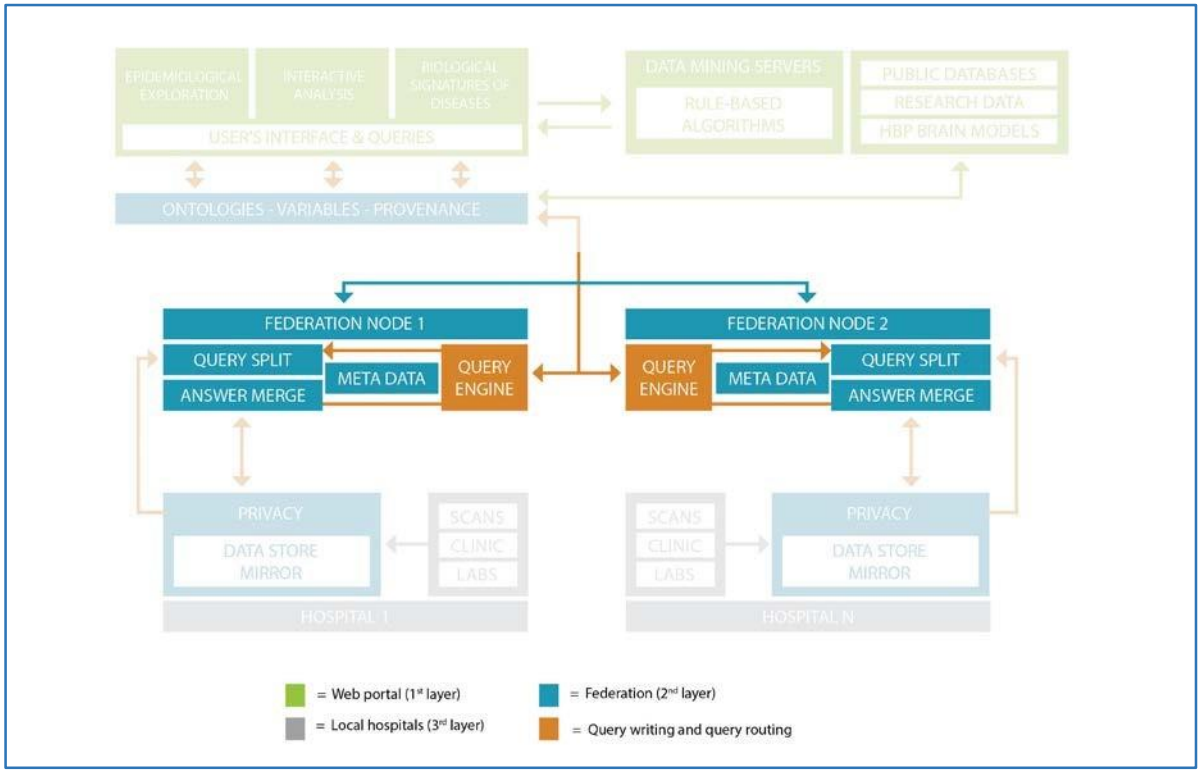

Figure 11: Data federation layer

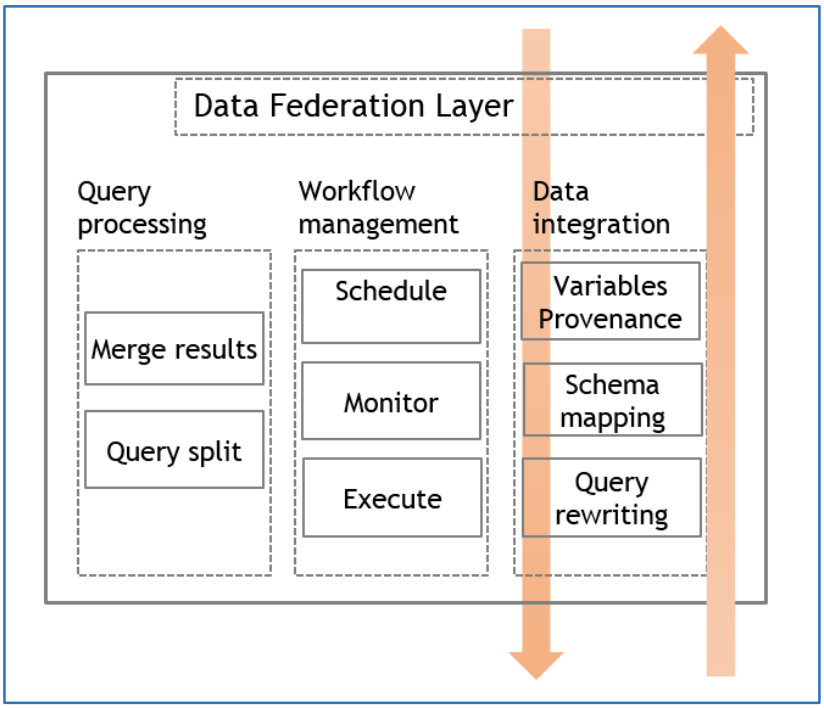

Figure 12: Components of the data federation architecture

5.6 Data Federation: Interfaces to Other Layers

The Data Federation layer receives queries from the Web Portal layer (either users or data mining applications) and dispatches them to local hospitals, where they are executed locally. Upon receiving results from local hospitals, the Data Federation merges results and returns them to the Web Portal. The Data Federation layer therefore connects the two other layers, the Web Portal and the Local Data layers.

5.7 Data Federation: Prerequisites

The SP8 team will need to interact with an external company for the development and the deployment of a dedicated software that will support the Data Federation layer functionalities, as described in this document. The external company will provide help for interfacing with the multiple database systems (e.g. Oracle, Microsoft SQL, IBM) used in local hospitals.

5.8 Data Federation: Necessary Parallel Activities

The SP8 team will need to specify the interfaces to the other layers, along with the development of the anonymisation filter and the data integration approach.

6. Local Data (3rd Layer)

6.1 Local Data: Functional Requirements

6.1.1 Data Integration

Queries submitted at the Web Portal layer will be redirected to local hospitals to gather required information.

Hospital data is often stored locally in disparate sources, of various technologies and with different schemata. Each participating hospital will be requested to build a local data warehouse. This will allow data to be accessed and integrated by the MIP in the most efficient and secure manner, fully maintaining its accuracy, consistency and privacy.

An Extract/Transform/Load (ETL) process will be used for populating local data warehouses (called Local Data Store Mirrors). Requirements for this process are listed below.

Requirement ref.	Description	Applies to Users/Use Case
SP8-FR-041-LD (WP 8.1)	**Extract:** The system shall extract the available research data from the disparate and heterogeneous data sources of each hospital.	All users SP8-UC-003 & SP8-UC-004 SP8-UC-006, SP8-UC-007 & SP8-UC-008
SP8-FR-042-LD (WP 8.1, WP 8.3)	**Transform:** The system shall convert the extracted data into a format as simple as possible (e.g. flat files), with the aid of rules or functions, to a format that is compliant with the schema of each hospital's Data Store Mirror.	All users SP8-UC-003 & SP8-UC-004 SP8-UC-006, SP8-UC-007 & SP8-UC-008
SP8-FR-043-DF ? WP 8.3	**Transform:** The process shall perform data cleaning and normalisation operations, entity resolution, application of various aggregate functions and finally anonymisation.	All users SP8-UC-003 & SP8-UC-004 SP8-UC-006, SP8-UC-007 & SP8-UC-008
SP8-FR-044-LD (WP 8.1)	**Load**: Load processes shall populate each hospital Data Store Mirror with data produced during the "Transform" process, after anonymisation.	All users SP8-UC-003 & SP8-UC-004 SP8-UC-006, SP8-UC-007 & SP8-UC-008

Table 18: Local data integration functional requirements

6.1.2 Variable Description and Provenance

Requirement ref.	Description	Applies to Users/Use Case
SP8-FR-045-LD (WP 8.1, WP 8.2)	Variable descriptions shall be produced during the ETL process. Metadata shall describe the extracted data, the transformation process undergone, and finally how they were loaded onto each hospital Data Store Mirror.	All users SP8-UC-003 & SP8-UC-004 SP8-UC-006, SP8-UC-007 & SP8-UC-008
SP8-FR-046-LD (WP 8.1, WP 8.2)	Provenance information shall include descriptions of the data creator, the dates of export and, the transformation and loading of data to a Data Store Mirror. Provenance provides information about the origin of the input data, i.e. proprietary system and its location, software or means used to produce them, as well as standards used to describe them.	All users SP8-UC-003 & SP8-UC-004 SP8-UC-006, SP8-UC-007 & SP8-UC-008

Table 19: Variable description and provenance requirements

6.1.3 Data Anonymisation

The anonymisation module of the Local Data layer uses a two-pronged approach to accomplish its task:

Requirement ref.	Description	Applies to Users/Use Case
SP8-FR-047-LD (WP 8.1)	Personal identifiers shall be removed when exporting data from hospital information systems, i.e. even before the MIP accesses the data for the first time.	All users SP8-UC-001 to SP8-UC-008
SP8-FR-048-LD (WP 8.1)	The MIP shall only allow aggregate queries to run and shall filter all results (ensuring they do not contain personal patient information) before returning them to users of the platform.	All users SP8-UC-001 to SP8-UC-008

Table 20: Data anonymisation requirements

6.1.4 Image Feature Extraction

Current algorithms for dealing with MRI data are designed for high quality images.

