

Figure 1: The Medical Informatics Platform (MIP) links brain-science research, clinical research and patient care into an accessible space, providing the scientific and clinical community with the infrastructure and tools to improve knowledge, diagnosis, early prediction and innovative treatment of brain diseases. The MIP collaborates with a core network of five European-wide hospitals to tailor its services to health and science users.

Project Number:	720270	Project Title:	Human Brain Project SGA1
Document Title:	D8.6.2 SP8 Medical Informatics Platform - Results for SGA1 Period 1		
Document Filename:	D8.6.2 (D48.2 D23 - SGA1 M12) ACCEPTED 20180709 PU.docx		
Deliverable Number:	SGA1 D8.6.2		
Deliverable Type:	Report		
Work Package(s):	WPs 8.1, 8.2, 8.3, 8.4, 8.5, 8.6		
Dissemination Level:	PU (= Public) with CO (= Confidential) Annexes		
Planned Delivery Date:	SGA1 M12 / 31 March 2017		
Actual Delivery Date:	SGA1 M14 / 23 May 2017, Submission to EC 30 May 2017; Accepted 09 Jul 2018		
Authors:	Ferath KHERIF, CHUV (P27) Mira MARCUS-KALISH, TAU (P57)		
Compiling Editors:	Tea DANELUTTI, Jacek MANTHEY CHUV (P27)		
Contributors:	SP8 WP leaders, SP8 task leaders		
SciTechCoord Review:	EPFL (P1): Jeff MULLER, Martin TELEFONT, Marie-Elisabeth COLIN UHEI (P47): Martina SCHMALHOLZ, Sabine SCHNEIDER		
Editorial Review:	EPFL (P1): Guy WILLIS, Colin McKINNON		
Abstract:	<p>This report provides the first year's achievements of SP8. Under the new leadership the work had to focus on delivering a complete MIP solution with high clinical impact. Most of the work involved implementation of the Fast Track strategies and delivery of the first biological signatures of diseases beyond dementia. To produce a sound infrastructure, we built upon both SP8's internal expertise and on hospitals' personnel expertise. SP8 developed further collaborations with other SPs and the neuroscience community to make use of state-of-the-art neuroscientific tools for the analysis of clinical data. SP8 also established links with additional medical platforms supported by further EU projects to mutualise data and resources. The Medical Informatics Platform (MIP) has already generated wide interest and thanks to its federated infrastructure has the potential to be a critical clinical data analysis hub.</p>		
Keywords:	Medical Informatics Platform, SP8, Biological signature of disease		

Table of Contents

1. SP Leader's Overview	6
1.1 Key Personnel	6
1.2 Progress	6
1.3 Deviations	7
1.4 Impact of work done to date	7
1.5 Priorities for the remainder of the phase	9
2. WP8.1 - Federated Clinical Data Infrastructure (FCDI)	10
2.1 Key Personnel	10
2.2 WP Leader's Overview	10
2.3 Priorities for the remainder of the phase	11
2.4 Milestones	12
2.5 T8.1.1 - Infrastructure to Support Just-in-Time Analytics on Raw Medical Data	13
2.6 T8.1.2 - Infrastructure for Distributing Query Engine Locally and over Hospital Hubs	14
2.7 T8.1.3 - Installation and Adaptation of Hospital Bundle at Local Hospitals with Upgrades and Support	16
2.8 T8.1.4 - Data Integration	19
2.9 T8.1.5 - Distributed Query Workflow Engine	23
2.10 T8.1.6 - SQL-Based Data Mining and Query Templates	24
2.11 T8.1.7 - Management of Query Templates and Workflows	26
2.12 T8.1.8 - Hospital Bundle Integration Strategy and Business Model	26
3. WP8.2 - Data Selection and Community Engagement	28
3.1 Key Personnel	28
3.2 WP Leader's Overview	28
3.3 Priorities for the remainder of the phase	28
3.4 Milestones	29
3.5 T8.2.1 - Data Selection and Governance	30
3.6 T8.2.2 - Hospitals and Information System Departments Relationship Management	31
3.7 T8.2.3 - Research Initiatives and Community Engagement	32
4. WP8.3 - Data Features, Tools and Biological Signatures of Disease	35
4.1 Key Personnel	35
4.2 WP Leader's Overview	35
4.3 Priorities for the remainder of the phase	35
4.4 Milestones	37
4.5 T8.3.1 - Tools to Mine Replicable Selection and Integration of Hierarchical Features, Inter and Across Domains using FDR	39
4.6 T8.3.2 - Developing Methods for High-Dimensional Data with Possible Informative Missing Values	41
4.7 T8.3.3 - Introducing Selective Inference into Dimensionality Reduction and Clustering Methods	43
4.8 T8.3.4 - Statistical Methods for "Disease Signature" Confidence Assessment	44
4.9 T8.3.5 - Methods for Distributed Rule-Based Disease Signature Discovery	45
4.10 T8.3.6 - Methods for Redescription Mining	46
4.11 T8.3.7 - Methods for Heterogeneous Networks	47
4.12 T8.3.8 - Methods for Disease Progression Modelling	48
4.13 T8.3.9 - Ontologies for Describing Data on Neurological Diseases in Patients	50
4.14 T8.3.10 - Methods for Linkage of Local SNP Data (Individual SNPs) to Imaging Data through SNP	51
4.15 T8.3.11 - Brain Morphological Features	52
4.16 T8.3.12 - Genetic, Proteins and Neurological Features	54
5. WP8.4 - Theory, Disease Models & Big Data Engineering	56
5.1 Key Personnel	56

5.2	WP Leader's Overview	56
5.3	Priorities for the remainder of the phase	56
5.4	Milestones	57
5.5	T8.4.1 - Brain Scale High Performance Deep Phenotyping	58
5.6	T8.4.2 - Brain Scale Disease Bayes Modelling	60
5.7	T8.4.3 - Tools for Macro- to Micro-Scale Data Analysis and Atlasing	62
5.8	T8.4.4 - Case Studies in Discovering Disease Signatures and Modelling Disease Progression	64
5.9	T8.4.5 - Large-Scale Data Analytics on Massively Parallel Architecture	66
6.	WP8.5 - The Medical Informatics Platform.....	67
6.1	Key Personnel.....	67
6.2	WP Leader's Overview	67
6.3	Priorities for the remainder of the phase	68
6.4	Milestones	69
6.5	T8.5.1 - Web-Based Medical Data Analyses Foundation.....	71
6.6	T8.5.2 - Web API and Microservice Architecture for Community-Driven Data Analyses and Workflows.....	74
6.7	T8.5.3 - Integration and Technical Coordination of WP8.5, Integration into Collaboratory.....	86
7.	WP8.6 - Scientific Coordination	89
7.1	Key Personnel.....	89
7.2	WP Leader's Overview	89
7.3	Priorities for the remainder of the phase	89
7.4	Milestones	90
7.5	T8.6.1 - Scientific Coordination and SP Coordination	91
7.6	T8.6.2 - Medical Informatics Platform Strategy and Business Model	91
8.	Publications	95
9.	Dissemination.....	96
10.	Education	97
11.	Ethics	98
12.	Innovation	98
12.1	MIP in the data sharing world.....	98
13.	Open Research Data.....	99
13.1	Open Research data delivered by the project	99
13.2	Open Research data to be made accessible by the project	99

List of Figures

Figure 1:	The Medical Informatics Platform (MIP) links brain-science research, clinical research and patient care into an accessible space, providing the scientific and clinical community with the infrastructure and tools to improve knowledge, diagnosis, early prediction and innovative treatment of brain diseases. The MIP collaborates with a core network of five European-wide hospitals to tailor its services to health and science users.	1
Figure 2:	FCDI Node	15
Figure 3:	Hospital Hub	16
Figure 4:	Overlay Network	18
Figure 5:	Example of common definitions for two variables across datasets	30
Figure 6:	Measurements of stability between healthy mutations carriers and healthy subjects	40
Figure 7:	Variables matching	41
Figure 8:	Missing values in hospital's clinical measurements	42
Figure 9:	Genotype vs symptoms vs dopamins.....	44
Figure 10:	Genotype P-values across all families	44
Figure 11:	Redescriptions (derived from different parts of the ADNI dataset), each describing at least 100 patients.	47

Figure 12: Overview of the proposed methodology for heterogeneous network	48
Figure 13: Example option tree learned on test data. The highlighted area represents the best tree.	49
Figure 14: Excerpt from the mid-level ontology for describing data on neurological diseases in patients	51
Figure 15: Screenshot from the developed matlab heatmap generation module	52
Figure 16: Set and variability modelling on a dataset example.	54
Figure 17: Atlas of TLE brain features	59
Figure 18: Atlas of PD brain features	60
Figure 19: Atlas of healthy ageing	60
Figure 20: Association between cognitive decline, personality traits and biological changes in the Medial Temporal Lobe (MTL).	61
Figure 21: qMRI maps before and after optimisation of qMRI software.....	64
Figure 22: MIP overall architecture.....	68
Figure 23: MIP-Local functional architecture showing Data Factory, Database bundle, Algorithm Factory and Web Analytics.	71
Figure 24: Data Factory using Airflow engine workflows for processing EDS, PPMI and CLM datasets	76
Figure 25: Image processing in Airflow engine	80
Figure 26: Number of datasets per country	100
Figure 27: Types of data in percentage.....	100

