SP8: Software Requirements Specification (SRS) for SGA2

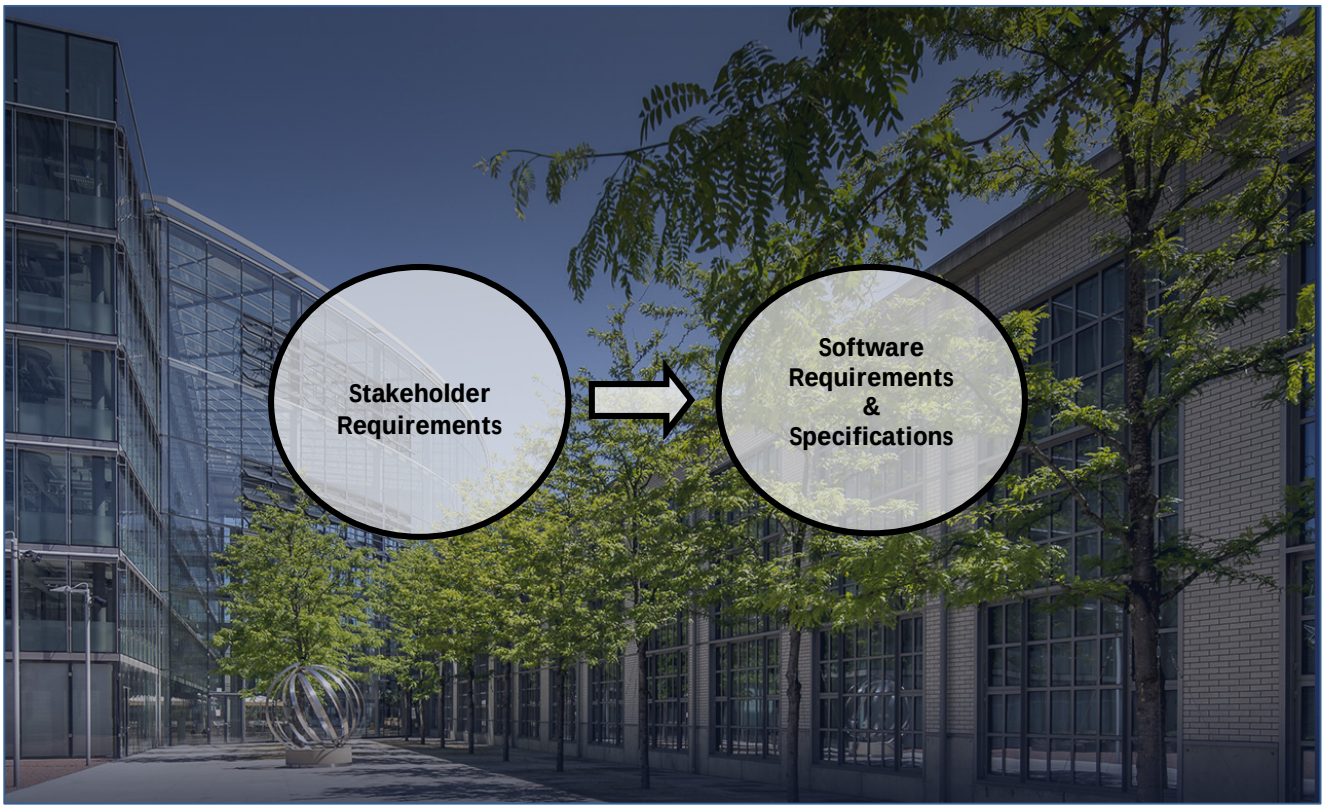

Figure 1: Software requirements and specifications flow

Figure 1. Software requirements (functional and non-functional) for the SGA2 Medical Informatics Platform that are the main focus of this Deliverable were based primarily on the discussions that took place with the key stakeholders involved during the SP8 SGA2 kick-off meeting in Geneva, at the Campus Biotech (depicted above).

Project Number:	785907	Project Title:	Human Brain Project SGA2

Document Title:	Software Requirements Specification (SRS)
Document Filename:	D8.5.1 (D52.1 D4) SGA2 M3 SUBMITTED 181217.docx
Deliverable Number:	SGA2 D8.5.1 (D52.1 D4)
Deliverable Type:	Report
Work Package(s):	WP8.5
Dissemination Level:	PU (= Public)
Planned Delivery Date:	SGA2 M3 / 30 Jun 2018
Actual Delivery Date:	SGA2 M9 / 17 Dec 2018

Authors:	Eleni ZACHARIA, UoA (P43), Evdokia MAILLI, UoA (P43)
Compiling Editors:	
Contributors:	Kostis KAROZOS, AUEB (P4), Yannis FOUFOULAS, UoA (P43)
SciTechCoord Review:	Jeff MULLER, EPFL (P1), Marc MORGAN, EPFL (P1), Yannick MOREL, TUM (P56)
Editorial Review:	Guy WILLIS, EPFL (P1)

Abstract:	This document presents the high-level software requirements and specifications of the Medical Informatics Platform (MIP) in SGA2. These requirements were based primarily on the discussions that took place during the SP8 SGA2 kick-off meeting in Geneva on 5-6 July 2018, and the Use Cases described in the Specific Grant Agreement SGA2, signed on 3 May 2018. Requirements discussed in Geneva also incorporated feedback from the SGA1 M24 review of SP8 in Stockholm, May 2018.
Keywords:	Software functional requirements, Medical Informatics Platform

Table of Contents

Table of Figures

Summary

During SGA1, a first version of the Medical Informatics Platform (MIP) was specified and developed, leading to the delivery of a functional MIP at the end of SGA1, as reported in HBP SGA1 Periodic Report (Month 1 to Month 24) - Part B.

In SGA2, the main ambitions of the Medical Informatics Platform (MIP) regarding software are the following:

- Address privacy issues regarding MIP Federate analyses
- Enrich and consolidate the analytical tools available in the MIP
- Provide an upgraded version of the MIP scaling-up its functionalities

The purpose of Deliverable D8.5.1 is to present the high-level software requirements and specifications of the MIP in SGA2. These requirements are based primarily on the discussions that took place during the SP8 SGA2 kick-off meeting in Geneva on 5-6 July 2018, and the Use Cases described in the second Specific Grant Agreement (SGA2) signed on 3 May 2018. Requirements discussed in Geneva also incorporated feedback from the SGA1 M24 review of SP8 that took place in Stockholm in May 2018.

This requirements analysis will outline high-level functional requirements (specify what the software has to do from a user's perspective), that are in agreement with the scope and planned effort for software development in SGA2. The user requirements will guide new developments in the MIP, as well as enhancements of its Components. The document will include also non-functional requirements (general criteria that the software has to meet). Finally, we present a high-level overview of the software Components of the MIP, as its architecture is evolving to capture the new requirements.