Requirement ref.	Description	Applies to Users/Use Case
SP8-FR-049-LD (WP 8.3)	Image Feature Extraction shall identify suitable scans and shall extract useful information from metadata in the relevant DICOM files.	All users SP8-UC-001 to SP8-UC-008
SP8-FR-050-LD (WP 8.3)	Image Feature Extraction shall provide volume measurements of different brain structures from structural MRI scans, using robust and sophisticated algorithms.	All users SP8-UC-001 to SP8-UC-008
SP8-FR-051-LD (WP 8.3)	Image Feature Extraction shall reduce the information from each dataset of images to a relatively small and manageable number of features. The system shall be built to allow for future increases in the range of features analysed as queries become more complex and images can be reduced to empirically more informative features.	All users SP8-UC-001 to SP8-UC-008
SP8-FR-052-LD (WP 8.3)	The system shall automatically identify and label brain regions to permit automatic retrieval of the MRI features. This shall be done by applying: 1. Automatic segmentation of brain structures from 35 MRI training scans, that have been manually labelled (from the CC-BY-NC licensed "MICCAI 2012) 2. The nonlinear registration of the same 35 MRI training scans onto new scan data that needs to be segmented and labelled.	All users SP8-UC-001 to SP8-UC-008
SP8-FR-053-LD (WP 8.3)	Clinical scans are acquired with a variety of very different image contrasts. So, the actual label propagation shall use a patch-based pattern recognition approach based on similarities between tissue maps.	All users SP8-UC-001 to SP8-UC-008
SP8-FR-054-LD (WP 8.3)	The Image Feature Extraction shall extend super-resolution approaches to combine data obtained using a variety of image contrasts.	All users SP8-UC-001 to SP8-UC-008

SP8-FR-055-LD (WP 8.3)	The Image Feature Extraction shall combine scans with different pulse sequences by fitting a probabilistic generative (forward) model to the data, accounting for different signal intensities in scans collected with different pulse sequences.	All users SP8-UC-001 to SP8-UC-008

Table 21: Image feature extraction requirements

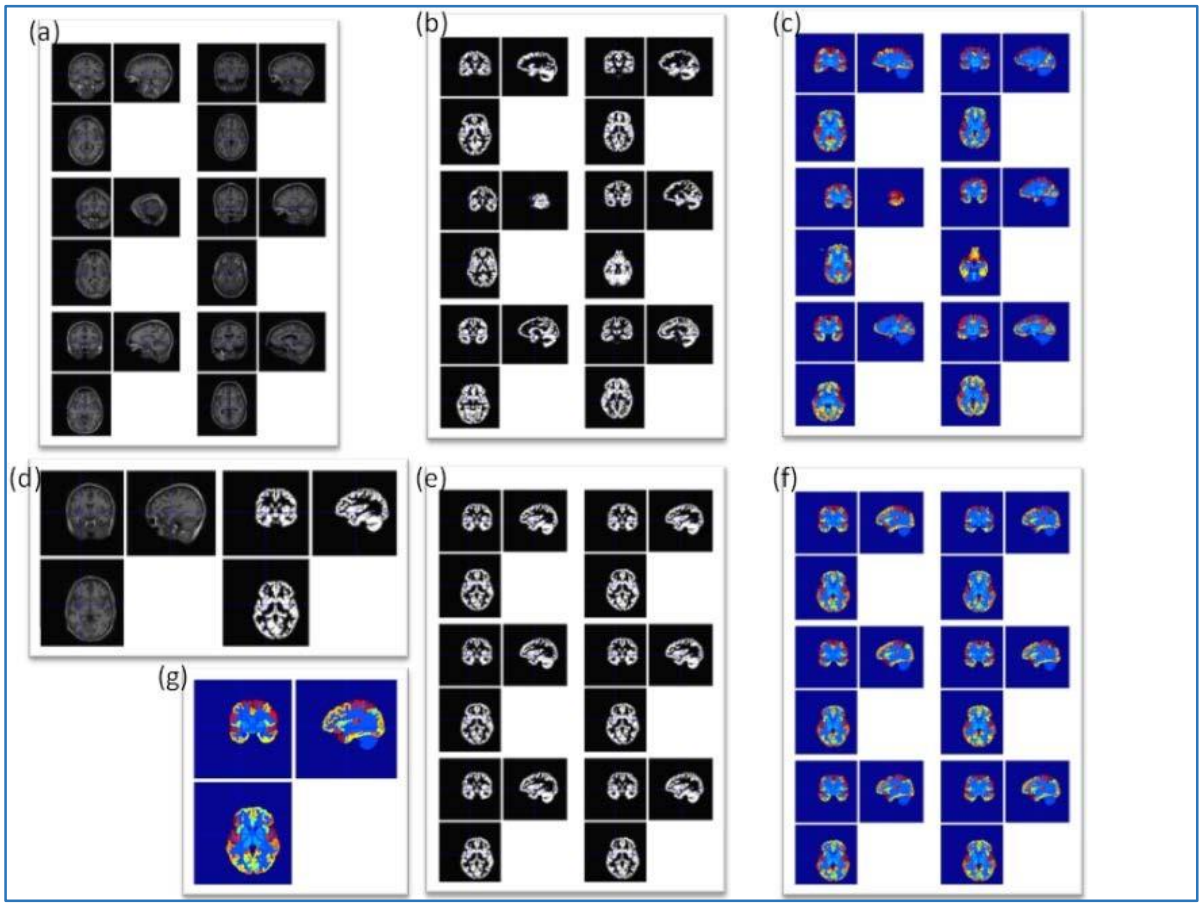

Figure 13: Structure labelling procedure

a) Six of the training scans; (b) Tissue maps (grey matter) segmented from training scans; (c) Manually defined labels of training scans; (d) Patient scan data to label, and tissue maps segmented from them; (e) Tissue maps from training data warped into alignment with patient data; (f) Training labels warped into alignment with patient data; (g) Labels propagated on to patient data by patch-based pattern recognition approach.

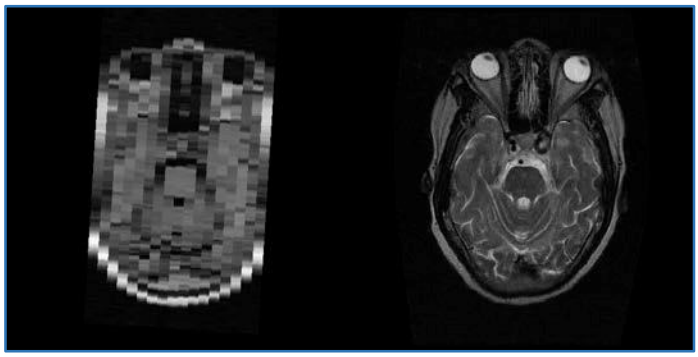

Figure 14: An axial section through T1- and T2-weighted hospital scans (same subject)

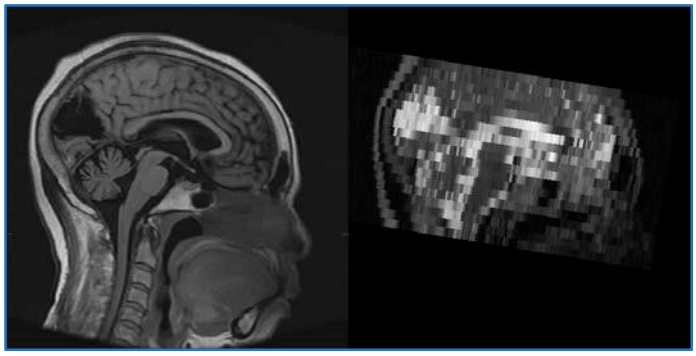

Figure 15: A sagittal section through the same data

6.1.5 Query Engine

The Query Engine receives queries from the Data Federation layer and runs them at local hospitals (Local Data layer). The Query Engine retrieves and posts back results, at the Data Federation level first and then at the Web Portal level.