List of Tables

Table 1: Milestones for WP8.1 - Federated Clinical Data Infrastructure (FCDI)	12
Table 2: Milestones for WP8.2 - Data Selection and Community Engagement.....	29
Table 3: Milestones for WP8.3 - Data Features, Tools and Biological Signatures of Disease	37
Table 4: Milestones for WP8.4 - Theory, Disease Models & Big Data Engineering.....	57
Table 5: Milestones for WP8.5 - The Medical Informatics Platform	69
Table 6: Milestones for WP8.6 - Scientific Coordination.....	90



1. SP Leader's Overview

1.1 Key Personnel

Subproject Leader: Ferath KHERIF (CHUV)

Subproject Deputy Leader: Mira MARCUS-KALISH (TAU)

Subproject Manager: Tea DANELUTTI (CHUV)

1.2 Progress

The mission of the new SP8 leadership is to maintain SGA1 engagement while completing unmet goals from the Ramp-Up Phase (RUP). All deliverables from the RUP were resubmitted and accepted. Above all, new strategies were developed to ensure success of the Medical Informatics Platform (MIP) as detailed below.

MIP Platform building and deployment:

To overcome the shortcomings of the initial design and recover from delays in deployment resulting from the RUP as well as comments from the June 2016 review, a Fast-track plan was defined and approved by Expert Review Report of November 2016. In the Fast-track plan the following actions are underway:

- Improvement of the Platform's architecture and elimination of redundant components and work, using proven technology and streamlining software development. The improvements automate manual steps (e.g. image pre-processing) and define the MIP-Local and MIP-Federation levels. The architecture changes are currently being implemented into the MIP and reflected in the continuous releases.
- Fast deployment of the MIP-Local and MIP-Federation to the five collaborating hospitals by involving hospital personnel (IT and clinicians) at all steps of the process. The complete plan, i.e. data, software, hardware, security needs, collaboration efforts expected and ethical agreement needs was devised with the hospitals representatives. The processes are summarised in the Evaluation Agreement (legal agreement between SP8 and Hospitals for the deployment process and the evaluation period of MIP by the local clinicians). MIP deployment is on track (see details below). The evaluation agreement can be used for other hospitals in the future.
- Strengthening the infrastructure: The infrastructure delivering the central service is located in the CHUV, Lausanne and comprises several machines along with a mini HPC, providing a centralised service for end users (in hospitals, research institutes or outside of scientific establishments). The infrastructure was scaled-up to accommodate access to the research data and the federated network.

MIP platform software development:

SP8 delivered v2 of the MIP with important new functionalities including revised variable exploration, the "Experiments Builder" module allowing to apply parameters and cross-validation to the MIP data mining methods, new data mining methods to MIP Algorithm Library (incl. Naïve Bayes, K-nearest neighbours) and overall increased usability.

Disease modelling achievements:

SP8 made significant progress to bridge neuroscience, clinical research and clinical practice with the MIP. We linked clinical and fundamental neuroscience to produce clinical benefits by 1) adding Neuroimaging-based biomarkers to the existing multiple clinical features, to facilitate interpretation of clinical symptoms and to bridge gaps between the different data scales, and 2) by developing Use Cases in collaboration with other SPs to ensure transfer of fundamental neuroscience expertise. The most significant results are:



- Published methods to identify disease signatures for application in Dementia, Parkinson's and Epilepsy (see Publications section);
- Published results on new disease models showing significant interaction between cognitive state and risk factors and predicting MTL abnormality.
- A new atlas of brain diseases (PD, AD and ageing) using methods developed in the project was generated on a large cohort of over 1000 subjects. Combination of this atlas with other brain atlases (e.g. brain region) provide neuroscientific knowledge.

Deliverables:

Following the Expert Review comments, the revised D8.6.1 including “Document-1” was delivered. Based on the Fast-Track plan, it details MIP-Local and MIP-Federated deployment strategy and process, as well as the Agreement sent to Hospitals. Before submission the Medical Platform Advisory Team approved “Document-1”.

Furthermore, the document addresses foreseeable risks associated with this plan and methods enabling risk control to ensure staying on track. The RASCI matrix as well as the detailed work plan was the result of a consultative work across the whole SP8 team.

1.3 Deviations

Additional effort was necessary to define and execute the Fast-Track plan in parallel to planned SGA1 activities, with effect on some of the Tasks and Deliverables.

It took SP8 some time to define the hospitals' tasks and therefore the budget required, and to prepare amendments and legal contracts for inclusion of the hospitals as third parties: The budget reallocation (budget transfer) between SP8 and the hospitals is underway.

The ongoing administrative, data protection, patient safety, ethical and legal issues have been done in parallel with the technical advances and have not slowed down the five hospitals deployment plan.

1.4 Impact of work done to date

The improved architecture based on building blocks allows reduction of dependency between elements of the MIP stack. It also simplifies the inclusion of additional analysis methods. The improved architecture also allows also for more robust and automated deployment activities in various environments. The development and deployment activities between Local and Federated parts are uncoupled.

We generated a sound infrastructure, already reviewed by leaders of other European projects, (also Canadian and American projects) who after the proof-of-concept are confident of its value and prepared to adopt it. A contract agreement has been transmitted to the PCO for validation. From this demonstration we expect rapid contribution of disease related data from other EU projects to the MIP.

The current deployment made possible by MIP-Local allows closer contact to real-world patient data and thus boosts the important mission of collecting data from various sources to build unique biomarkers (as demonstrated in the research papers published).

Current Status of the hospitals' deployment is summarised below:

SP8 specification for hardware and software are available under:

- <https://hbpmmedical.github.io/deployment/hardware/>.
- <https://hbpmmedical.github.io/deployment/software/>.

CHUV-CLM:



- Hardware: Dedicated hardware was provided by the hospital to SP8 specifications.
- Data: The clinical data requirements were discussed and agreed upon with the clinicians. Representative clinical data has been extracted and is available for pre-processing.
- Installation: The MIP-Local software was delivered and installed (including: Data Factory, Data bases, Mapping tools, Web portal).
- Pre-processing of neuro-imaging data and extraction of brain features is underway.
- Contacts: Daniel DAMIAN (Project coordinator and IT lead), Prof. Jean-Francois DEMONET (Clinical lead).

LILLE:

- Hardware: Dedicated hardware was provided by the hospital to SP8 specifications.
- Data: The clinical data requirements were discussed and agreed upon with the clinicians.
- Remote Access: Hospital IT staff provided secure remote access and passwords to the SP8 team and installation of the Virtual Machine containing the MIP software was achieved.
- Pre-processing will start once the hospital data is available.
- MOU was signed 29 March 2016.
- Contacts: Melanie LEROY (Project coordinator), Pascal VIVIER (IT lead), Prof. Florence PASQUIER (Clinical lead).

MILAN:

- Hardware: Dedicated hardware was provided by the hospital to SP8 specifications.
- Data: The clinical data requirements were discussed and agreed upon with the clinicians.
- Remote Access: Hospital IT staff provided secure remote access and passwords to the SP8 team and installation of the Virtual Machine containing the MIP software was achieved.
- The specific agreement was signed.
- Pre-processing will start once the hospital data is available.
- Contacts: Claudia BESAGNA (Project coordinator), Giani Origgi (IT), Prof. Gabriela BOTTINI (Clinical lead)

FREIBURG:

- Hardware: Dedicated hardware was provided by the hospital to SP8 specifications.
- Data: The clinical data requirements were discussed and agreed upon with the clinicians.
- Remote Access: Hospital IT staff provided secure remote access and passwords to the SP8 team and installation of the Virtual Machine containing the MIP software was achieved.
- MOU signed 31 July 2015.



- Contacts: Dr. Christian HAVERKAMP (Project coordinator), Volkmar Glauche (IT), Dr. Dr. Karl EGGER (Clinical lead).

TLVMC: (Now due in M18 after the change of the hospital's general manager).

- Hardware: Dedicated hardware designated By TAU (servers).
- Data: The specification of the clinical data was discussed and agreed upon with the clinicians.
- MIP: Virtual Machine specification in advanced discussions with the hospital staff.
- MOU signed 23 March 2016.
- Contacts: Dr. Ahuva MEILIK (Head of operation research and quality manager department), Dr. MD. Alexis MITELPUNKT (Neurology department).

1.5 Priorities for the remainder of the phase

We believe that our strategy to implement the MIP-Local while continuing the infrastructure for the MIP-Federated and our commitment to the reviewer's requests provided in Document-1 are valid steps towards delivering federation between several hospitals before the end of the year (2017). After achieving deployment of MIP-Local to at least three hospitals, the next priority is to train local users and to make development closer to clinical needs.