Requirements elicitation is a continuous process and as such, inputs will be gathered from all potential MIP users, as the pool of associated hospitals increases, as well as from new Partners in SP8 (in WP8.7, WP8.9 WP8.10). This feedback will be used to update D8.5.1 and ensure that all critical needs are met.

1. Introduction

The MIP aims to enable breakthrough medical progress in the field of brain diseases through federated analysis of data residing in a wide network of hospitals.

More precisely, the goal for the MIP in SGA2 is to capture clinical data from more than 30,000 patients with brain diseases. Most of the data to be incorporated within the MIP will be provided by a few large, existing patient cohorts, via signed partnerships (e.g. CREACTIVE will provide 5,000 patients with TBI, while WP8.10 will provide cohorts of several thousand patients with psychiatric disorders). Other hospitals will contribute smaller cohorts, of around 1,000 patients.

The SP8 SGA2 Medical Informatics Platform contributes to SGA2 Key Results KR8.1, KR8.2, KR8.3, and KR8.4, as described in the SGA2 Grant Agreement (GA).

2. Glossary

API – Application Programming Interface

CDE - Common Data Element (https://github.com/HBPMedical/mip-cde-meta-db-setup/blob/master/variables.json)

DAG - Directed Acyclic Graph

DICOM - Digital Imaging and Communications in Medicine

EAV - Entity Attribute Value model (https://en.wikipedia.org/wiki/Entity%E2%80%93attribute%E2%80%93value_model)

EDC - Electronic Data Capture

EEG - Electroencephalogram

EHR - Electronic Health Record

GA - Grant Agreement SGA2 - signed May 28

GDPR - General Data Protection Regulation (v5853/12, dated 27 January 2012) http://ec.europa.eu/justice/data-protection/reform/files/regulation_oj_en.pdf Applies from May 2018.

MIP - Medical Informatics Platform

NFS - Network File System

POC - Proof Of Concept

REST - REpresentational State Transfer (http://en.wikipedia.org/wiki/Representational_state_transfer)

SGA2 - Second Specific Grant Agreement

UDF - User Defined Function

3. Definitions

MIP Data Governance Steering Committee (MDGSC)	The MIP Data Governance Steering Committee is the body responsible for producing the applicable guidelines and policies as well as the Publication and Authorship policy. It is led by key stakeholders of the Human Brain Project from both clinical and ethics/legal area.
MIP	The MIP installed in each hospital based on the list of components that can be found at: https://hbpmedical.github.io/software-catalog/
MIP local	Part of the MIP installed in each hospital. Local data owners and authorized staff of a hospital access and analyse the pseudonymised datasets which are stored in a local server per hospital. This server is not connected to the MIP federate network.
MIP federate node	Part of the MIP installed in each hospital. It is connected to the MIP federate network. Its software components access the fully anonymized datasets of the recruited hospitals.
MIP federate network	The network of all authorized, active and connected MIP federate nodes. It can only be accessed through the web-based user interface of the MIP federate network by staff authorised by the MDGSC.

4. Use Cases

The GA describes three main Use Cases (UCs) for the MIP in SGA2, as well as the CDP6-related Work Package. More clinical Use Cases can be expected, as new clinicians will be introduced to the MIP in SGA2.

5. SGA2-SP8-UC001 - Advanced phenotyping of the ageing brain cognitive diseases at early stage

Description

Automated prediction of typical or mixed Alzheimer's disease (AD) using combined routine brain volumetry and cognitive assessment for a single patient.

This case will predict whether a patient at a University Memory Clinic with suspected neurocognitive disorders suffers from either typical or mixed forms of AD. We will be applying machine learning techniques, based on: Brain volumetry using T1-weighted magnetic resonance imaging, and the Montreal Cognitive Assessment (MoCA).

The dependent variables that will be used are age, gender, MoCA score, left and right temporal lobe volumes, and white matter abnormality volume. An Independent variable will be Alzheimer broad category, containing AD and mixed AD.

A Gaussian-naive Bayes algorithm will be used to classify typical AD vs mixed AD, using six predictive features considered as continuous variables (age, gender, MoCA score, left and right temporal lobe volumes, and white matter abnormality volume), with a reference dataset.

The above analysis will be stored and reproduced, if necessary, by clinicians.

To aid clinical decision-making, the MIP will provide a visualisation displaying a single patient's data alongside the previous analysis.

Challenges

To enhance the MIP's federated analytics capabilities with new algorithms (Naive Bayes), and to enhance algorithmic visualisations, to allow comparison with a single patient's data.

5.1 SGA2-SP8-UC002 - Methods for discovery of novel or refined disease models and their application to Parkinson's disease

Description

The objective of the Use Case is to combine multimodal data from different domains (genetic, gait and motor testing, clinical information and imaging) to identify disease subtype signatures, disease projection trajectories and validate the models created in large populations. Such data sets are extensive and analysis has to account for many effects and covariates, thus calling for novel approaches.

The input will be derived from large-scale federated hospital data. The data collected will include clinical information, patient and family history, genetic and biomic data, clinical assessments (such as neuropsychological examination, gait and balance testing), as well as reports on co-morbidities, such as mood, behaviour and autonomic function. Furthermore, features extracted from SPECT, PET and MRI imaging data will be analysed. The models built in one hospital will be applied and stratified using data from other federated hospitals, thereby enabling validation of the models. The large-scale data will be used to impute missing data and to adjust partial models to predict the most appropriate subtype.

The prospective results of this analysis will unveil differences between genotypes and subtypes beyond clinical phenotypes and contribute to identification of potential biomarkers indicating different disease evolution and prognosis.

Challenges

This Use Case requires feature extraction from medical images, as well as support with additional algorithms for 3-C (Categorise, Cluster, Classify) strategy. Enhancement of federated analysis with at least Random Forest, Naive Bayes, KNN, ID3.

5.2 SGA2-SP8-UC003 - Federated Analysis of large-scale intracerebral EEG data from patients with epilepsy

Description

This Use Case will be based on the work developed during the last four years by the group of Olivier David (Grenoble), within the framework of the ERC F-tract. F-Tract aimed at gathering intracerebral EEG data from patients with epilepsy undergoing stereo electroencephalography (SEEG) in various European centres. More specifically, F-tract focuses on cortico-cortical evoked potential (CCEPs) obtained while performing 1 Hz brain stimulation of intracerebral leads, while recording responses to that stimulation in all other leads (typically between 128 and 256 per patient). These CCEPs offer unique information on brain functional connectivity, including its directionality. SEEG typically samples about 1% of a patient's brain, so that every patient's data contribute to a very limited amount of brain regions. F-Tract aims at building an atlas of CCEPs for the entire brain by pooling data from several hundred patients.

SEEG-MIP represents the development of F-Tract within SP8.