Requirement ref.	Description	Applies to Users/Use Case
SP8-FR-056-LD (WP 8.1)	The Query Engine shall directly access data from different flat files in the Data Store Mirror, and hence avoid the traditional Database Management System (DBMS) data duplication.	All users SP8-UC-003 & SP8-UC-004 SP8-UC-006 & SP8-UC-007
SP8-FR-057-LD (WP 8.1)	The Query Engine shall use code generation techniques to efficiently execute queries on unindexed and unstructured data within the flat files of the Data Store Mirror.	All users SP8-UC-003 & SP8-UC-004 SP8-UC-006 & SP8-UC-007
SP8-FR-058-LD (WP 8.1)	The current prototype of the Query Engine shall facilitate querying and analysing multi-dimensional datasets by arraybased implementations like SciDB.	All users SP8-UC-003 & SP8-UC-004 SP8-UC-006 & SP8-UC-007
SP8-FR-059-LD (WP 8.1)	The Query Engine shall provide a declarative language interface (AQL), as well as bindings to popular statistical analysis tools such as R.	All users SP8-UC-003 & SP8-UC-004 SP8-UC-006 & SP8-UC-007
SP8-FR-060-LD (WP 8.1)	The Query Engine shall address the issues of data anonymisation, data duplication and data heterogeneity in Local Data Store Mirrors.	All users SP8-UC-003 & SP8-UC-004 SP8-UC-006 & SP8-UC-007

Table 22: Query engine requirements

6.2 Local Data: Use Cases

Use case	System functions of the layer	Functional requirements
SP8-UC-001 **Epidemiological Exploration**	NA	NA
SP8-UC-002 **Epidemiological Exploration**	NA	NA
SP8-UC-003 **Interactive Analysis**	The system executes the query at the Local Data Store Mirrors using the local schemata in the relevant local hospitals. The results are sent to the Data Federation layer.	SP8-FR-041-LD to SP8-FR-044-LD and SP8-FR-045-LD & SP8-FR-046-LD and SP8-FR-047-LD & SP8-FR-048-LD
SP8-UC-004 **Interactive Analysis**	The system executes the query at the Local Data Store Mirrors using the local schemata in the relevant local hospitals. The results are sent to the Data Federation layer.	SP8-FR-041-LD to SP8-FR-044-LD and SP8-FR-045-LD & SP8-FR-046-LD and SP8-FR-047-LD & SP8-FR-048-LD
SP8-UC-005 **Interactive Analysis**	NA	NA
SP8-UC-006 **Biological Signatures of Diseases**	Continuous automated data mining runs with all the data sources	SP8-FR-041-LD to SP8-FR-044-LD and SP8-FR-045-LD & SP8-FR-046-LD and SP8-FR-047-LD & SP8-FR-048-LD

SP8-UC-007 **Biological Signatures of Diseases**	Continuous automated data mining runs with all the data sources.	SP8-FR-041-LD to SP8-FR-044-LD and SP8-FR-045-LD & SP8-FR-046-LD and SP8-FR-047-LD & SP8-FR-048-LD
SP8-UC-008 **Data Mining**	The system executes the query at the Local Data Store Mirrors using the local schemata in the relevant local hospitals. The results are sent to the Data Federation layer.	SP8-FR-041-LD to SP8-FR-044-LD and SP8-FR-045-LD & SP8-FR-046-LD and SP8-FR-047-LD & SP8-FR-048-LD

Table 23: Local data use cases

6.3 Local Data: Non-Functional Requirements

Requirement Reference	Description
SP8-NFR-017_LD	Efficiency – scalability, optimise resource and time management
SP8-NFR-018_LD	Reliability - fault tolerance, recoverability
SP8-NFR-019_LD	Interoperability with other systems
SP8-NFR-020_LD	Compliance
SP8-NFR-021_LD	Security
SP8-NFR-022_LD	Maintainability
SP8-NFR-023_LD	Extensibility

Table 24: Local data non-functional requirements

6.4 Local Data: Software

6.4.1 Operations Software

- The prototype of the Query Engine is currently based on SciDB, with interface to R.
- Image Feature Extraction will use tools developed in the framework of Statistical Parametric Mapping12(SPM).
- The ETL process (also mentioned in Section 6.1.1) will use data transformation tools and methods to bring the hospital raw data to formats compliant with each hospital's Data Store Mirror, and then upload it so that it can be used within the MIP.

6.4.2 User-Facing Software

The system will provide management tools to Authorised Users for database housekeeping.

6.5 Local Data: Physical Architecture

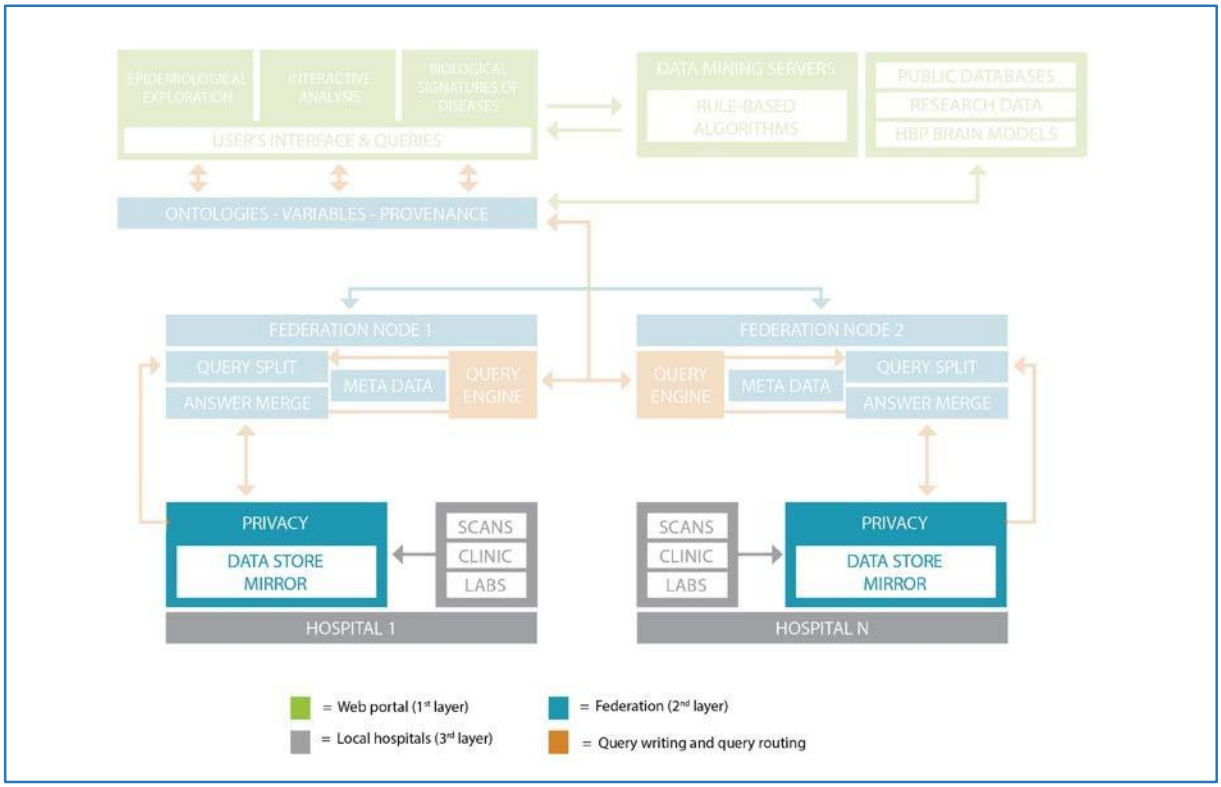

Figure 16: Local Data layer

The Data Store Mirror in each hospital will be a federation node in the Data Federation layer. This implies that users will be able to use the MIP to perform queries, data analysis and data mining with respect to data that reside in all federation nodes, i.e. all hospitals that provide data.