The priority for the deployment is to pre-process the hospital's clinical data and make it available to the local clinicians in the MIP-Local Web Analytics module. The next priority is to make the summaries of this clinical data from several hospitals available on the MIP-Federated level. The research priority is to integrate clinical hospital with additional research datasets listed in the Open Research Data chapter. The combined datasets will be used to improve and cross-validate disease models.

After empirically validating already created biological signatures of disease across hospitals, we will organise a broad campaign across Europe's premiere clinical neuroscience centres centred around demonstrating the added diagnostic and prognostic value of MIP-derived disease signatures (one large event is already planned).

We will pursue our strategy to include more contributions from external communities. We have already planned several workshops are planning the next HBP school on future medicine, on the topic of disease-related-neuroscience including modelling, in collaboration with all other SPs.



2. WP8.1 - Federated Clinical Data Infrastructure (FCDI)

2.1 Key Personnel

Work Package Leader: Anastasia AILAMAKI (EPFL)

2.2 WP Leader's Overview

During the last reporting period, there are ongoing modifications to the development work plan (please see details below). Nevertheless, the WP8.1 team has responded by dynamically adapting to the modifications. Moreover, every Partner made significant progress in their research agenda, resulting in exciting prototypes and very interesting paper submissions (please find more details in the Component reporting). In addition, the collaboration among the WP8.1 partners was smooth and productive. The communication via email or other electronic methods was always effective and all questions or comments were addressed.

In October 2016, an additional review of SP8 took place. As a result of this review and its recommendations, the SP8 leadership made decisions which had a significant impact on our planned work. This process was time-consuming and caused delays.

In particular, SP8 leadership, based on review recommendations, has completely deprioritised WP8.1 research, in order to use WP8.1 resources for Platform development and deployment tasks that are not the responsibility of WP8.1. While we have accommodated these needs to the largest extent possible, this did mean fewer resources devoted to research and to expanding the functionality of the platform.

For DIAS, the situation was even more problematic, for the following reasons:

After a redefining process, following the comments of the EC reviewers at the end of 2016 the SP8 leadership decided to shift the focus of SP8 and concentrate SP8 efforts around the deployment of the Medical Informatics Platform at the Hospitals. In October 2016 the SP8 leader changed the role of the EPFL DIAS Lab from being responsible for the Hospital Bundle deployment to being responsible for the deployment of the entire MIP at the hospitals. As a result, significant additional effort is required from DIAS, and consequently, task descriptions have to be modified. The SP8 leadership is still modifying the priorities concerning DIAS' tasks (the last modification was announced in the project management tool 28/2/2017). Nevertheless, since there is no official amendment to the DoA yet (planned to be submitted to the EU in June 2017), the task descriptions for EPFL DIAS in this report are the ones defined in the DoA which was agreed and signed in 2015.

Despite the continuous modifications on WP8.1 Tasks, the WP8.1 work for the M12 Deliverable in the Work Package has progressed in a satisfactory way. The work defined in WP8.1 tasks has been completed up to the point where it is impeded by the following external factors: (a) delay in hospital funding for acquisition of servers, (b) delay in signature of agreements between hospitals and SP8 leadership and (c) the HBP Legal Counsel, in collaboration with their CHUV counterparty, is re-processing terms in the non-disclosure agreements between technical teams.

WP8.1 has worked towards the improvement of their product in the context of the MIP. Moreover, our engineers have integrated this product within the MIP.

We have been in contact with the hospitals via meetings where we presented the HBP project and provided a live demo of our product (i.e., Prof. AILAMAKI's presentation to TAU Hospital on 21 September 2016 via Skype). As a result, the hospitals were very interested in participating in the project.

In addition, our team members (young engineers and postdocs) have had the opportunity to learn and practice the development of tools for medical research.



Finally, and importantly, WP8.1 generated multiple research results of importance to the MIP and to clinician users of the MIP, as outlined in the report.

2.3 Priorities for the remainder of the phase

In the remainder of SGA1, our WP will continue their work on the optimisations of the hospital bundle components, in order to deliver a state-of-the-art product to the medical scientific community as part of the MIP. We will also continue to join the effort to realise the deployment of the MIP on the Hospitals infrastructure as initially planned.

IMPORTANT NOTE: the DIAS team has reviewed the naming and definitions of their PLA components to better reflect the final orientation of work. In order to create a clear reporting for an external audience, only the updated components are presented here.

The most important modifications concern the change of concept for the Hospital Bundle, the renaming and improvement of description of several components, and the decision to support the Brain Imaging Data Structure format (BIDS) instead of genetic data (as requested by CHUV - SP8 Leader and Partner). Detailed information regarding the modifications and renaming is available upon request.

The main modifications are reflected with some additional comments within the report. WP8.6 and the WP8.1 coordinator will convene a meeting to finalise the naming of the Components.

2.4 Milestones

Table 1: Milestones for WP8.1 - Federated Clinical Data Infrastructure (FCDI)

MS No.	Milestone Name	Leader	Task(s) involved	Expected Month	Achieved Month	Comments
MS6	MS8.1.1 Implementation Plan of the Hospital Bundle	P1 - EPFL	T8.1.8	M02	M02	In SGA1 there were five Milestones defined for WP8.1 which also represent the Deliverables of each period. For M02 (MS6), our team was asked to prepare the Implementation Plan of the Hospital Bundle. The SP Leader has specified that it would be enough to define the plan as Milestones. This was done and documented in the signed SGA1 document. Furthermore, EPFL DIAS has prepared a document called MIP deployment which is associated with this first Milestone. The document is a living document and will be continuously updated until the end of SGA1 (Dimitra TSAOUSSIS-MELISSARGOS 16 September 2016).
MS38	MS8.1.2 Service integration for online interaction through web portal together with WP8.5	P1 - EPFL	T8.1.4	M06	n/a	The Hospital bundle tools are completed, tested and packaged and available for installation at the hospital side. The installation has already been done at CHUV in Lausanne, Switzerland. This marks the completion of MS38 (M06) as defined in the DoA.
MS111	MS8.1.3 Version 1 of Hospital Database Bundle released and integrated	P1 - EPFL	T8.1.3	M12	n/a	The Hospital Bundle components have been integrated and tested to work together. Moreover, the packaged software has been integrated with the WEB portal of the MIP. The Milestone is therefore achieved.

2.5 T8.1.1 - Infrastructure to Support Just-in-Time Analytics on Raw Medical Data

2.5.1 *Key Personnel*

Task Leader: Anastasia AILAMAKI (EPFL)

2.5.2 *SGA1 DoA Goals*

- 1) Develop advanced code-generation techniques that can be integrated into the query engine to allow native support of imaging and genetic data.
- 2) Create a function library with common operations for imaging data.
- 3) Determine the applicability of randomised algorithms and synopses for optimising time-consuming queries.

2.5.3 *Component Progress*

2.5.3.1 DATA > Test data > BIDS test data files

Description: Obtain Brain Imaging Data Structure examples for development and testing.

Progress: Completed. The specification of the data format provides several examples.

Links: <http://bids.neuroimaging.io/>

2.5.3.2 SOFTWARE > HDB > Query Engine > Query plug-in for BIDS Data

Description: The goal of this Component is to develop a plug-in to enable the local query engine to perform queries directly on Brain Imaging Data Structures (BIDS) folders.

Progress: Design has started. Due to the focus shift of the project (see above), the report on this Component will appear in future progress reports.

Note: This Component replaces "SOFTWARE > Query Engine > Query plug-in for Genetic Data".

2.5.3.3 SOFTWARE > HDB > Query Engine > BIDS Data Library in Query Engine

Description: The goal of this Component is to develop a library of functions for common operations on Brain Imaging Data Structures in the local query engine.

Progress: Due to the focus shift of the project (see above), a more detailed report on this Component will appear in future progress reports.

Note: This Component replaces "SOFTWARE > Query Engine > Genetic Data Library in Query Engine".

2.5.3.4 DATA > Test Data > Nifti test data files

Description: Obtain example medical files for development and testing.

Progress: Completed, example files based on real data were obtained by SP8 partner.

2.5.3.5 SOFTWARE > HDB > Query Engine > Query plug-in for Medical Imaging Data

Description: The goal of this Component is to develop a plug-in to enable local query engine to perform queries directly on Medical Imaging files (e.g., MRI and Nifti).

Progress: Work on the conceptualisation and an initial implementation has started.

Due to the focus shift of the project (see above), a more detailed report on this Component will appear in future progress reports.



2.5.3.6 SOFTWARE > HDB > Query Engine > Nifti Library in Query Engine

Description: The goal of this Component is to develop a library of functions for common operations on imaging data/Nifti files in the local query engine.

Progress: This Component has been delayed because of an upstream component.

Quality Control:

- SOFTWARE > HDB > Query Engine > Query plug-in for Medical Imaging Data - T8.1.1 - no input received from this component

2.5.3.7 SOFTWARE > HDB > Query Engine > Extended Array Query Support

Description: This Component provides basic primitives for queries and computation over multidimensional arrays in the query engine. This will enable queries such as selecting a sub-range in arrays in n dimensions, transpose arrays or multiply matrices.