It aims at achieving the following main objectives:

- Capturing CCEPs data from more epilepsy centres to significantly increase the number of cases contributing to the CCEPs atlas,
- Decentralizing the processing of SEEG and related neuroimaging data to each participating hospital in contrast with F-Tract where non-anonymized raw data are centralized before being processed,

- Pooling only anonymised CCEP data within a centralised atlas, and
- Sharing these data within the HBP and, in particular, with the Human Brain Atlas built in SP2.

Challenges

To achieve these goals, the entire flow of raw data processing will be implemented within the MIP software package and deployed together in at least 10 hospitals during SGA2. Most importantly, SEEG-MIP will not use the current MIP algorithm factory (see Section 12 - Annex). Conversely, we aim to share the following functionalities between the current MIP and SEEG-MIP:

- Storage facilities of raw and pre-processed neuroimaging and SEEG data (LORIS is being explored for this purpose),
- Generic MRI pre-processing pipelines (co-registration, segmentation, etc.)
- Web-based user interface.

These elements represent the main developments underlying SEEG-MIP integration into SP8.

5.3 CDP6 "Modelling for Drug Discovery"

From the requirements analysis so far, WP8.6 (CDP6) does not use the MIP or its data.

6. Functional Requirements (FRs)

In the Geneva kick-off meeting, priorities were established for requirements regarding the software. These are highlighted below and fall into three categories: Top priority, Important, and Desirable.

6.1 Top priority

FR1: Full anonymisation of clinical data for federation usage

Data currently include typical clinical information (e.g. diagnosis), and results from various investigations such as scores at neuropsychological tests (e.g. MOCA), laboratory findings (e.g. presence or absence of a specific gene polymorphism such as ApoE4), and regional brain volumes extracted from brain T1-MRI. In the near future, scalar values reflecting individual connectomes calculated from MRI diffusion tensor imaging through The Virtual Brain shall also be available. The number of variables considered for a particular condition or cohort might reach several hundred (e.g. epilepsy surgery cohort from the EpiCARE European Reference Network).

Privacy issues were addressed in SGA1 by designing the MIP so that it ensures data privacy "by design". During SGA1, only pseudonymised datasets were imported into the MIP, with their corresponding lookup table being stored outside the MIP. We need to move one step further during SGA2, by creating two databases in each hospital, located on different servers. The first database will be accessed through MIP local and will contain the pseudonymised datasets. The second database will contain a fully anonymised version of the pseudonymised datasets. The fully anonymised database will not have any associated lookup table and, consequently, it will not be possible to link the fully anonymised datasets back to the pseudonymised ones. Analyses performed through the MIP federate network will only have access to fully anonymised datasets, while access to pseudonymised data will be restricted to local investigators (i.e. from the hospital owning the data) using MIP-local.

Full anonymisation will be achieved through the development of an automated anonymisation tool. This tool will be applied on the pseudonymised datasets. Once the federated datasets are updated (i.e. including new data, addition of new patients or removal of patient data upon request), a new, fully anonymised database will be produced to replace the current one. The

different versions of the anonymised database will be kept using a version control system to support provenance. The life span of such versioning will be decided accordingly.

FR2: Statistical models / algorithms

Supported algorithms in the MIP fall into two broad categories:

- Available in MIP local (local algorithms), and
- Available in MIP federate network (federated algorithms).

Federated algorithms are essentially machine learning algorithms that comply with additional privacy constraints. (i.e. only aggregate results can be accessed outside a hospital node)

Currently, MIP supports a limited number of local and federated algorithms. Local algorithms are divided into three categories:

- Statistical Analysis (PCA, Linear regression, Correlation heatmap, ANOVA),
- Feature Extraction (TSNE, k-means, as well as algorithms developed by SP8 partners: TAU HEATMAPLY, TAU GGPARCI, JSI HINMINE, JSI HEDWIG), and
- Predictive modelling (SGD neural networks, SGD linear model, naive Bayes, k-nearest neighbours, gradient boosting).

Federated algorithms are the following: histogram, linear regression, k-means as well as algorithms developed by SP8 partners (model tree and regression tree). The above algorithms are also listed in the MIP web portal (https://collab.humanbrainproject.eu/#/collab/50/nav/242). The MIP is password protected; to gain access, please contact: (support@humanbrainproject.eu).

The list of local algorithms can be increased by integrating local existing libraries (i.e. scikit-learn) with the MIP. However, due to privacy constraints, this is not the case for the federated algorithms. The growth of federated algorithms list is more challenging, as they need to be implemented from scratch.

At the SGA2 kick off discussions, it was determined that, in order to support Use Cases SGA2-SP8-UC001 and SGA2-SP8-UC002, it would be necessary that the following algorithms should be available in federation: covariance matrix, PCA, ID3, Random Forest, Naive Bayes, KNN, logistic regression, ANOVA.

The federation engine Exareme[1] will be enhanced with new user-defined functions in order to support the implementation of new federated algorithms/workflows, as well as the integration of algorithms developed by other SP8 Partners. The scientific workflow engine (described below in FR3) will also offer new tools to support the algorithm implementation.

In MIP local, new algorithms will be available with the integration of existing state-of-the-art machine learning libraries (i.e. scikit-learn).

FR3: Scientific Workflow Engine

The biomedical community uses scientific workflow systems widely. Such systems help the construction and automation of scientific problem-solving processes that include executable sequences of software components (algorithms) and data flows.

A workflow engine should offer the following functionality to MIP users:

- **Management of workflow components**: MIP users must be able to register their algorithms (components) to the workflow engine.
- **Management of workflows**: users should be able to easily create, edit and delete workflows, using the registered algorithms.

[1] http://madgik.github.io/exareme

- **Execution of workflows**: the users must be able to execute a workflow, monitor its progress and also manage the execution (pause, resume or even cancel the execution).
- **User space**: users should be able to have a personal space where they can create/edit and test their workflows before publishing them.

A scientific workflow engine (Galaxy[2]) will be used to support the implementation of such complex workflows in SGA2. It moves reusability of software components one step forward, since it maintains a library of existing components that users may employ to compose their complex algorithms. Moreover, Galaxy supports the reproducibility of the processes by keeping a detailed account of the execution of a workflow: the user that executed it, the input and output dataset, and the workflow that was executed.

In order to integrate Galaxy in MIP the following work has to be done:

- Integration with MIP Federate Network. The workflow engine will be integrated with the federation engine through its analysis API (Rest API).
- PFA[3] format will be used as the layout of input/output data of both systems.
- Integration with the MIP portal. Galaxy offers a web interface to support creation of workflows. This interface will be integrated with the portal so that authorised expert users (i.e., scientists, clinicians, etc.) have access to the engine's tools.
- Implementation/integration of other supporting components of Galaxy as described in the workflow engine's architecture (Annex).