Data Store Mirrors will reside in servers located in local hospitals. They will be accessible only through the MIP. The specifications for local servers (such as computational power,

data store space and memory size) will be studied and defined by the time the size of the data to be shared by each hospital is assessed and known.

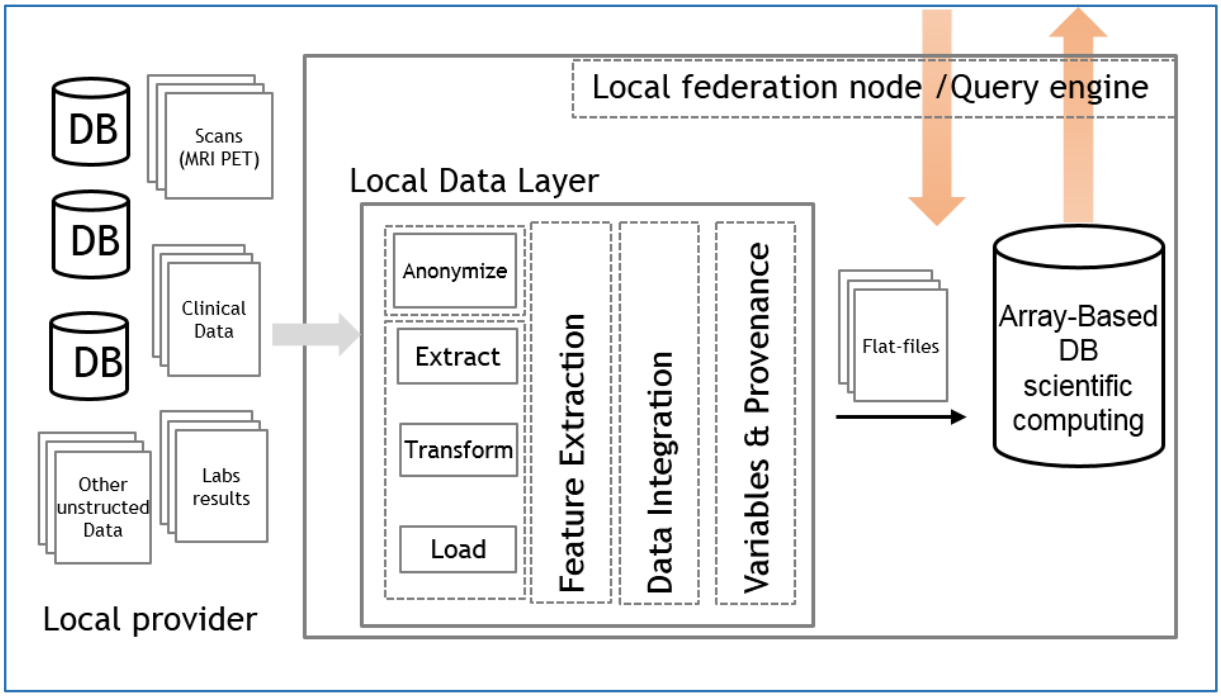

Figure 17: Components of the Local Data architecture

6.6 Local Data: Interfaces to Other Layers

Local Data tools will be accessible through the Data Federation layer and will return results to that layer. The exact interface still needs to be specified.

6.7 Local Data: Prerequisites

The SP8 team will interact with hospitals' IT and data warehouse management teams for the development and deployment of tools that will support the creation of the Data Store Mirror. A hospital's IT and data warehouse management team will provide help for interfacing with the multiple standards used in each hospital (e.g. DICOM, HL7) and with proprietary formats (e.g. Siemens). Tools for privacy and anonymisation will also be developed with the hospitals' security officers according to local policies and protocols.

6.8 Local Data: Necessary Parallel Activities

The tasks contributing to a Local Data Store Mirror (i.e. implementing Image Feature Extraction, Anonymisation and Tool Integration) need to be executed in parallel.

7. Functions

The MIP functionalities will be developed in the SP8 work packages according to the specifications described in this document. The functions are distributed over the 30-month duration of the Ramp-Up Phase.

To summarise, WP 8.1 and WP 8.3 will implement functions related to the creation of the Local Data Store Mirror and the development of the query engine. The functions related to Ontology and Variables descriptions will be created by WP 8.1, in collaboration with WP 8.2 for the validation and the release of Variables. WP 8.1 will also develop functions related to the Data Federation layer, in collaboration with WP 8.4 for the implementation of third-party tools (sub-contracting). WP 8.3 will implement functions related to services and data mining, which are hosted at the Web Portal layer, in collaboration with WP 8.2 for the application of the Biological Signatures of Diseases and WP 8.4 for maintenance. WP 8.5 will develop functions related to user support, training, clinical outreach and hospital recruitment.

7.1 WP 8.1

Task No:	8.1.1	Partner:	EPFL-DIAS
Function No:	8.1.1.1	Leader:	Anastasia AILAMAKI
Function Name:	Requirement analysis		
Use Case:	Analysis of requirements for the *in-situ* query engine (deployment environment, queries etc.).		
Planned Start Date:	October 2013	Planned Completion Date:	January 2014
Requires Functions:	None		

Table 25: Function 8.1.1.1

Task No:	8.1.1	Partner:	EPFL-DIAS
Function No:	8.1.1.2	Leader:	Anastasia AILAMAKI
Function Name:	Export mechanisms and data formats		
Use Case:	Tools and scripts for the export and analysis of data, i.e. data sources, export mechanisms and data formats.		
Planned Start Date:	December 2013	Planned Completion Date:	March 2014
Requires Functions:	None		

Table 26: Function 8.1.1.2

Task No:	8.1.1	Partner:	EPFL-DIAS
Function No:	8.1.1.3	Leader:	Anastasia AILAMAKI
Function Name:	Interface to Data Federation layer		
Use Case:	Implementation of data exchange (queries and results) between the Data Federation layer and Query Engine on the side of the Query Engine. Includes the specification and pertains to query language and result formats.		
Planned Start Date:	March 2014	Planned Completion Date:	June 2014
Requires Functions:	8.1.1.1 – Requirement analysis 8.1.1.2 - Export mechanisms and data formats		

Table 27: Function 8.1.1.3

Task No:	8.1.1	Partner:	EPFL-DIAS
Function No:	8.1.1.4	Leader:	Anastasia AILAMAKI
Function Name:	Implementation for flat CSV files and imaging data		
Use Case:	Prototypic implementation of query engine for flat CSV files (resulting from imaging data).		
Planned Start Date:	March 2014	Planned Completion Date:	November 2014
Requires Functions:	8.1.1.1 – Requirement analysis 8.1.1.2 - Export mechanisms and data formats		

Table 28: Function 8.1.1.4

Task No:	8.1.1	Partner:	EPFL-DIAS
Function No:	8.1.1.5	Leader:	Anastasia AILAMAKI
Function Name:	Definition and implementation of test cases		
Use Case:	Implementation of test cases for the query engine, testing for correctness of results with a broad range of datasets and query benchmarks. Bug fixing and improvements.		
Planned Start Date:	November 2014	Planned Completion Date:	February 2015
Requires Functions:	8.1.1.4 - Implementation for flat CSV files and imaging data		

Table 29: Function 8.1.1.5

Task No:	8.1.1	Partner:	EPFL-DIAS
Function No:	8.1.1.6	Leader:	Anastasia AILAMAKI
Function Name:	Interface to hospital datawarehouse		
Use Case:	Implementing and deploying the first version of the query engine for accessing raw imaging data at CHUV.		
Planned Start Date:	July 2015	Planned Completion Date:	August 2015
Requires Functions:	None		

Table 30: Function 8.1.1.6

Task No:	8.1.1	Partner:	EPFL-DIAS
Function No:	8.1.1.7	Leader:	Anastasia AILAMAKI
Function Name:	Exporting systems to other hospitals		
Use Case:	Packaging the query engine for integration at hospitals that participate in the Ramp-Up Phase. Implementation of scripts for deployment.		
Planned Start Date:	September 2015	Planned Completion Date:	December 2015
Requires Functions:	8.1.1.6 – Interface to hospital datawarehouse		