Progress: Conceptualisation has started. Traditional databases work on bags (unordered sets) and not on lists. This Task requires to add the notion of lists in the query engine. The first implementation will be based on a special function which adds indexes to bags.

Using this strategy, we will be able to perform common array operations such as crop, transpose, matrix multiplication in nifti files for instance. Changes in the optimiser will be necessary to accommodate the use of ordered list (i.e. arrays).

Due to the focus shift of the project (see above), a more detailed report on this Component will appear in future progress reports.

2.5.3.8 MIP - SOFTWARE > HDB > Query Engine

Description: This software component is a database management system which is used as part of the MIP. The engine serves as a database backend at each participating hospital and is responsible for executing queries on raw hospital data.

Progress: This Component has been provided at the end of RUP.

Links:

- <https://github.com/HBPMedical/PostgresRAW>
- <https://github.com/HBPMedical/PostgresRAW-UI>

2.6 T8.1.2 - Infrastructure for Distributing Query Engine Locally and over Hospital Hubs

2.6.1 Key Personnel

Task Leader: Anastasia AILAMAKI (EPFL)

2.6.2 SGA1 DoA Goals

- 1) Develop and integrate the FCDI (Hospital Bundle) software support for local data hubs.
- 2) Identify cache design options and invalidation algorithms that require low maintenance, while guaranteeing data consistency and accuracy.

Note: As mentioned in the introduction, the Hospital Bundle concept has been modified. It is no longer a packaged software but different components which are integrated in the MIP separately.

2.6.3 Component Progress

2.6.3.1 SOFTWARE > HDB > Query Engine > Distributed Query Engine Over HPC

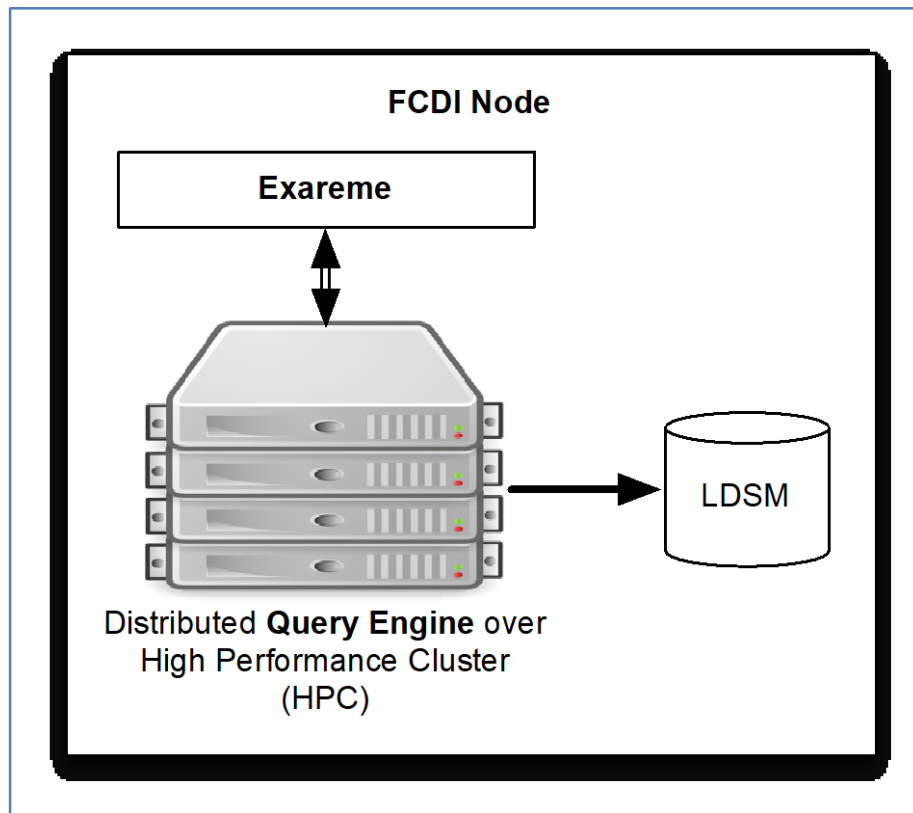


Figure 2: FCDI Node

Description: This Component will extend the local query engine to enable the use of distributed computing frameworks (such as Spark).

Progress: A first Proof of Concept has been implemented. It exploits the capacity of the Query Engine to connect to external relational databases. The Query Engine connects to SparkSQL, which is an existing implementation of a distributed database engine over Spark.

2.6.3.2 SERVICE > Hospital Hub > Secure Connection between Members of Hospital Hubs

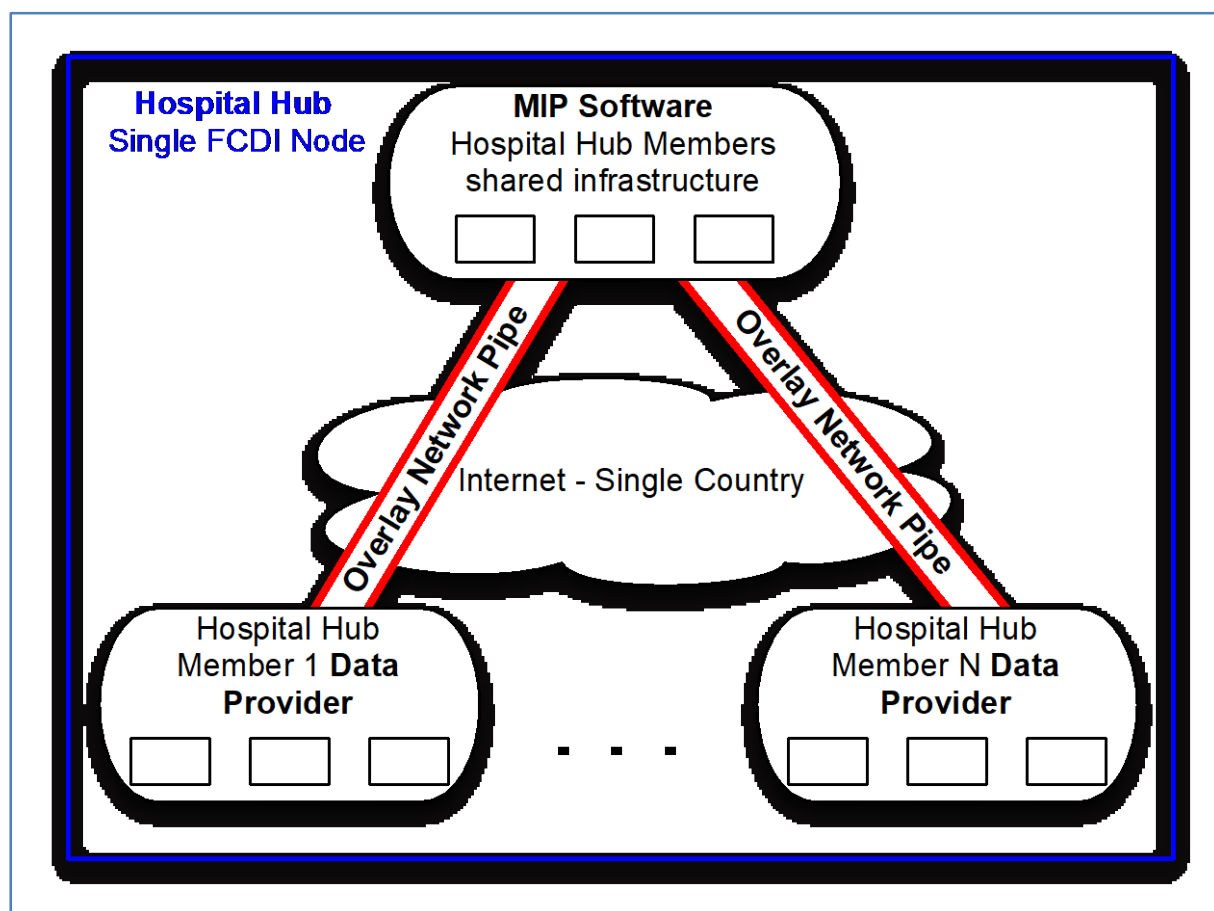


Figure 3: Hospital Hub

Description: This Component will ensure secure connections between two members of a hospital hub.

This Component depends directly on the MIP therefore it should be renamed “SOFTWARE > MIP > Secure Connection between Members of Hospital Hubs”.

Progress: The Docker “Overlay Network” technology has been selected; a Proof of Concept is being implemented. The results will be available at:

<https://github.com/HBPMedical/RAW-deploy>

Work on the reference document will start once the first release of MIP Local will be available, in order to take into account the complete architecture of the platform, in parallel to the setup of the MIP Federated.

2.6.3.3 SERVICE > Hospital Hub

Description: The goal of this Component is to extend the MIP architecture to support multiple data providers in a single FCDI node.