FR4: Web portal enhancements

To support analysis as described in FR2, the MIP portal needs the addition of the following:

- Visualisations of the statistical models/algorithms mentioned in FR2. The output format that has been retained is PFA (Portable Format for Analytics). The visualisations can take the following forms: static images, interactive charts and dynamic charts. As the MIP will render results in a web browser, only the following images and JavaScript chart libraries will be supported: PNG images, SVG images, Highcharts[4], and VIS[5].
- The visual Workflow Editor for the Workflow Engine described in FR3. The MIP' s Workflow Editor should provide MIP users with a graphical user interface for specifying what data to operate on, what steps to take, and what order to do them in. These workflows are produced and shared by expert users through the scientific workflow engine's drawing canvas[6].

FR5: Organizing metadata with Data Catalogue

Clinicians, researchers and data officers need a single point of truth that provides descriptive information (metadata). Description includes: Hospitals' variables (the information stored in hospitals), CDEs (Common Data Elements - elements that have been decided by clinical experts to be common to all datasets).

A Clinical Data Catalogue will be integrated in the MIP portal to cover this need.

MIP users will be able to search for variables, read variables metadata from the GUI, download hospital variables metadata, explore visually the local variables-to-CDEs transformations, and choose between versions. The MIP Data Factory team will be responsible for inserting variables

[2] https://galaxyproject.org

[3] http://dmg.org/pfa

[4] https://www.highcharts.com/demo

[5] http://visjs.org

[6] https://galaxyproject.org/learn/advanced-workflow

for a new hospital, inserting/editing/deleting variables in an existing hospital, and managing metadata versions.

For every HBP hospital, there will be descriptions and metadata for its variables. The metadata presentation will also depict their hierarchical nature. Version control will be supported by storing metadata versions in the system's database. Metadata import will be done with an XLSX file (metadata schema) which will define all variables metadata for data of a specific hospital and will be provided by the hospital itself. The XLSX file will have the following columns:

1) *csvFile*: The name of the dataset file that contains the variable
2) *name*: The name of the variable
3) *code*: The variable's code
4) *type*: The variable's type
5) *values*: The variable's values. It may have an enumeration or a range of values.
6) *unit*: The variable's measurement unit
7) *canBeNull*: Whether the variable is allowed to be null or not
8) *description*: The variable's description
9) *comments*: Comments about the variable's nature
10) *conceptPath*: The variable's concept path
11) *methodology*: The methodology the variable has come from
12) *mapFunction*: The function that transforms the variable's value into the value of its corresponding CDE
13) *mapCDE*: The corresponding CDE

Download of the metadata will be available in JSON format defined in https://github.com/HBPMedical/mip-cde-meta-db-setup/blob/master/variables.json

FR6: Visualising brain scans with LORIS

Clinicians and researchers should be able to explore and annotate MRI scans.

MIP allows running of analytics and experiments, not on actual brain scans, but on numerical features derived from the scans, describing the brain's morphology. Integrating LORIS' (Longitudinal Online Research and Imaging System) features will allow access to brain scans. LORIS[7] is a web-based data and project management software for neuroimaging research studies.

The MIP will be complemented with some specific LORIS modules dealing with image browsing, visualisation, and quality control (QC) flagging and commenting. MIP users will have access to their own hospital's MRIs through LORIS. The users will be able to: browse images, visualise images in all angles both in 2D and 3D, examine images' quality and make annotations.

The LORIS interface will have to be available through the MIP portal.

In every hospital's MIP setup, brain scans will be imported into a LORIS database instance that complies to LORIS database schema[8]. LORIS modules will run only on a MIP local which conforms fully to privacy issues as described in FR1, FR2. Importing patients' brain scans will be automated, following LORIS' pipeline scripts[9].

[7] https://github.com/aces/Loris

[8] https://github.com/aces/Loris/tree/master/SQL

[9] https://demo.loris.ca/LORIS_Imaging_Pipeline_flowchart_ZM_20150608.png

For browsing and visualising MRIs, we will use LORIS' Imaging Browser and the BrainBrowser[10] LORIS feature for submitting QC feedback[11] will be integrated as well.

For SGA2, the next steps for the Data Factory processing will not be affected by LORIS integration and use. The output from the clinicians' remarks on the brain scans is not yet taken into account and stored in the central data factory pipeline's database (see ANEX).

FR7: Mapping datasets

Heterogeneity in hospitals' data sources is an obstacle to uniform analysis. All data have to be standardised in order to be imported into the MIP.

For mapping and harmonisation of data, the Data Factory team uses MIPMap[12], a data integration and data exchange tool. Mapping and harmonisation of hospitals' datasets will be performed according to rules and correspondences defined by the hospital's clinicians, along with SP8 experts. These rules and correspondences will be published in the Data Catalogue (FR5). When a hospital provides new data for the first time, an engineer will use the MIPMap GUI to set MIPMap to enforce these rules on the local data, so that they conform to the global schema MIP uses. This configuration is called mapping-task.

MIPMap engine takes as input the source dataset along with the mapping-task and populates the target dataset which is imported to the MIP.

6.2 Important

FR8: Time-based query support

The MIP shall be used to select patient cohort samples for longitudinal studies – a selection of a group of patients who experienced the same event in a defined period of time. The MIP shall also provide data support for time-based queries and time-aligned data fetching. This support will help clinicians with a time-based selection of patient cohorts, follow-up observational studies, exploration of trends, extension of current machine learning tools for diagnostics with prognostic models, and development of disease progression models.

To support longitudinal analysis, there has to be data with clinical observations and measurements in different timestamps for the same patients. Having these available, since other machine learning algorithms need a unique tuple per patient, the system will create views suitable for every type of analysis we want to execute. Privacy issues are addressed through FR1, FR2.

FR9: Access control

Each user should have role-based access rights. For example, researchers (e.g. statisticians, scientists, clinicians and public) will be able to analyse data using the MIP, visualise their results and observe what other users have executed. System administrators will be able to monitor the system usage and load, and take the appropriate actions. To facilitate registration of datasets in the MIP, to cope with the larger number of datasets expected by the end of SGA2, data importers should be able to pre-process and upload data onto the MIP with a straightforward procedure. For verification purposes, users need to know what experiments have been done and who has accessed the data. Verification of erasure of patient data should also be provided to authorised users and therefore a monitoring tool and usage log are also required. All the user-related data should be GDPR-compliant. Authentication and authorisation will be achieved through the use of user credentials (passwords, access keys, etc.).

[10] https://brainbrowser.cbrain.mcgill.ca

[11] https://github.com/aces/Loris/wiki/Imaging-Database#8-quality-control-within-the-imaging-browser

[12] https://github.com/HBPMedical/MIPMap

Currently, HBP single sign on (SSO) is used for user authentication, and in SGA2 it will also be used for authorisation.