Table 31: Function 8.1.1.7

Task No:	8.1.2	Partner:	AUEB
Function No:	8.1.2.1	Leader:	Vasilis VASSALOS
Function Name:	Extract/Transform/Load process		
Use Case:	Design the Extract/Transform/Load process that will result in the extraction of hospital data and the population of its Data Store Mirror, as well as the schema for each Data Store Mirror.		
Planned Start Date:	January 2014	Planned Completion Date:	July 2014
Requires Functions:	None		

Table 32: Function 8.1.2.1

Task No:	8.1.2	Partner:	AUEB
Function No:	8.1.2.2	Leader:	Vasilis VASSALOS
Function Name:	Ontology/schema mapping		
Use Case:	Design the schema/ontology for the MIP that will provide the necessary vocabulary to users so that they can perform queries as well as investigate relations among its entities.		
Planned Start Date:	March 2014	Planned Completion Date:	October 2014
Requires Functions:	None		

Table 33: Function 8.1.2.2

Task No:	8.1.2	Partner:	AUEB
Function No:	8.1.2.3	Leader:	Vasilis VASSALOS
Function Name:	Manual schema mapping		
Use Case:	Develop/Integrate software to the Web Portal to allow users to create their own mappings across each hospital and the MIP schema/ontology.		
Planned Start Date:	March 2014	Planned Completion Date:	December 2014
Requires Functions:	None		

Table 34: Function 8.1.2.3

Task No:	8.1.2	Partner:	AUEB
Function No:	8.1.2.4	Leader:	Vasilis VASSALOS
Function Name:	Schema mapping across schemata and ontologies		
Use Case:	Create mappings across hospital schemata and the MIP ontology/schema. These mappings will be created semi-automatically or by users at the Web Portal.		
Planned Start Date:	August 2014	Planned Completion Date:	June 2015
Requires Functions:	8.1.2.1 - Extract/Transform/Load process 8.1.2.2 – Ontology/schema mapping		

Table 35: Function 8.1.2.4

Task No:	8.1.2	Partner:	AUEB
Function No:	8.1.2.5	Leader:	Vasilis VASSALOS
Function Name:	Query rewriting across hospital schemata		
Use Case:	Develop Query Rewriting algorithms. These algorithms will be able to rewrite queries posed over the vocabulary of the MIP ontology/schema, to the various schemata that describe the data. They will aid the Query Answering process.		
Planned Start Date:	April 2015	Planned Completion Date:	February 2016
Requires Functions:	8.1.2.4 - Schema mapping across schemata and ontologies		

Table 36: Function 8.1.2.5

Task No:	8.1.3	Partner:	UoA (P37)
Function No:	8.1.3.1	Leader:	Yannis IOANNIDIS
Function Name:	Complex Dataflow Processing Engine		
Use Case:	Design and develop a complex dataflow-processing engine with native UDF support. The engine will support dataflow processing without moving or storing data coming from local data providers.		
Planned Start Date:	November 2013	Planned Completion Date:	October 2015
Requires Functions:	None		

Table 37: Function 8.1.3.1

Task No:	8.1.3	Partner:	UoA (P37)
Function No:	8.1.3.2	Leader:	Yannis IOANNIDIS
Function Name:	User Defined Functions (UDFs) for Complex Dataflow Support		
Use Case:	Develop complex UDFs to be used in SQL-based dataflows executed by the Complex Dataflow Processing Engine. UDFs will provide functionality such as validation (character encoding, date format) and statistics (Pearson).		
Planned Start Date:	May 2014	Planned Completion Date:	May 2015
Requires Functions:	8.1.3.1 - Complex Dataflow Processing Engine		

Table 38: Function 8.1.3.2

Task No:	8.1.4	Partner:	EPFL-DIAS
Function No:	8.1.4.1	Leader:	Anastasia AILAMAKI
Function Name:	Anonymisation of patient data		
Use Case:	Analysis of requirements for the anonymisation of patient data. Formal specification of requirements.		
Planned Start Date:	October 2013	Planned Completion Date:	March 2014
Requires Functions:	None		

Table 39: Function 8.1.4.1

Task No:	8.1.4	Partner:	EPFL-DIAS
Function No:	8.1.4.2	Leader:	Anastasia AILAMAKI
Function Name:	Anonymisation module		
Use Case:	Deploy anonymisation module at CHUV. Functional and performance tests on site.		
Planned Start Date:	January 2016	Planned Completion Date:	March 2016
Requires Functions:	None		

Table 40: Function 8.1.4.2

7.2 WP 8.2

Task No:	8.2.1	Partner:	CHUV
Function No:	8.2.1.1	Leader:	Ferath Kherif
Function Name:	Clinical data analysis and preparation – on representative datasets		
Use Case:	Investigate the required vs. available data in order to provide maximum information to scientists. Define the data types needed, describe variables of existing data and prepare for insertion into the local nodes.		
Planned Start Date:	Oct 2013	Planned Completion Date:	March 2014
Requires Functions:			

Table 41: Function 8.2.1.1

Task No:	8.2.1	Partner:	CHUV
Function No:	8.2.1.2	Leader:	Ferath Kherif
Function Name:	Clinical data analysis and preparation – on extended datasets		
Use Case:	Investigate the required vs. available data in order to provide maximum information to scientists. Define the data types needed, describe variables of existing data and prepare for insertion into the local nodes.		
Planned Start Date:	April 2014	Planned Completion Date:	March 2015
Requires Functions:	8.2.1.1		

Table 42: Function 8.2.1.2

Task No:	8.2.1	Partner:	CHUV
Function No:	8.2.1.3	Leader:	Ferath Kherif
Function Name:	Clinical data analysis and preparation – automation of processes		
Use Case:	Investigate the required vs. available data in order to provide maximum information to scientists. Define the data types needed, describe variables of existing data and prepare for insertion into the local nodes. Identify the processes and implement automated workflows for generating variables.		
Planned Start Date:	April 2015	Planned Completion Date:	March 2016
Requires Functions:	8.2.1.2		

Table 43: Function 8.2.1.3

Task No:	8.2.1	Partner:	CHUV
Function No:	8.2.1.4	Leader:	Ferath Kherif
Function Name:	Research data analysis and preparation – on representative datasets		
Use Case:	Investigate the required vs. available data in order to provide maximum information to scientists. Define the data types needed, describe variables of existing data and prepare for insertion into the local nodes.		
Planned Start Date:	Oct 2013	Planned Completion Date:	March 2014
Requires Functions:			

Table 44: Function 8.2.1.4

Task No:	8.2.1	Partner:	CHUV
Function No:	8.2.1.5	Leader:	Ferath Kherif
Function Name:	Research data analysis and preparation – on extended datasets		
Use Case:	Investigate the required vs. available data in order to provide maximum information to scientists. Define the data types needed, describe variables of existing data and prepare for insertion into the local nodes.		
Planned Start Date:	April 2014	Planned Completion Date:	March 2015
Requires Functions:	8.2.1.1		

Table 45: Function 8.2.1.5

Task No:	8.2.1	Partner:	CHUV
Function No:	8.2.1.6	Leader:	Ferath Kherif
Function Name:	Research data analysis and preparation – automation of processes		
Use Case:	Investigate the required vs. available data in order to provide maximum information to scientists. Define the data types needed, describe variables of existing data and prepare for insertion into the local nodes. Identify the processes and implement automated workflows for generating variables.		
Planned Start Date:	April 2015	Planned Completion Date:	March 2016
Requires Functions:	8.2.1.2		

Table 46: Function 8.2.1.6

7.3 WP 8.3

Task No:	8.3.1	Partner:	Tel Aviv University
Function No:	8.3.1.1	Leader:	Mira MARCUS-KALISH
Function Name:	Descriptive function to provide summary statistics		
Use Case:	Epidemiological exploration (see SP8-UC-001).		
Planned Start Date:	December 2013	Planned Completion Date:	October 2014
Requires Functions:	None		

Table 47: Function 8.3.1.1

Task No:	8.3.1	Partner:	Tel Aviv University
Function No:	8.3.1.2	Leader:	Mira MARCUS-KALISH
Function Name:	Parametric tests		
Use Case:	Interactive analysis (see SP8-UC-003 and SP8-UC-004)		
Planned Start Date:	October 2014	Planned Completion Date:	April 2015
Requires Functions:	8.3.1.1 - Descriptive function to provide summary statistics		