Progress: Design has started. Each data provider node will upload data to a central node, which will be a full MIP-Federated node. Extensions will be required to keep track of the origin data provider.

2.7 T8.1.3 - Installation and Adaptation of Hospital Bundle at Local Hospitals with Upgrades and Support



2.7.1 *Key Personnel*

Task Leader: Anastasia AILAMAKI (EPFL)

2.7.2 *SGA1 DoA Goals*

- 1) To prepare and install the first working version of the FCDI (Hospital Bundle) at the participating hospitals in a bottom-up fashion.
- 2) To integrate new features into the FCDI as they become available from other Tasks in WP8.1.

Note: As the Hospital Bundle does not exist anymore, these goals need to be updated. The current focus of the work can be summarised as:

- 1) To contribute to the creation of the MIP by integrating WP8.1 software in the new architecture, supporting anonymisation and improving network security.
- 2) To deploy the MIP, once it is provided by SP8 partners

2.7.3 *Component Progress*

2.7.3.1 SERVICE > Data Capture > Data Access > Deployment

Description: Enable connection to local data (scripts, adaptation to node-specific data and storage structure).

Progress: At CHUV hospital, the online connection to the information system is underway, several issues are being worked on:

- 1) Access rights / server configuration to allow the anonymiser to retrieve the information
- 2) Adaptation of the connection system, as instead of only having access to the agreed data, the systems allows us to see everything. Thus, an input filter has been finalised to make sure we do not try to retrieve information we are not expected to.

We have not yet had access to the remaining hospital systems, preventing further progress.

2.7.3.2 SERVICE > Data Capture > Data Anonymisation > Integration and deployment

Description: On-site installation at individual FCDI nodes, to enable automatic production of anonymised data. This Component includes adaptations to the local information systems and provides local hospital anonymisation validation tests.

Progress: The software has been provided, local adaptation for each hospital is underway as access is granted and the required authorisations delivered.

The Data Governance and Data Selection (DGDS) Committee has taken over the task of providing the list of variables which should be retrieved from each hospital, as well as how they should be depersonalised. The list has still not yet been communicated and is critical for the proper configuration of the data depersonalisation.

2.7.3.3 SERVICE > Deployment > MIP Local

Description: On-site installation and configuration MIP local at FCDI nodes (i.e. participating hospitals). First deployment step, before federation.

Progress: CHUV has delivered the MIP Local installation script only partly, with no installation documentation, nor complete configuration example. EPFL DIAS is in the process of collecting the different configuration resources and merging them for a full production deployment of the MIP Local on a single server.

Links: <https://github.com/HBPMedical/mip-microservices-infrastructure>

2.7.3.4 SERVICE > Deployment > MIP Federated

Description: On-site installation and configuration of MIP global at FCDI nodes (i.e. participating hospitals). Second deployment step: federation.

Progress: The integration of Exareme within the MIP is under way. Support has been provided for development and debugging of the packaged version of Exareme in a simplified setup.

Links:

- <https://github.com/HBPMedical/RAW-deploy>
- <https://github.com/HBPMedical/ExaremeLocal-docker>

2.7.3.5 SOFTWARE > Remote Starting of Services

Description: Selection, adaptation and integration of remote management tools for FCDI.

Progress: The technologies have been selected. First PoC is being designed.

Links:

- <https://github.com/HBPMedical/RAW-deploy>

2.7.3.6 SOFTWARE > Encrypted Overlay Network

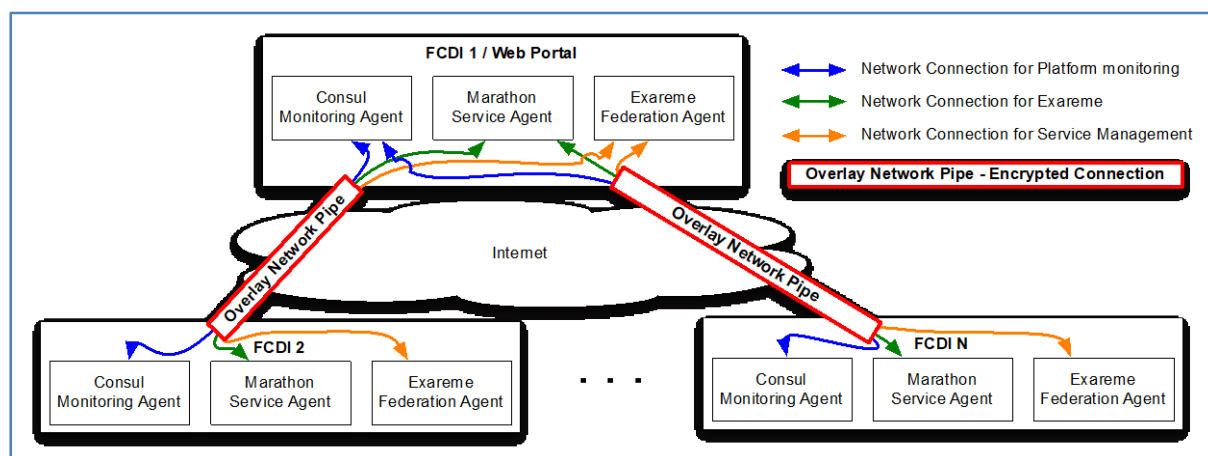


Figure 4: Overlay Network

Description: Selection, adaptation and integration of an encryption solution for communications between FCDI nodes.

Progress: The Docker “Overlay Network” technology has been selected in collaboration with the UoA team, to ensure their requirements were met. Thus, a single technology allows to secure and crypt communications at different levels in MIP and provides isolation between services (cf. Hospital Hubs).

The first PoC is being implemented.

Links:

- <https://github.com/HBPMedical/ExaremeLocal-docker>

2.7.3.7 SERVICE > HDB > Query Engine > Integration

Description: Integration of query engine with other components of MIP. (Selection and adaptation of the Query Engine to be used in MIP was done as part of RUP).

Progress: DIAS has done the packaging under the new MIP requirements, as well as prepared the extended documentation about the software compilation, installation as well as usage as packaged in docker images.

Both a REST API and standard JDBC connections are available.

Configuration of the data location and layout is still a work in progress in the context of the MIP.

Links:

- <https://github.com/HBPMedical/PostgresRAW-docker>
- <https://github.com/HBPMedical/PostgresRAW-UI-docker>
- <https://github.com/HBPMedical/jdbc-driver-raw>

2.7.3.8 SERVICE > HDB > Query Engine > Upgrades

Description: Continued development of query engine in evolving context of MIP. (Selection and adaptation of the Query Engine to be used in MIP was done as part of RUP).

Progress: Several revisions providing enhancements as well as bug fixes have been provided, packaged, and are available in the context of the MIP.

Links:

- <https://github.com/HBPMedical/PostgresRAW>
- <https://github.com/HBPMedical/PostgresRAW-UI>

2.7.3.9 SOFTWARE > Data Factory > Data Anonymizer

Description: Gnubila FedEHR software.

Progress: The software has been provided at the end of RUP.

2.8 T8.1.4 - Data Integration

2.8.1 *Key Personnel*

Task Leader: Vasilis VASSALOS (AUEB)

2.8.2 *SGA1 DoA Goals*

T8.1.4 focuses on the following challenges:

- 1) Advanced data integration capabilities for the MIP. The outcome will be a mediator engine that executes complex database-like queries over the aggregate hospital information, with a focus on handling more complex access control schemes, schema constraints and mappings, tree-structured and semi-structured data, as well as ontologies. This engine will allow queries to be processed while respecting rules about who has access, and showing where and how information maps across hospitals, in a way that also takes advantage of the additional information in the multiple schemas, such as keys and foreign keys, parent-child relationships, and ontologies (providing, for example, ISA, PART OF, and equivalence relationship information, as well as taxonomies).
- 2) Enabling online data integration by providing novel techniques for the continuous, incremental integration of incoming hospital data. This will enable the move from manual to automatic data cleaning, and the transformation and merging of actions whenever hospitals add new data to the MIP.
- 3) Enabling community schema curation by allowing community members to create, validate and reuse schema mappings. This will increase the scope of the MIP significantly, and allow it to deal more easily with the variety of clinical data available.



2.8.3 *Component Progress*

2.8.3.1 SOFTWARE > HDB > Ontology Based Data Access

Description: Ontologies play a key role in semantically defining a domain of interest. Their use in the medical domain has been extensive since they provide a standard terminology with well-defined semantics and relations among its components that allows interoperability. Bridging ontologies and data is of paramount importance for MIP. Given a query, provide answers that reflect both the data and the knowledge captured by the ontology. This Component will produce such a system that will reformulate posed queries to capture the knowledge of HBP and other ontologies while also providing access to data stored on the LDSMs.