FR10: Web portal

The important features of the web portal are the following:

- Sharing and Collaboration tool. Allows users to share their analyses with the scientific community. Specifically, a scientific format should be provided to users to allow export of the analyses. Also, pdf export of the results of the analyses should be supported.
- Research log. Allows users to apply previously asked research questions to new imported data.

6.3 Desirable

FR11: Frequently asked questions (FAQ) / Guidelines

The users of the MIP will benefit from a list of answers to all questions that have already been asked, and from more detailed guidelines/help. It would be useful to add these features to the web portal.

7. Non-functional requirements (NFRs)

7.1 Top priority

NFR1: Automated clinical data extraction pipeline

In order to be imported into the MIP, clinical data has to be run through the Data Factory pipeline, components of which are described in ANNEX. The processing steps have to be interlinked, scheduled and automated so as to expedite data import. Requirements are:

- Schedule and log the data processing
- Guarantee data quality
- EHR data and imaging data are generated from different branches of the pipeline. When they are stored into a relational database, they need to be linked sequence-irrelevantly, meaning it should support importing either of them first.

The Data Factory scripts and processes will be executed through Airflow[13]. Airflow authors workflows as DAGs of tasks. Its scheduler executes the tasks on an array of workers while we are able to monitor the execution via its interface.

For the quality of the data, we will incorporate in the pipeline our Quality Control tools for outliers and error detection. As a first step, there will be some profiling tools that generate statistical reports for the input datasets. In addition to that, we will develop a data cleansing tool for tabular data that will give recommendations for value corrections to hospital personnel.

NFR2: Scalability

The MIP currently supports the following:

[13] https://airflow.apache.org

- Local execution supported by Woken[14], which uses well-established scalable technologies like Chronos[15], Mesos[16], and Docker[17] to integrate third-party algorithms.
- Federated execution supported by Exareme is based on local processing engines that process the raw fully anonymised data and return aggregate results to the federation node. The processing engine is built on top of SQLite with Python extensions. SQLite[18] is an embedded, serverless and portable DBMS, without background threads or processes, that allows fast data access due to its storage format (reads 35% faster than the file system). The system supports databases up to 140TB in each hospital and is able to run up to 100,000s SQL statements per second. Moreover, in the processing engine, both Python and SQLite run in the same process, eliminating the communication cost.

The MIP should be able to scale up to more than 30 hospitals by the end of SGA2.

While the system is already scalable, its scalability will be further increased during SGA2 by using:

- Python's JIT compiler[19] (instead of interpreter)
- Fusion of UDFs (User Defined Functions) in queries at runtime.

8. Quality

Software quality in SGA2 will be ensured by documentation and dissemination of technical and operational standards, quality control and testing standards for software that will be developed/modified by SP8, according to the HBP Software Engineering and Quality Assurance Approach[20].

9. Relations to other platforms

9.1 The Virtual Brain neuroinformatics platform

'The Virtual Brain' is an open-source simulation platform that allows operators to combine experimental brain data from a wide range of sources in order to improve their understanding of the brain's underlying mechanisms. By entering data from an individual patient into the model, operators can produce personalised brain models.

The MIP in SGA2 will address the challenge of integrating some of its data into the Virtual Brain, which will allow researchers to identify the complex interactions that contribute to brain function.

9.2 SEEG MIP

The MIP - SEEG MIP integration is described in Section 5.3 SGA2-SP8-UC003 - Federated Analysis of large-scale intracerebral EEG data from patients with epilepsy.

[14] https://github.com/LREN-CHUV/woken

[15] https://mesos.github.io/chronos

[16] http://mesos.apache.org

[17] https://www.docker.com

[18] Richard Hipp, "SQLite: The most used and most misunderstood database", SIGMOD 2017 Systems Award

[19] https://pypy.org

[20] D11.3.3 (D62.2, D17 - SGA1 M10) HBP Software Engineering and Quality Assurance Approach

9.3 Blue Brain Nexus

A data-driven and ontology-reference model is required, to facilitate understanding and organisation of the assembled data inside the MIP. We will explore Blue Brain Nexus[21], a data management platform based on a knowledge graph which is already used by the HBP's SP5 Neuroinformatics Platform (NIP). An effort will be made to benefit to the fullest from Blue Brain Nexus' capabilities and semantically enrich datasets in the MIP. There will be a Proof of Concept Deliverable for storing and organising EHR and imaging data using the Blue Brain Nexus knowledge graph. The integration of the Data Factory pipeline (see ANNEX) and the Blue Brain Nexus will proceed according to the POC's results.

10. Accessibility

All working/completed versions of SP8 Medical Informatics Platform SGA2 software components source code are available on the GitHub[22] development platform. The HBP MIP source code can be found at the following address: https://github.com/HBPMedical. This is an open access repository.

11. Necessary Parallel Activities

In order to incorporate the functional requirements mentioned above into the next versions of the MIP, the following activities should be carried out in parallel with software development and integration:

- Recruitment of new hospitals (T8.2.2)
- Registration of new datasets (Hospitals, CREACTIVE consortium, epilepsy centres - WP8.2)
- Import of data (WP8.2)
- User acceptance tests (T8.2.3)
- Requirements and Specification revision (WP8.1, WP8.5)

12. Conclusion / Outlook

The MIP will evolve in SGA2 and will be enhanced with new federated analytics capabilities, such as new algorithms (FR2), enhanced visualisations and workflow management. New types of data will also be supported (The Virtual Brain, EEG). Additionally, in SGA2, the MIP will strengthen privacy guarantees (FR1, FR2) regarding MIP-federated analyses. By combining the new features described in this document with the goal of installing the MIP in over 30 hospitals by the end of SGA2, we strongly believe that, by then, the MIP will reach the goal of being an essential open-access component of clinical neuroscience research in Europe.