Table 48: Function 8.3.1.2

Task No:	8.3.1	Partner:	Tel Aviv University
Function No:	8.3.1.3	Leader:	Mira MARCUS-KALISH
Function Name:	Supervised and unsupervised data mining algorithms		
Use Case:	Biological signatures of diseases (see SP8-UC-006 and SP8-UC-007)		
Planned Start Date:	October 2014	Planned Completion Date:	April 2016
Requires Functions:	None		

Table 49: Function 8.3.1.3

Task No:	8.3.2	Partner:	UCL
Function No:	8.3.2.1	Leader:	John ASHBURNER
Function Name:	Label propagation framework		
Use Case A:	Given a patient image segmented into various tissues, with the mapping to/from reference space, the function projects structurally pre-defined labels onto images.		
Use Case B:	A user can compute volumes of brain structures, or other image statistics derived from regional volumes.		
Planned Start Date:	February, 2014	Planned Completion Date:	April, 2014
Requires Functions:	None		

Table 50: Function 8.3.2.1

Task No:	8.3.2	Partner:	UCL
Function No:	8.3.2.2	Leader:	John ASHBURNER
Function Name:	Model construction for toy data, i.e. small representative clinical dataset		
Use Case A:	The user may eventually compute higher resolution tissue maps from collections of scans of the same subject acquired in diverse frames of reference (with thick slices) and pulse sequences.		
Use Case B:	A user may evaluate the effectiveness of the joint image modelling procedure under idealised conditions.		
Planned Start Date:	May, 2014	Planned Completion Date:	September, 2014
Requires Functions:	None		

Table 51: Function 8.3.2.2

Task No:	8.3.2	Partner:	UCL
Function No:	8.3.2.3	Leader:	John ASHBURNER
Function Name:	Sparse matrices and Intra-session alignment		
Use Case A:	Higher resolution image reconstructions may be projected back onto the original images, such that they can be compared against other scan data.		
Use Case B:	Rigid-body alignment may correct intra-session head motion in patient scans.		
Planned Start Date:	October, 2014	Planned Completion Date:	March, 2015
Requires Functions:	8.3.2.2 - Model construction for toy data		

Table 52: Function 8.3.2.3

Task No:	8.3.2	Partner:	UCL
Function No:	8.3.2.4	Leader:	John ASHBURNER
Function Name:	Integration with segmentation		
Use Case A:	A generative model will be combined within an existing image segmentation framework such that more reliable information can be extracted, based on the incorporation of prior knowledge about the contents of brain images and modelling of the intensity of non-uniformity artefacts found in MRI scans.		
Use Case B:	Spatial transformations to/from reference space may be computed.		
Planned Start Date:	April, 2015	Planned Completion Date:	August, 2015
Requires Functions:	8.3.2.2 - Model construction for toy data 8.3.2.3 - Sparse matrices and Intra-session alignment		

Table 53: Function 8.3.2.4

Task No:	8.3.2	Partner:	UCL
Function No:	8.3.2.5	Leader:	John ASHBURNER
Function Name:	Testing and optimisation		
Use Case A:	The effectiveness of the generative model will become known to the user.		
Use Case B:	Weaknesses will be identified, and corrected when possible.		
Planned Start Date:	September, 2015	Planned Completion Date:	December, 2015
Requires Functions:	8.3.2.2 - Model construction for toy data 8.3.2.3 - Sparse matrices and intra-session alignment 8.3.2.4 - Integration with segmentation		

Table 54: Function 8.3.2.5

Task No:	8.3.2	Partner:	UCL
Function No:	8.3.2.6	Leader:	John ASHBURNER
Function Name:	Redefine features		
Use Case A:	The effectiveness of existing image features will be assessed empirically.		
Use Case B:	Additional image features will be added, according to evidence of additional information leading to improved diagnostics or better sensitivity at defining disease signatures.		
Planned Start Date:	January, 2016	Planned Completion Date:	March, 2016
Requires Functions:	8.3.2.1 - Label propagation framework 8.3.2.2 - Model construction for toy data 8.3.2.3 - Sparse matrices and intra-session alignment 8.3.2.4 - Integration with segmentation 8.3.2.5 – Testing and optimisation		

Table 55: Function 8.3.2.6

7.4 WP8.4

Task No:	8.4.1	Partner:	CHUV
Function No:	8.4.1.1	Leader:	Ferath Kherif
Function Name:	Build Web UI – prototype/internal release		
Use Case:	*Infrastructure:* set up prototype Web infrastructure (small scale) *Build & Integrate:* Build UI, integrate with the other MIP components. *Launch:* create deployment and system plan and tests		
Planned Start Date:	End Jan 2015	Planned Completion Date:	End March 2015
Requires Functions:			

Table 56: Function 8.4.1.1

Task No:	8.4.1	Partner:	CHUV
Function No:	8.4.1.2	Leader:	Ferath Kherif
Function Name:	Build Web UI – public release		
Use Case:	*Infrastructure: scale up,* purchase & set up web servers, set up connections etc, set up development – test - production environments & staging processes *Build & Integrate: Scale up 1st prototype, improve usability, look and feel, add functionality and implement new use cases;* integrate with the scaled up MIP distributed architecture (i.e. creating Publishing Services/APIs). *Launch:* create deployment and system plan and tests		
Planned Start Date:	End March 2015	Planned Completion Date:	End March 2016
Requires Functions:	8.4.1.1		

Table 57: Function 8.4.1.2

Task No:	8.4.1	Partner:	CHUV
Function No:	8.4.1.3	Leader:	Ferath Kherif
Function Name:	Acquisition of standardised federation software infrastructure from an industrial provider		
Use Case:	Define selection criteria, evaluate potential solutions on the market and select most appropriate (define and implement Proof-of-Concepts, close-to-production prototypes).		
Planned Start Date:	June 2014	Planned Completion Date:	March 2015
Requires Functions:			

Table 58: Function 8.4.1.3

Task No:	8.4.1	Partner:	CHUV
Function No:	8.4.1.4	Leader:	Ferath Kherif
Function Name:	Define detailed technical MIP architecture and component integration plan		
Use Case:	Define how the MIP components will practically connect to result in a fully functional end-product. Consequently, identify technologies, resources.		
Planned Start Date:	November 2014	Planned Completion Date:	February 2015
Requires Functions:			

Table 59: Function 8.4.1.4

Task No:	8.4.1	Partner:	CHUV
Function No:	8.4.1.5	Leader:	Ferath Kherif
Function Name:	Set up MIP Operational processes		
Use Case:	Define incident management & user support processes, platform support, and change management.		
Planned Start Date:	July 2015	Planned Completion Date:	End March 2016
Requires Functions:			

Table 60: Function 8.4.1.5

Task No:	8.4.1	Partner:	CHUV
Function No:	8.4.1.6	Leader:	Ferath Kherif
Function Name:	Project manage the integration (continuous task)		
Use Case:	Define scope and plans, allocate tasks, build backlog and sprints, ensure the different releases of different components are well coordinated, organise and coordinate the team and follow through tasks to completion according to given deadlines and specification.		
Planned Start Date:	March 2014	Planned Completion Date:	End March 2016
Requires Functions:			

Table 61: Function 8.4.1.6

7.5 WP 8.5

Task No:	8.5.3	Partner:	FBF
Function No:	8.5.3.1	Leader:	Giovanni FRISONI
Function Name:	Build the MIP Training and Support Centre		
Use Case A:	Design and create the MIP Training and Support Centre open e-learning platform for the benefit of users.		
Use Case B:	Once the MIP Training and Support Centre are fully deployed, users will be helped to learn and share contents and procedures.		
Use Case C:	Users may also use the MIP Training and Support Centre to encourage external researchers to share knowledge collaboratively, based on their experience. The ultimate goal is to get researchers to synergistically study, talk, and learn from each other regardless of time and physical location, in order to foster knowledge about brain diseases with a new critical mass of data.		
Planned Start Date:	May 2014	Planned Completion Date:	December 2015
Requires Functions:	None		

Table 62: Function 8.5.3.1

Task No:	8.5.3	Partner:	FBF
Function No:	8.5.3.2	Leader:	Giovanni FRISONI
Function Name:	Preparing training materials		
Use Case A:	Creation of ad-hoc documentation and sandboxes regarding (I) Analysis tools (II) Data Mining tools to be printed/distributed to users during HBP courses, summer schools, workshops, etc.		
Use Case B:	Deployment of web-contents and use-cases analyses to be plugged and played directly by users in the MPI Training and Support Centre.		
Planned Start Date:	June 2014	Planned Completion Date:	December 2015
Requires Functions:	None		