Progress: To implement the ontology-based data access module a number of query rewriting systems have been tested and reviewed. AUEB-RC focused on the systems Rapid and IQAROS two systems supporting ontology-based data access over data annotated using OWL 2 QL ontologies because, first, members of AUEB-RC team have been involved in their design and implementation and, second, because these systems are two of the most efficient query rewriting systems. Ontology-based data access is comprised of two steps. First the user query is rewritten into a union of conjunctive queries (UCQ) taking into account the axioms of the ontology and the second is further rewriting the UCQ given the mappings of the database to the ontology entities. Both systems could handle the first step but neither could handle the second. AUEB-RC extended IQAROS with a mapping rewriting step in order to be able to access the data stored in a database. Further to that AUEB-RC has developed a number of optimisations for improving the efficiency of the query answering/access step. Many of these optimisations were scientifically novel and hence these research results on were published in the following paper: Venetis T, Stoilos G, Vassalos V. Rewriting Minimisations for Efficient Ontology-Based Query Answering. ICTAI 2016: 1095-1102.

Moreover, an extended version of the above paper was invited at the International Journal on Artificial Intelligence Tools and is currently under review.

To test the system and approach in the Human Brain Project AUEb-RC is in the process of the following activities:

- Reviewed various medical ontologies from the BioPortal (<http://bioportal.bioontology.org/>) in order to use them to annotate and map the project data to ontologies. From this process AUEB-RC identified several interesting ontologies, but so far AUEB-RC has selected SNOMED CT.
- Started mapping the ADNI and other available hospital data of the project to the SNOMED CT ontology.
- Started reviewing the RestAPI of RAW and Exareme in order to integrate our IQAROS system with them and provide the ontology-based data access service.

The M12 planned release has been achieved.

Links: Venetis T, Stoilos G, VassalosV. Rewriting Minimisations for Efficient Ontology-Based Query Answering. ICTAI 2016: 1095-1102.

2.8.3.2 SOFTWARE > HDB > Access Right Module

Description: This is a module that enables the local database and the mediator engine to execute complex database-like queries over the hospital data while respecting complex access control schemes and schema constraints and mappings. This module will allow queries coming from the Web Portal to the LDSMs to be processed while respecting rules about who has access, and showing where and how information maps across hospitals, in a way that also takes advantage of the additional information in the multiple schemata, such as keys and foreign keys and parent-child relationships. Hence this Component will take as input the access rights of the users that perform various tasks on the Platform, in an appropriate



format, and various known schema constraints of MIP data. This Component will affect the way the Local Database and the Federation Engine work, meaning that it will only allow users to query them according to their access rights.

Progress: AUEB-RC has designed a centralised access control system that meets the Platform requirements concerning availability of its resources (that is the hospital data stored in the local database for MIP and the MIP services). For this purpose, AUEB-RC has reviewed state-of-the-art access control systems and proposed a model that can satisfy MIP policy requirements. The model that was adopted is a role-based access control model (RBAC) as it has been recognised as an efficient access control mechanism that has been used by web applications and healthcare systems. Moreover, AUEB-RC has decided on the use of semantic technologies to define the access control system. Ontologies can be used in the context of an access control mechanism to express the structure of the organisation, the role of each user, and the relationship between people working in the organisation. They can be extended to capture future requirements of the system and they allow the use of existing domain specific vocabularies. For example, in healthcare there are several vocabularies already used to describe medical data, such as the JuBrain Atlas, the Allen Atlas that are used for brain modelling.

The proposed access control model assigns permissions to MIP users according to their role. Roles have been defined by considering the MIP user groups (Clinician, Researcher, Statistician, Scientific Developer, Platform Developer, Medical Research Writers, General Public). Permissions are assigned to roles and are related to the actions that each user group can perform that is, access available data, store and update models. According to the model, a user can be assigned to many roles and a single permission can be assigned to more than one roles.

Moreover, AUEB-RC has used a basic ontology to describe the elements of the proposed access control model. The basic ontology includes the concept Action that can be either a PermittedAction or a ProhibitedAction. An Action is related to the concept Subject (that describes the user performing the action) and Object (that describes the resource). Moreover, AUEB-RC has defined a domain specific ontology that describes the structure of the Platform. Its concepts are subclasses of the classes in the basic ontology. The user groups are represented in the ontology as subclasses of the Subject concept, and the data of MIP are subclasses of the Object class. This ontology can be extended with more concepts that describe MIP resources if necessary. By using the vocabulary of the ontology we can describe access control rules. By using the vocabulary of the ontology we can describe access control rules. For example, a rule can describe that only psychiatrists and neurologists can obtain data related to brain scans, or that only psychiatrists can get examination values of psychiatric exams.

Since the core of the access control model of the database has been designed, the set access rules can be easily extended to capture the requirements of the MIP concerning the secure sharing of resources. Finally, the set of access rules will be integrated in the OBDA in order to filter the query answering according to the rights of each user.

The planned M12 Milestone has been achieved and can be summarised in a document that can be found using the link in the next section.

Links: <https://github.com/aueb-wim/AccessRightModule>

2.8.3.3 SOFTWARE > HDB > Online Data Integration Module

Description: This Component will enable the move from manual to automatic cleaning, and the transformation and merging of actions whenever hospitals add new data to the Medical Informatics Platform. More precisely this Component will extend MIPMap, developed during the RUP of HBP, to support incremental Data Exchange. This means that instead of re-integrating data to the hospital's LDSM, whenever new data are exported from participating hospitals (following the standard pipeline of anonymisation, etc.), they will be integrated into the already existing data taking into account the information that has been integrated



before. Hence this Component will vastly affect the way Data is integrated to the Platform (Data Integration & Schema Mapping/Data Exchange) and the way metadata will be enriched. The functionality provided is incremental integration of data from hospitals.

Progress: To implement the Online Data Integration module AUEB-RC has performed a state-of-the-art review in the area of Incremental View Maintenance, a scientific area closely related to the Online Data Integration component. Moreover, since MIPMap used for the translation of data is a Data Exchange tool, i.e. it implements an Export Transform Load (ETL) procedure using declarative rules (Tuple Generating Dependencies - TGDs), AUEB-RC has also performed an evaluation of how the aforementioned technologies can be utilised in this scope.

Moreover, AUEB-RC has put effort in the deployment of MIPMap in the MIP, the design of mapping tasks and data translation for hospital and research data, as well as has developed new MIPMap functionalities that were the outcome of specifications provided by CHUV, due to focus shift decided by SP8 leadership.

The planned M12 Milestone has been achieved and can be summarised in a document that can be found using the link in the next section.

Links: <https://github.com/aueb-wim/OnlineDataIntegration>

2.8.3.4 SOFTWARE > Web Exploration and Analytics > Community Schema Curation

Description: This Component will allow MIP users to create, share, validate and reuse schema mappings. More precisely, MIP users (provided they have specific access rights) will be able to share their mappings, making them global. This will allow all other MIP users to view these global mappings and endorse them partially or completely to their own. Additionally, users will be able to combine and extend existing mappings by adding/removing tables. Finally, users will be able to “friend” other users allowing them access to their (non-global) mappings. This Component will overall increase the scope of the MIP significantly, and allow it to deal more easily with the variety of clinical data available. This Component is based on extending WebMIPMap with crowd sourcing functionalities. This Component affects the ontology& standards component as it will make standardisation easier. Moreover, it affects the Information and Scientific References component as it will affect the ontologies and variables used and finally it will affect schema mapping and data integration as it will affect the way mappings (that could potentially run on MIPMap) are created.

Progress: This Component builds on the existing WebMIPMap component, developed in the RUP. For M12 AUEB-RC has completed the planned M12 Milestone that allows administrators (in general users with privileges) to create and store global schema mappings that are made available to all WebMIPMap users. Additionally, WebMIPMap has been enhanced with the ability of (regular) users to view/edit global schema mappings and additionally store them to their own workspace. More precisely, users can view global schema mappings, displayed using different colour, and modify them according to their needs. In the following users are able to store the modified global schema mapping in their workspace and load them later on as their own. Another functionality that has been added to WebMIPMap is that of allowing user-defined mappings (not administrator global ones) to be viewed by other users. To achieve this a preliminary version of the “users I trust” feature has been implemented. This feature borrows from social networks and provides WebMIPMap users the ability to define a list of users that they trust and share with them schema mappings, that are called public. Trusted schema mappings can be again edited and stored in the users’ workspace.

The development of this Component is well ahead of schedule and link to the provided code can be found in the next section.

Links: <https://github.com/HBPMedical/WebMIPMap/tree/SGA1>



2.8.3.5 SOFTWARE > HDB > MIPMap refinement

Description: MIPMap is a data exchange tool developed by AUEB-RC during the Ramp Up phase, based on the open source ++Spicy¹¹ data exchange tool. MIPMap supports and automates the required ETL processes in order to translate the data provided by the hospitals to the MIP schema and thus populate each hospital's Local Data Store Mirror. MIPMap has been successfully delivered during the RUP, however new specifications on the Data Factory workflow, necessitate the refinement of MIPMap. Such specifications include exporting translated instances directly to RDBMS as well as supporting i2b2 schema.