[21] https://github.com/BlueBrain/nexus

[22] https://github.com

13. Annex

13.1 MIP logical architecture

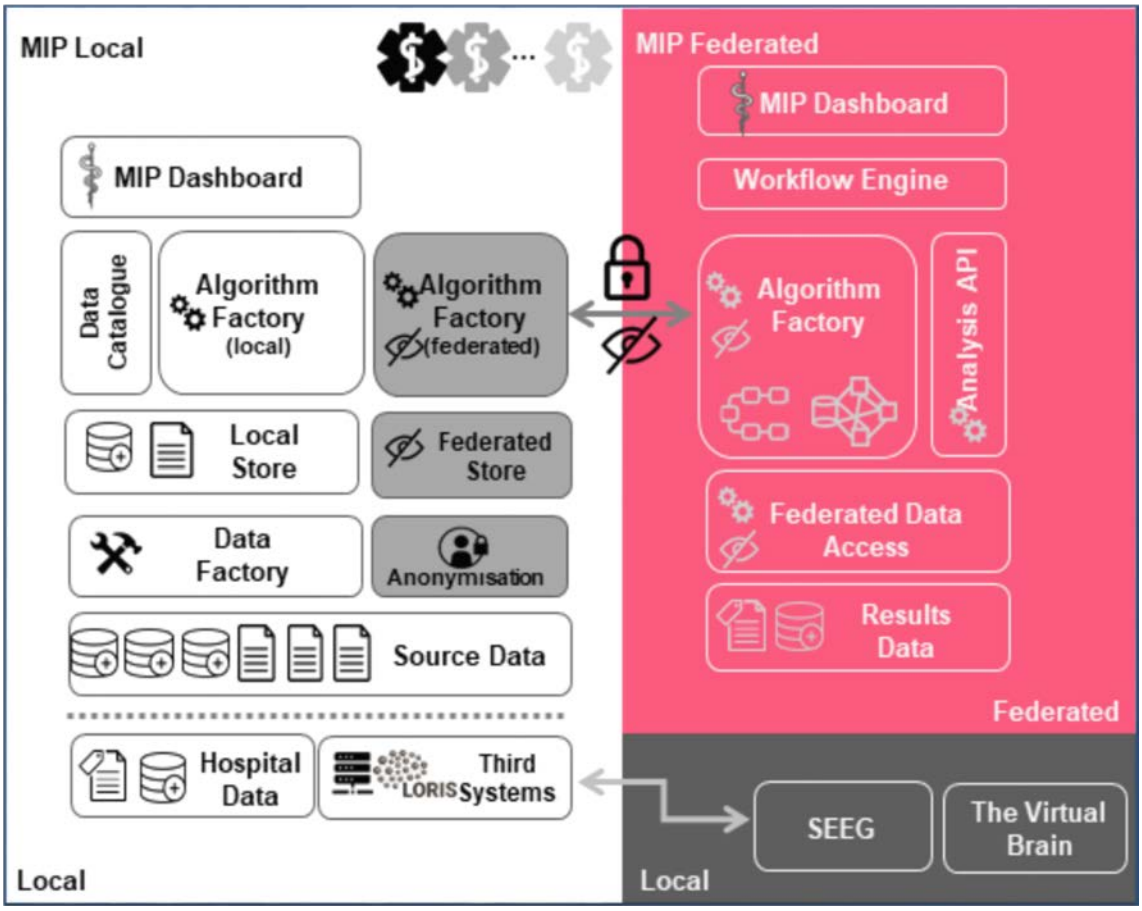

Figure 2: MIP condensed logical architecture

The above schema illustrates the logical architecture of MIP in SGA2. The grey coloured MIP Local Components represent the anonymised federated store / database that is described in FR1.

An expanded logical architecture will follow in future Deliverables; this figure is not intended to guide development, but rather to communicate MIP's key characteristics and operation, in combination with MIP processing flows in Figure 3.

According to the schema, there are two different data stores. The local store / database contains the original pseudonymised hospital data (as it is explained in FR1) that are accessible through the MIP Local, while the federated store / database contains fully anonymised data that are available to the MIP Federate Network.

The MIP offers algorithms that run on local and/or federated data. The federated network supports the federated data access, analysis and the creation of scientific workflows, while the MIP Local supports the execution of algorithms on local data.

Both the MIP Local and MIP Federate networks have their own dashboards. While this is a unique software system, it may offer different algorithms that process local or federated data store accordingly.

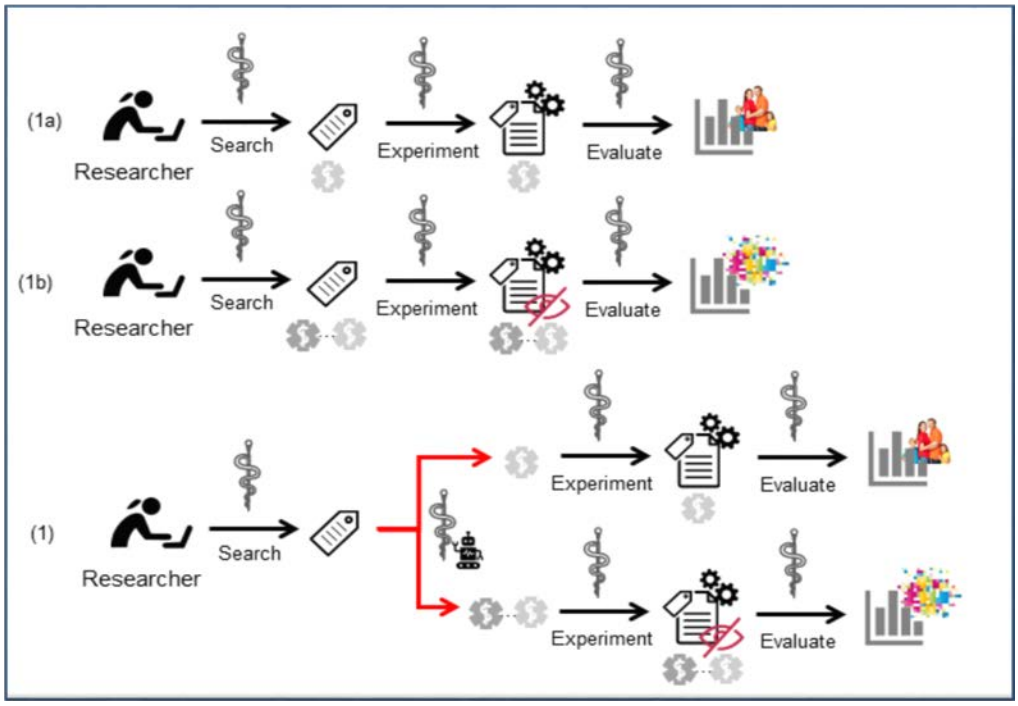

Figure 3: MIP processing flows

Figure 3 shows the analysis processing flows in the MIP. Flow (1a) represents the flow for an experiment that runs in one hospital, while flow (1b) represents the flow of an experiment that runs in multiple hospitals and addresses additional privacy concerns. In flow (1), we show that there is a single point of entry to the platform. The invocation of the appropriate processing flow (local, federated) is data-driven; when a user selects remote data, the evaluation moves to federated mode.

To support the above architecture (see Figure 2), the implementation/enhancement of the following related software components are required, namely: 1) Workflow engine, 2) Federated engine and 3) Anonymisation.

The scientific workflow engine is responsible for the orchestration and execution of the scientific workflows. These workflows will be designed using the Galaxy editor, and their metadata and definition will be stored in the Workflow Library. An overview of the workflow architecture is depicted in Figure 4.