Table 63: Function 8.5.3.2

Task No:	8.5.3	Partner:	FBF
Function No:	8.5.3.3	Leader:	Giovanni FRISONI
Function Name:	“Face 2 Face” & Basic "e-learning"		
Use Case A:	Face to face meetings with hospitals interested in sharing their own clinical/research databases.		
Use Case B:	Workshops hosted in research centres to show the power of the federated databases of the MIP.		
Use Case C:	Web seminar on GlobalCrossing to show users how to perform specific queries and analyses on the federated database.		
Use Case D:	GoToMeeting teleconferences to teach users the main functionalities concerning MR images analyses by exploiting the MIP Web Portal.		
Planned Start Date:	December 2014	Planned Completion Date:	December 2015
Requires Functions:	None		

Table 64: Function 8.5.3.3

Task No:	8.5.3	Partner:	FBF
Function No:	8.5.3.4	Leader:	Giovanni FRISONI
Function Name:	Professional e-learning		
Use Case A:	On line site with documentations and tutorials about all services and tools available on the MIP Web Portal.		
Use Case B:	Users perform ad-hoc quizzes to monitor their acquisition of knowledge and skills.		
Use Case C:	To collect grades, evaluations, and feedback from users to improve Web Portal contents and tools.		
Planned Start Date:	December 2015	Planned Completion Date:	March 2016
Requires Functions:	None		