Progress: This Component is progressing according to the specifications that are emerging following the Data Factory workflow. The development follows the needs that arise and currently all specifications are met. More precisely the functionality of exporting translated instances to RDBMSs has been implemented, tested and integrated to the tool, thus supporting export to any database that follows the i2b2 (or any other) schema. Moreover, new functionality that automates the process of harmonisation during the data translation is being developed and will be completed and integrated soon. This involves allowing MIPMap to accept as input except for the correspondence of the specific hospital and harmonised variables also the transformation function that converts one to the other. Up until now this was a process that would need to be performed manually, during the creation of the mapping task that would translate data originating from hospitals to the common MIP schema.

Links: <https://github.com/HBPMedical/MIPMap>

¹ <http://www.db.unibas.it/projects/spicy/>

2.9 T8.1.5 - Distributed Query Workflow Engine

2.9.1 Key Personnel

Task Leader: Yannis IOANNIDIS (UoA)

2.9.2 SGA1 DoA Goals

This Task will focus on the development of the distributed complex workflow engine developed at the University of Athens (UoA). The engine should support:

- a) Distributed execution of complex, resource, and time-consuming data processing flows,
- b) Streaming processing of data flows, and
- c) Iterative workflow computations on both the local and global level.

This Task will focus on the integration of all above functionalities of the federated engine with the Local Data Store Mirror and web interfaces. The Task will continuously evaluate and test all different parts and versions of the distributed complex workflow engine developed using data from the participant hospitals. The results of each evaluation phase will inform our developments in subsequent phases so that possible problems are corrected. This Task will contribute to the Platform, as well as to brain research by providing the module of the distributed complex dataflow engine.

2.9.3 Component Progress

2.9.3.1 Master component

Description: The Master Component will provide functionalities such as execution of distributed complex, resource and time-consuming data processing flows, streaming processing of data flows, and computation of iterative workflows on both the local and the global level. Furthermore, the Master Component will evaluate and test all different parts and versions of the distributed complex workflow engine developed using data from the participant hospitals.



Progress: We have begun working on an initial implementation of the input and output format of the distributed query engine. PFA format is now used: 1) as input parameter to be executed in pipeline fashion and 2) as output format to validate results. Furthermore, after evaluating implementation approaches of Iterative workflow computations on global level, we begun and completed the implementation of global iterations on the distributed complex workflow engine. Moreover, continuous integration has been set up to ensure the quality and functionality of Exareme's codebase. Last but not least, we finished the implementation of a docker container to package the master component with one worker. The planned M12 Milestone has been achieved and link to the provided code can be found in the next section.

Links:

- www.exareme.org
- <https://github.com/madgik/exareme/tree/mip>
- <https://arxiv.org/pdf/1501.01070.pdf>
- <https://hub.docker.com/r/hbpmip/exaremelocal/>
- <https://github.com/HBPMedical/ExaremeLocal-docker>
- <https://travis-ci.org/madgik/exareme>

2.9.3.2 Worker / Bridge Component

Description: The workers reside on the hospital nodes and act as a bridge with the RAW query engine which executes the queries in situ.

Progress: First, we have finished the integration with the Raw query engine. Second, we have created automated tests to ensure the integration and functionality of Exareme, the algorithms and the query engine (RAW). Furthermore, we evaluated optimisation approaches for global iterations on worker nodes. Last but not least, we finished the implementation of a docker container to package the Master Component with one worker and the Raw query engine. The planned M12 Milestone has been achieved and link to the provided code can be found in the next section.

Links:

- www.exareme.org
- <https://github.com/madgik/exareme/tree/mip>
- <https://github.com/HBPMedical/ExaremeLocal-docker>
- <https://hub.docker.com/r/hbpmip/exaremelocal/>

2.9.3.3 Web portal connector component

Description: This Component interfaces the Master Component with the web portal.

Progress: We have begun working in the integration with the web portal (both async and sync rest API are supported). Furthermore, we implemented global iterations for supporting the same interface to the web portal. The planned M12 Milestone has been achieved and link to the provided code can be found in the next section.

Links: <https://github.com/madgik/exareme/tree/mip>

2.10 T8.1.6 - SQL-Based Data Mining and Query Templates

2.10.1 Key Personnel

Task Leader: Yannis IOANNIDIS (UoA)



2.10.2 SGA1 DoA Goals

This Task will focus on:

- a) Developing complex user-defined functions (UDFs). UDFs are needed in SQL-based data mining workflows for adapting and supporting algorithms provided by SP8 data mining groups. In addition, UDFs can be used to call external libraries/systems such as NumKit, SciKit, R, external data streams, etc.
- b) Designing and implementing algorithm templates, which describe in a high level language, parameterised distributed workflows that can be instantiated and submitted by specifying particular values to their parameters. Templates contain SQL queries that use UDFs for data processing and describe different processing workflow graphs such as: a) local-global workflows, b) multistep local-global workflows, c) iterative local-global workflows.

This Task will be integrated with T8.1.5 on both the local hospital bundle and the Web UI (T8.5) levels.

2.10.3 Component Progress

2.10.3.1 Template composer component

Description: The template composer converts the template, which describes parameterised distributed workflows, into an ExaDFL query script. The template composer is responsible for the isolated execution of each algorithm template. The planned M12 Milestone has been achieved and link to the provided code can be found in the next section.

Progress: We finished the implementation of the template in order to model local-global workflows, multistep local-global workflows and iterative local-global workflows. Now we are able to support the global iterations of iterative algorithms composed of init, step, termination_condition & finalise phases. The planned M12 Milestone has been achieved and link to the provided code can be found in the next section.

Links:

- <https://github.com/madgik/exareme/tree/mip>
- <https://github.com/madgik/mip-algorithms/blob/master/README.md>

2.10.3.2 UDFs component

Description: UDFs Component focuses on the development of complex user-defined functions (UDFs) that are needed in SQL-based data mining workflows, adapting and supporting algorithms provided by SP8 data mining groups. UDFs that interface with external libraries/systems such as NumKit, SciKit, R will also be implemented.

Progress: We have implemented and tested a set of UDFs to support four use cases, three of which contain simple statistics and the fourth use case contain a complex data mining algorithm (privacy preserving distributed linear regression).

Furthermore, we started working on an initial implementation of UDFs integrated with SciKit-learn:

- a) UDFs for fitting data to predictive models
- b) UDFs for classifying new samples

We also started work on incorporating row UDFs written in R (in parallel with python). Last but not least, we started supporting aggregate UDFs written in R (in parallel with python). The planned M12 Milestone has been achieved and link to the provided code can be found in the next section.

Link:

<https://github.com/madgik/exareme/tree/mip/exareme-tools/madis/src/functionslocal>

2.11 T8.1.7 - Management of Query Templates and Workflows

2.11.1 Key Personnel

Task Leader: Yannis IOANNIDIS (UoA)

2.11.2 SGA1 DoA Goals

This Task will focus on the management of the query templates and workflow and it should support the following:

- 1) A repository for the query templates that will provide storage, reviewing, access control and audit trail/logging capabilities related to the template management. Authentication and authorisation will be achieved through the use of user credentials. A version control system will be used to host the repository, storing the query templates as well as the history of template modifications. A code review system will be used to review all changes of the repository. A bug tracking/ticketing system will be used to collect all users' submitted bugs and features.
- 2) The ability for users to review, register, unregister and update any algorithm. The users will be able to monitor statistics regarding their algorithm overall execution. Each user will have access rights. Credentials with limited access will be used for the web portal.

2.11.3 Component Progress

2.11.3.1 Query template repository

Description: The query template repository Component will provide storage, reviewing, access control (authentication and authorisation) and audit trail/logging capabilities. The repository will be hosted in a version control system (VCS).

Progress: Proof of concept has been implemented for the query template repository. Furthermore, the docker container described on Master Component is connected with the query template repository component. The planned M12 Milestone has been achieved and link to the provided code can be found in the next section.

Link: <https://github.com/madgik/mip-algorithms>

2.11.3.2 Management component of query template repository

Description: The management Component of query template repository will manage user access to the query template repository. Each user will be tagged by a role in order to be able to have the corresponding access rights. The users will also have the ability to review, register, unregister, update any algorithm and monitor statistics regarding their algorithm overall execution.

Progress: Progress on Component: First, we have created specifications. Second the repository supports user access rights, review process and versioning. Last, proof of concept for this repository has been implemented. The planned M12 Milestone has been achieved.

Link: <https://github.com/madgik/mip-algorithms>

2.12 T8.1.8 - Hospital Bundle Integration Strategy and Business Model

2.12.1 Key Personnel

Task Leader: Anastasia AILAMAKI (EPFL)

2.12.2 SGA1 DoA Goals

The DIAS team is in discussion with the SP8 leadership in order to confirm that the DoA goals are still relevant for SGA1. T8.6.2 was responsible for the creation of the Business Plan of



the MIP and an upstream dependency for T8.1.8. Component Progress. This is not valid anymore and needs to be reflected in the DoA.

2.12.3 Component Progress

2.12.3.1 SERVICE > MIP> MIP Deployment Coordination> Weekly Progress report

Description: This Component will include regular reporting (either verbally within the scope of the SP8 weekly meeting or written upon request) on the MIP deployment as per the agreed Milestones and Deliverables of SP8.