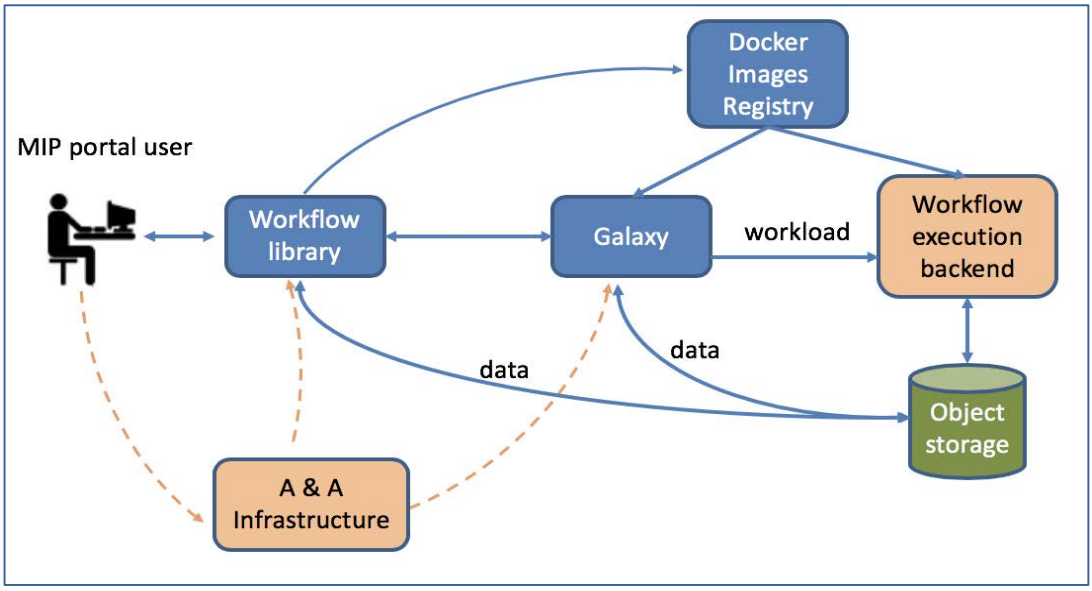

Figure 4: Workflow architecture overview

The *modus operandi* of workflow execution within this architecture is the following: The MIP users will use the portal to search for existing datasets, upload their own data, design workflows and submit them for execution to the workflow engine. Authentication actions rely on the Authentication and Authorisation Infrastructure supporting the platform.

In the MIP, processing components or whole consolidated applications are wrapped as Docker images and are registered on the platform's specific Docker Registry; the respective metadata for them are stored in the workflow library.

In all cases, the Docker images should follow a specification that facilitates the integration and execution in MIP. Specification for Docker images will be the based on the specification developed in SGA1 and extended, if needed. The registered processing components are available in the Galaxy workflow editor and can be used to create workflows by connecting them and setting component parameters (mandatory or not).

Galaxy will be used for managing the execution of applications and workflows. The actual execution takes place in the Workflow execution backend of the architecture. Finally, an object storage is providing a permanent cloud-based, large-capacity storage for storing input data and output results of the workflow execution.

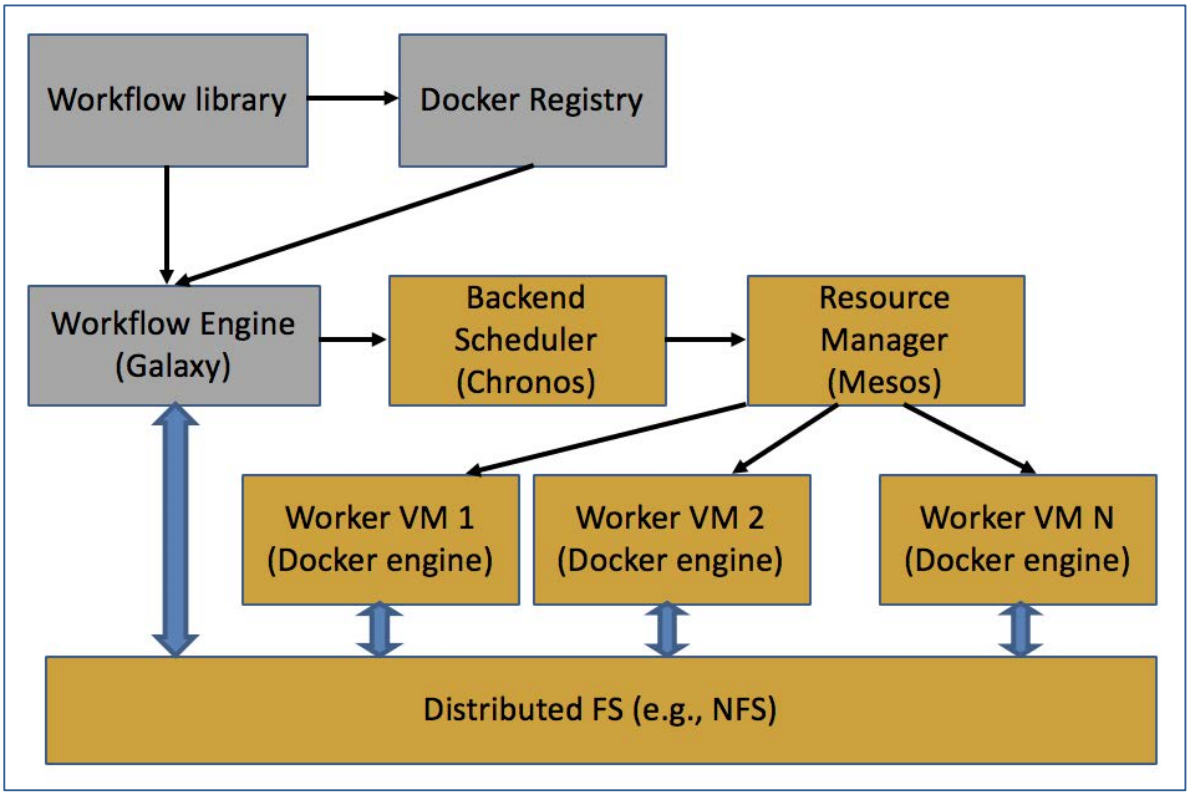

Figure 5: Workflow execution backend architecture

The main components of the workflow execution backend (Figure 5) are:

- **Back end Scheduler:** Responsible for receiving the workflows, managing priorities, and negotiating with the resource manager for the actual resources (worker VMs) needed for running each step of the application. The scheduler component will be implemented using Chronos.
- **Resource Manager:** Is aware of the resources made available by the cloud infrastructure. Allows pooling of VMs, network resources, storage, etc. Communicates with the scheduler allocating resources required every time by a specific workflow. Currently, the resource manager will be implemented using Apache Mesos.
- **Worker VMs:** Are responsible for the actual execution of each step (component) of the workflow.