Table 65: Function 8.5.3.4

7.6 Functions: Timeline

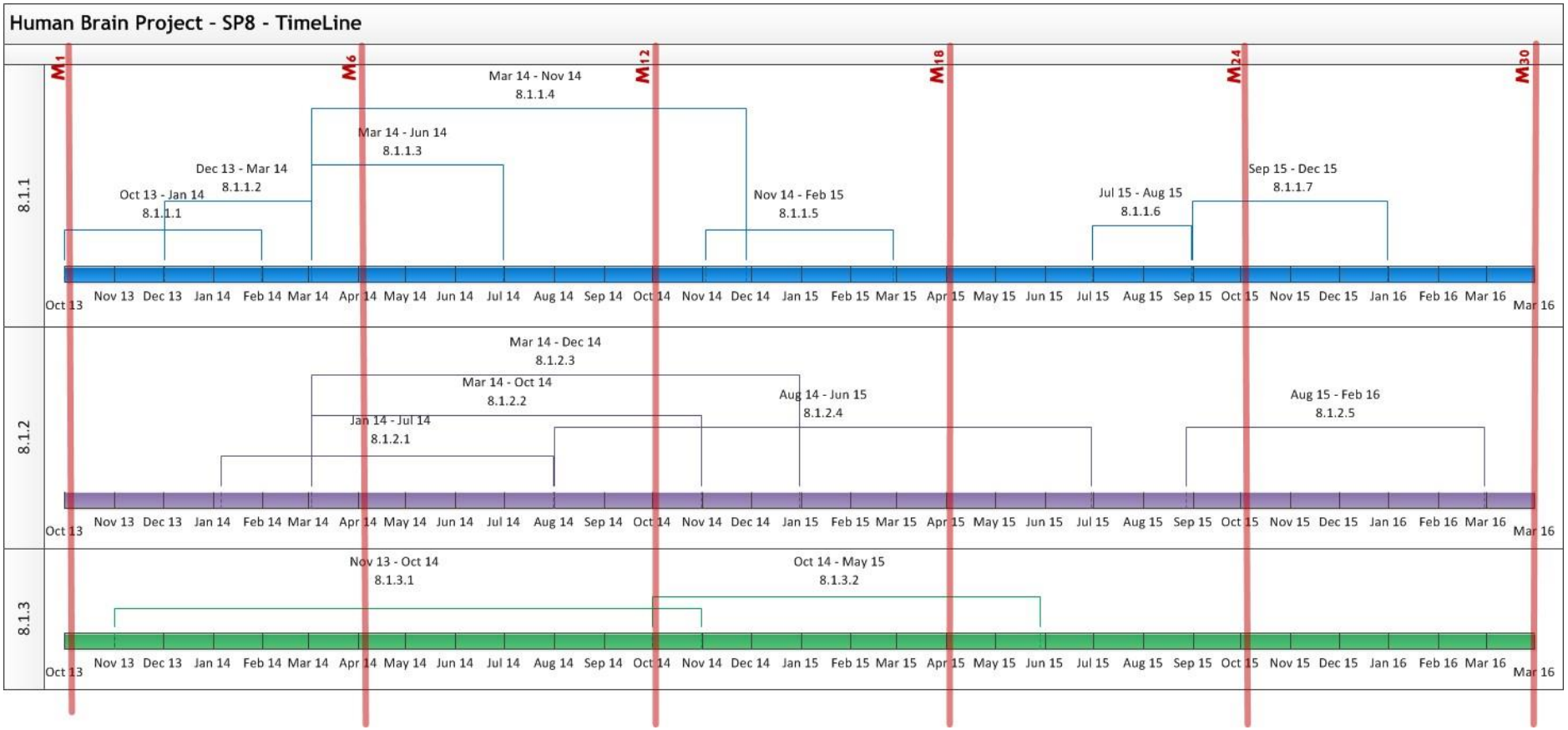

Table 66: Function Timeline 1

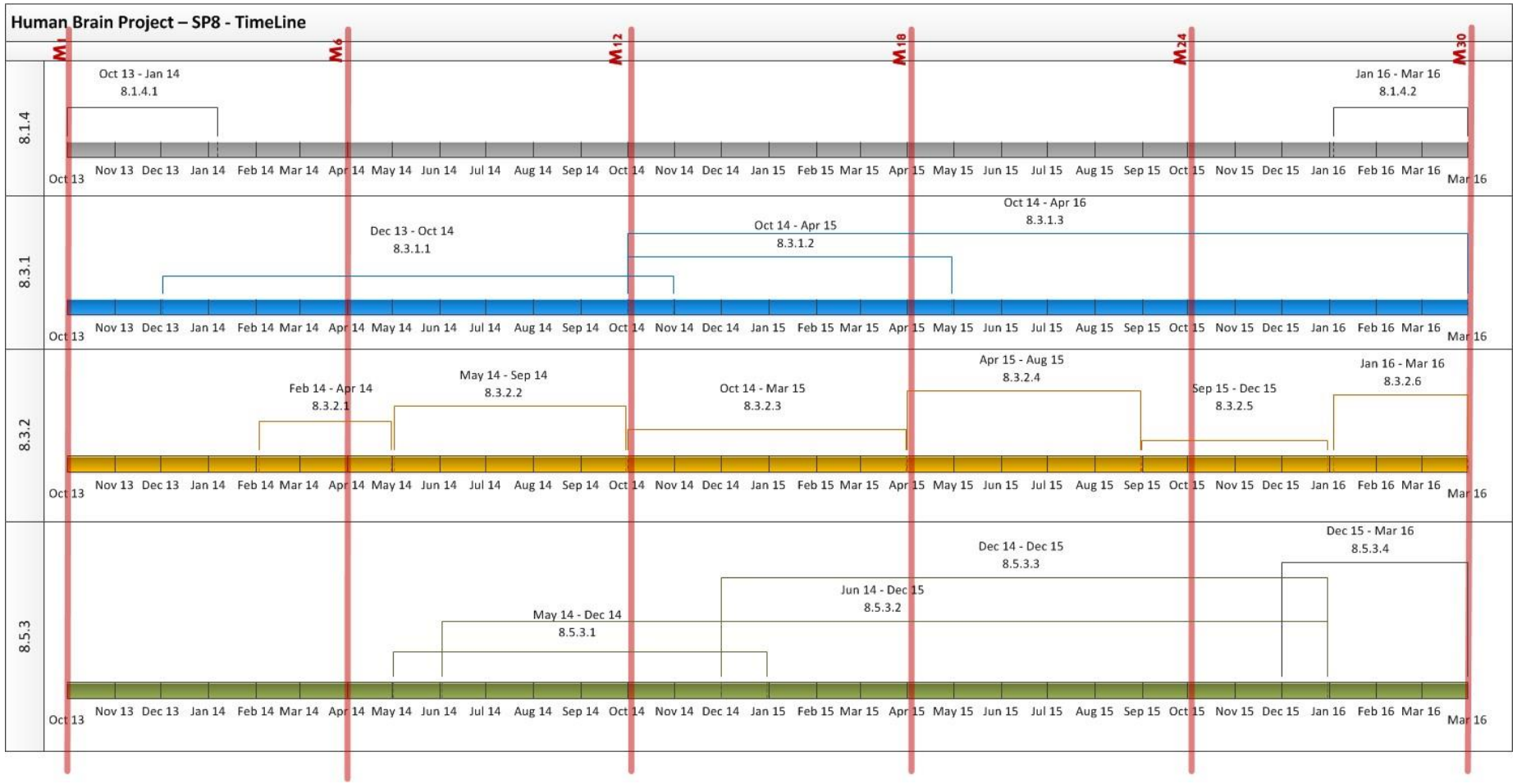

Table 67: Function Timeline 2

8. Key Performance Indicators

Ref.	Description of KPI	M12	M18	M24	M30
	Hospitals integrated into Federation:	1: CHUV	2	5	5
1	Communication initiated	17 (completed in M6)	-	-	-
	Data:				
2	Non-image data (n. of hospitals)	1 hospital: CHUV	2 hospitals: CHUV & other	2 to 5 hospitals	5 hospitals
3	Image data (n. of hospitals)		1 hospital: CHUV	2 to 5 hospitals	5 hospitals
4	Anonymisation of hospital data (n. of hospitals)	1 hospital: CHUV	2 hospitals: CHUV & other	2 to 5 hospitals	5 hospitals
5	Number of Images in the MIP		6000	8000	10000
6	Number of research collaboration agreements/data transmission agreements with large scale studies				4
7	Number of research collaboration agreements with pharmaceutical companies				2
8	Number of patient studies involved				11 (5 hospitals + 4 res. collab. + 2 pharm. comp.)
	System Functionality:	See Timeline at Section 7.4 for Functionality vs Timeline delivery			
	Interactive Analysis				

9	Number of pre-defined statistical functions	6	12	18	30

Table 68a: Key Performance Indicators

Ref.	Description of KPI	M12	M18	M24	M30
	Data Mining				
10	Number of standard features available in "Biological Signatures of Diseases"	200	500	5000	10000
11	Number of models identified and tested	2	4	6	10

Table 68b: Key Performance Indicators

Ref.	Description of KPI	M12	M18	M24	M30
	Variables, Ontologies & Provenance				
12	Number of Ontologies defined for the necessary vocabulary/datasets	4	5	5	5
13	Number of mappings between hospitals (via the Data Federation layer)		2 hospitals: CHUV & other	2 to 5 hospitals	5 hospitals
14	Number of pre-defined statistical queries for the epidemiological counts				6
	Training & Support Centre				
15	Number of workshops per year	2	2	2	2
	Software implementation:				
16	Test of usability of distributed processing engine	Initiated & validated	Completed		
17	Test anonymisation software (subcontractor) at CHUV	Completed			
18	Test for query rewriting algorithms across hospitals	Initiated	Validated	Completed	
	Federation software				

19	Software selection	3 providers evaluated	1 to 2 proof-of-concepts	1 solution implemented	Refined, enriched and fully integrated solution
	Anonymisation software				
20	Software selection	Evaluated & selected	Completed		

Table 68c: Key Performance Indicators

9. Glossary: MIP Terminology

Data Integration:

Data Integration is the process of merging data from different sources.

Data Federation:

Data Federation is a type of Data Integration. This is a process and virtual database that allows queries to heterogeneous and fragmented databases with a “no copy” and “no move” policy in relation to original source data. The virtual database contains only the Variables and the Provenance that describe the original data in the local databases at each site in the network of MIP associated infrastructures.

Biological Signatures of Diseases:

The Biological Signatures of Diseases are deterministic mathematical constructs that aim to describe both variability at the phenomenological level (clinical features with symptoms and syndromes) and that at the biological level (genetic, proteomic, etc.). The key property of a biological signature of disease is that it accounts for the fact that a symptom of brain dysfunction can be due to many biological causes (one-to-many symptom mapping) and that a biological cause can present with many symptoms (many-to-one symptom mapping). In reality, the situation is often one of many to many mappings between symptoms and biological causes. With advanced computing power and data mining, nearly exhaustive searches of a data space can be performed to identify sets of rules that describe homogeneous populations, to explain their biological data and to predict the pattern of symptoms.

Biological Signatures of Diseases are the results of a continuous dynamic data mining process of clinical data in local data sources. These results are aggregated to generate a single multi-modal description of the disease space.

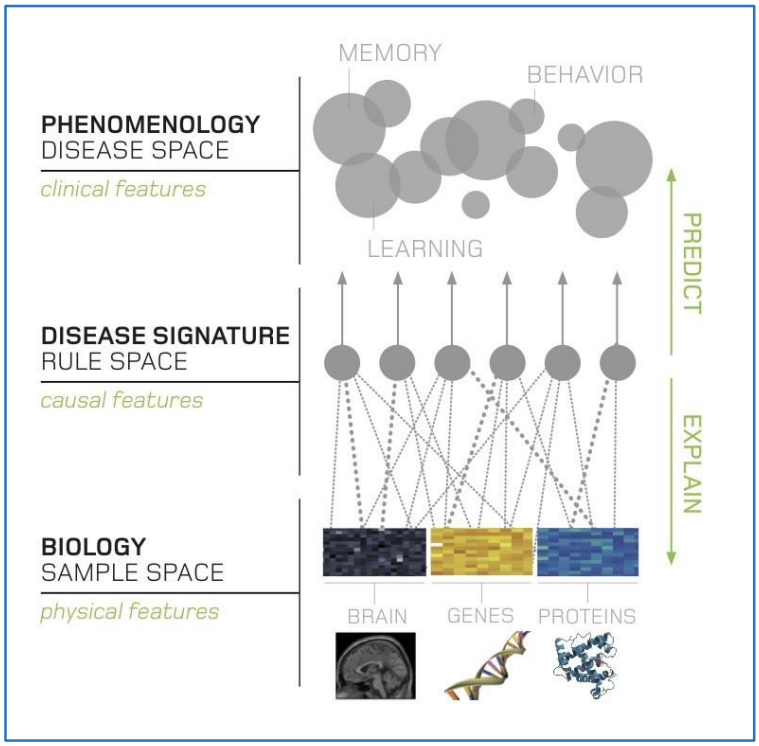

Figure 18: Biological signature of brain diseases/Continuous data mining process

Data Mining:

Data Mining is a computational process that uses machine learning tools to identify recurrent patterns in large datasets. The methods can be based on fast black-box approaches or on more power-hungry mathematical models.

Mining will redefine the multi-dimensional brain diseases space by populating it with disease signatures (constellations of parameterised biological, anatomical, physiological and clinical variables that define homogeneous clustered populations).

- The more data becomes available, the greater the discriminatory refinement.
- The result will be a continuously optimised constellation of disease signature clusters defining the disease space.

The MIP will implement data mining methods that provide causal generative models of data that are able to deal with complex, heterogeneous data.

10. References

[1] http://en.wikipedia.org/wiki/Sharable_Content_Object_Reference_Model

[2] www.valamislearning.com

[3] www.neugrid4you.eu

[4] Marnette, B., Mecca, G., Papotti, P., Raunich, S., Santoro, D. (2011) ++Spicy: an OpenSource Tool for Second-Generation Schema Mapping and Data Exchange. *PVLDB* 4(12): 1438-1441.

[5] http://www.w3.org/

[6] Bechhofer, S., van Harmelen, F., Hendler, J., Horrocks, I., McGuinness, D.L., Patel-Schneider, P.F. and L.A. Stein (Editors) (2004). OWL Web Ontology Language Reference, URL http://www.w3.org/TR/owl-ref/.

[7] Motik, B., Cuenca Grau, B., Horrocks, I., Wu, Z, Fokoue, A., and C. Lutz (Editors), (2009). OWL 2 Web Ontology Language Profiles.

[8] http://protege.stanford.edu/

[9] http://webprotege.stanford.edu/

[10] http://alignapi.gforge.inria.fr/

[11] Hernández, M., Miller, R.J., and L.M. Haas, *Clio: A Semi-Automatic Tool For Schema Mapping*, System Demonstration, ACM SIGMOD International Conference, 2001.

[12] www.fil.ion.ucl.ac.uk/spm