- **Shared File System:** Worker VMs and Galaxy need to have a shared file system for sharing data. For this reason, a distributed File System needs to be deployed using technologies like NFS.

Federated Engine (Exareme - supports Federated Data Access)

- Enhancements for algorithm implementation (see FR2)
- Scalability features (see NFR2)

Anonymisation

Anonymisation in the MIP is covered in FR1. Privacy awareness issues during the implementation/integration of a federated algorithm are addressed in FR2.

13.2 Data Factory

The Data Factory is responsible for integrating, processing and importing hospital data into the MIP. The updated SGA2 version of the Data Factory depicted in Figure 6 contains newly added tools, such as REDCap[23], LORIS[24], and Blue Brain Nexus knowledge graph, along with the Components developed in SGA1 (part of the MIP version described in HBP SGA1 Periodic Report (Month 1 to Month 24) - Part B.), such as the Brain features extraction tool and MIPMap[25].

[23] https://www.project-redcap.org

[24] http://lorisdb.github.io/

[25] https://github.com/aueb-wim/MIPMap

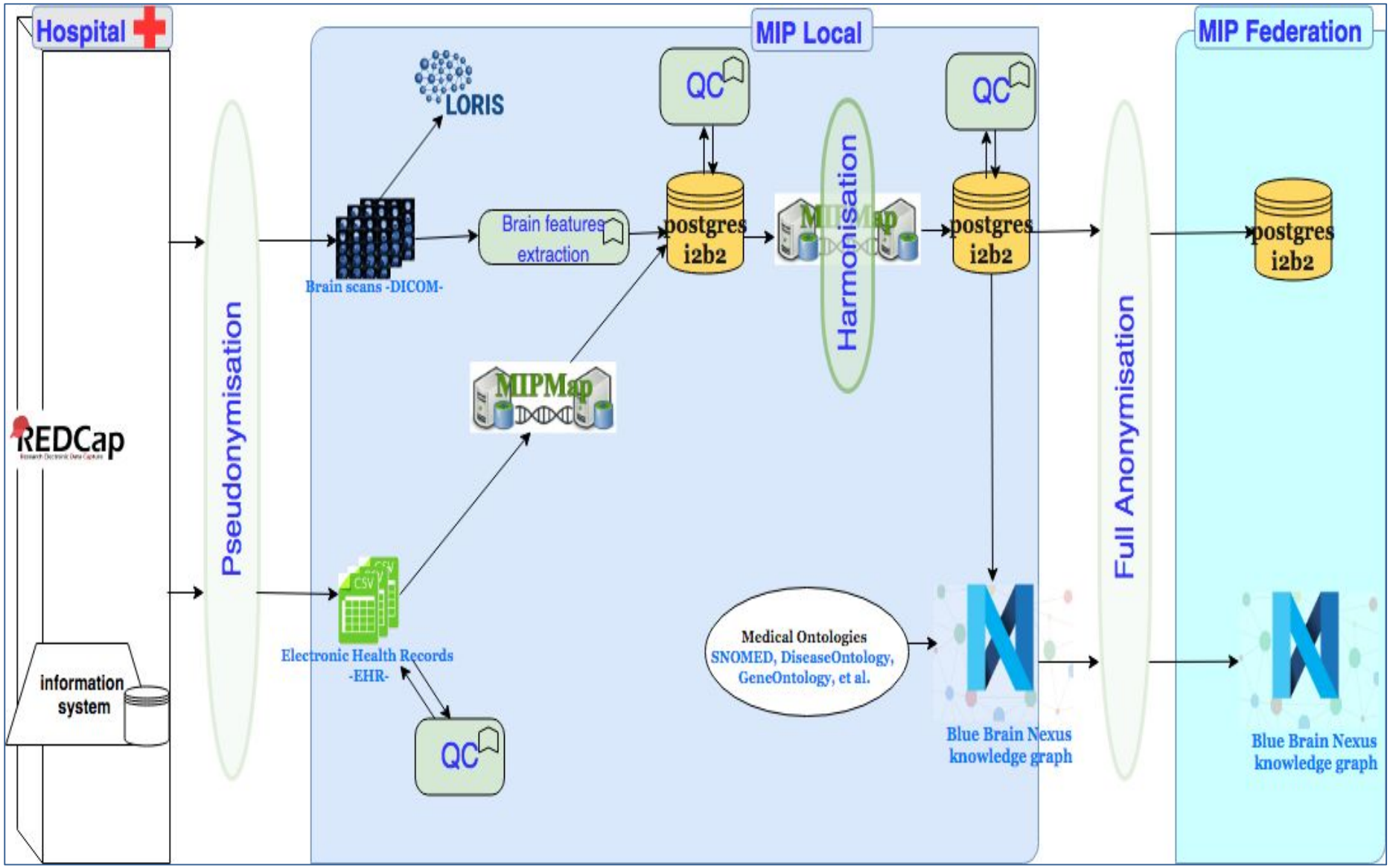

Figure 6: Data Factory pipeline

such as REDCap and other electronic data capture (EDC) systems. A proof of concept (POC) for automatic extraction of clinical data for a specific hospital will be implemented.

Before local data leave a hospital's information system, they will be de-identified (pseudonymised) in that hospital. When entering the Data Factory pipeline, brain scans (DICOMs) will be processed by John Ashburner's tool[26] which generates their brain morphometric features. These features have numerical values and will be stored in an I2B2 schema postgres database. EHR records will also be stored in the same database and linked to the imaging features after having been mapped to the I2B2 schema by MIPMap. This database will be the hospital's capture database since it captures data values as they are. Afterwards, data will again be processed by MIPMap which will apply the harmonisation rules (defined by clinicians) to create CDEs. A new I2B2 database will be populated having all harmonised values therefore it will be referred to as the harmonisation db. Data from this harmonisation db will be imported into the hospital's MIP Local node. The process from the pseudonymisation up to this step was designed in SGA1. All other Components mentioned in this section are new SGA2 requirements.

Only fully anonymised datasets are imported into the MIP Federate node, with no possibility to be linked back to the pseudonymised dataset, as explained in FR1.

Throughout the whole Data Factory processing pipeline, our quality control tools (also mentioned in NFR1 in Section 6.1) will have a critical role, generating statistical reports for imported data and locating outliers. When problematic or incomplete data are encountered, the hospital personnel will be prompted to make corrections, where possible. As far as the original DICOM images are concerned, except for the Imaging Metadata Processor, there will also be a LORIS module which, as stated in FR6, will facilitate image inspection from the clinicians.

Finally, as described in Section 8, there will be a proof of concept for storing and organising EHR and imaging data, using the Blue Brain Nexus knowledge graph.

[26] https://www.fil.ion.ucl.ac.uk/~john/