Progress: This Component is repetitive and work-in-progress throughout SGA1 and will remain so until the end of SGA1. Until M12, there have been presentations by the members of the DIAS team of the progress of their tasks. Sample of their presentations can be found in the SP8 Collaboratory attached to the weekly SP8 meeting minutes and are available upon request.

2.12.3.2 SERVICE > MIP> MIP Deployment Coordination

Description: This Component takes care of the coordination within WP8.1 with regard to the MIP deployment at the chosen hospitals. It includes weekly meetings between the stakeholders (CHUV, UoA, AUEB, DIAS-EPFL) and follow-up of assigned actions.

Progress: The main progress of this Component relies on the fact that there have been several discussions and meetings on the subject in order to determine the best way forward. The work is on-going and the meeting minutes are available upon request.

2.12.3.3 Hospital Bundle Industrialization Plan (to be confirmed and amended accordingly)

Description: This Component will create a draft market research document for the hospital bundle.

Progress: Until M12 DIAS has defined a draft plan template. There are ongoing discussions between the SP8 leadership and the owner of this Task which will determine the final form of the document to be delivered by the end of SGA1. The strategy discussed until now consists of defining the parts of the Hospital bundle (i.e., each WP8.1 partner has to provide the necessary information) and document the outcome. As the owner of the MIP, CHUV will lead the task. Email exchanges on the matter are available upon request.

2.12.3.4 SOFTWARE > HDB > Hospital Bundle Integration Strategy

Description: This Component will define the strategy for the Hospital Bundle integration.

Progress: Due to the lack of history on the PLA, we cannot see who has entered this Component in our task. We cannot report on this Component since we are not the ones who have entered the information.



3. WP8.2 - Data Selection and Community Engagement

3.1 Key Personnel

Work Package Leader: Bogdan DRAGANSKI (CHUV)

3.2 WP Leader's Overview

According to the agreed "Fast-Track" plan, we established a Data Governance and Data Selection (DGDS) Committee that coordinates data selection across the recruited hospitals. The DGDS produced a "minimal" data set of clinical variables common to all hospitals as baseline for further technological and scientific development. The DGDS established unified data governance that resolved issues related to ethics, legal framework and security requirements.

Our efforts to augment Community Engagement should be better aligned with the HBP activities on the topic. The emphasis to engage basic and clinical neuroscientists has to include also patients' organisations representatives and health care decision makers.

Data selection was one of the critical elements of the "Fast-Track" plan that allowed us to bring the MIP project to the next level. This step secured the optimal functioning of the Data Factory that is now feeding a number of up- and downstream components linked to all major technological and scientific developments on the MIP.

3.3 Priorities for the remainder of the phase

Given the success of the agreed "Fast-Track" strategy that required a proactive data selection from hospitals, our priority for the remainder of SGA1 is to expand the "minimal" data set to a number of clinically relevant fields. This should follow the needs of MIP users that will give their feedback on the current set of available data.

Additionally, we will focus on community engagement through building channels for user-feedback on MIP features and professional communication to patients' organisations, health policy makers and the public. After empirically validating the already created biological signatures of disease across hospitals, we will organise a broad campaign across Europe's premiere clinical neuroscience centres centred on demonstration of the added diagnostic and prognostic value of the disease signatures of the MIP. This will be led by the CHUV's SP8 team of specialists in PR and community engagement.

3.4 Milestones

Table 2: Milestones for WP8.2 - Data Selection and Community Engagement

MS No.	Milestone Name	Leader	Task(s) involved	Expected Month	Achieved Month	Comments
MS34	MS 8.2.1 Report on Data Governance Risks and Issues, Innovation Plan and Community Engagement	CHUV	T8.2.1	M06	M06	WP8.2 developed a unified data governance framework, including data governance policies, rules, standards, best practices, processes and technologies to ensure secure delivery of trusted data Monitor the data flow from data selection to data curation and data analyses. WP8.2 formed a Medical Data Governance Committee, with stakeholder representatives, including data providers, local Institutional Review Boards (IRBs) and patient groups

3.5 T8.2.1 - Data Selection and Governance

3.5.1 Key Personnel

Task Leader: Bogdan DRAGANSKI (CHUV)

Other Researcher: Alexis MITELPUNKT (TAU)

3.5.2 SGA1 DoA Goals

Description: T8.2.1 will recruit hospitals to acquire access to data sources that the MIP will make available to users. This will facilitate access to cognitive, imaging, biological, genetic, and molecular datasets from large-scale European studies and initiative after establishing agreements or MoUs with hospitals and biobanks. The research and development will be conducted in collaboration with users from participating centres. The Task will handle all ethical and legal issues arising from the building the Clinical infrastructure and the Platform.

3.5.3 Component Progress

3.5.3.1 DATA > MDR > Common Variables & Metadata

Description: The Data Governance and Data Selection (DGDS) Committee selected brain relevant variables and metadata common to the participating hospitals to provide a "minimum" data set including demographic, behavioural and clinical features. The common variables and metadata will be used by the MIP for statistical analysis together with brain anatomy features.

Progress:

- Definition of variables and metadata from each of the participating hospitals with description of a "minimal" data set (available at <https://drive.google.com/drive/folders/0B5K3IDNQ5PbrbFdnamE0UWoycE0>).

Medical Informatics Platform's CDEs			Revision	0.2	Date
CDE	Gender	Group	Demographics		
Description		Type	Values	Format	
Gender of the patient - Sex assigned at birth		Binominal	Male; Female	"M"; "F"	
Data Set	Source	Column name	Type	Values	Format
ADNI	PTDEMOG.csv	"PTGENDER"	Binominal	Male; Female	1; 2
EDSD	EDSD-multicenter.csv	"Gender"	Binominal	Male; Female	"M"; "F"
PPMI	Screening_Demographics.csv	"Gender"	Polynomial	Female of child bearing potential Female of non-child bearing potential Male	0; 1; 2
Niigarda	Demo-Anonym.csv	"SEX"	Binominal	Male; Female	"M"; "F"
Lille	TBD	"P_Sex"	Binominal	Male; Female	"H"; "F"
CLM	Demographics_2016_08_03.xls	"SEX"	Binominal	Male; Female	"Homme"; "Femme"
CDE	Handedness	Group	Demographics		
Description		Type	Values	Format	
Describes the tendency of the patient to use either the right or the left hand more naturally than the other.		Polynomial	Right; Left; Ambidextrous	"R"; "L"; "A"	
Data Set	Source	Column name	Type	Values	Format
ADNI	PTDEMOG.csv	"PTHAND"	Binominal	Right; Left	1; 2
EDSD	EDSD-multicenter.csv	"Handedness"	Polynomial	Right; Left; Ambidextrous	"R"; "L"; "A"
PPMI	Socio-Economics.csv	"HANDED"	Polynomial	Right; Left; Ambidextrous	1; 2; 3
Niigarda	NA	NA	NA	NA	NA
Lille	TBD	"Lateralite"	Polynomial	Right; Left; Ambidextrous	"D"; "R"; "A"
CLM	NA	NA	NA	NA	NA

Figure 5: Example of common definitions for two variables across datasets

Contributions:



- All hospital and MIP representatives from DGDS participated actively in the definition of the “minimal” data set that is now used in the MIP for testing and validating data harmonisation procedures at the Meta Data Register (MDR) level. EPFL, CHUV, TAU and UCL contributed to the selection of the variables.

3.5.3.2 DATA > Hospital Clinical Data - Brain imaging-Genetic-Clinical (EHR)

Description: Anonymised Data-only available via the MIP aggregation and analytic functions.

Progress:

- For cases with available genetic information (e.g. ApoE status), brain MRI and clinical test results, we provide anonymised data accessible via the MIP in the form of aggregates on top of the corresponding disease signature (here, Alzheimer’s disease).

Contributions:

- CHUV SP8 team prepared the data from open data bases (ADNI) and clinical data.

3.6 T8.2.2 - Hospitals and Information System Departments Relationship Management

3.6.1 Key Personnel

Task Leader: Giovanni FRISONI (UNIGE)

Other Researcher: Pegah SARKHEIL (UKAACHEN)

3.6.2 SGA1 DoA Goals

- To initiate, monitor and coordinate interactions with participating hospitals, which are in connection with MIP and its network.
- To establish appropriate conditions for rapidly deploying the MIP solution throughout Europe and constitute a maintained database of contacts for the project lifetime.
- To follow interactions with hospitals interested in connecting with MIP and its network.
- A database exposing useful information must be exposed through MIP Knowledge Base.

3.6.3 Component Progress

3.6.3.1 SOFTWARE > Hospital Databases Bundle (HDB) > database containing information of MIP solutions adopted at hospital level

Description: We constituted and are currently maintaining and updating a database of variables, contacts, and standards adopted in the SGA1 engaged Hospitals (i.e.: Lille Hospital, Tel Aviv Hospital, Milano Hospital, Freiburg Hospital, CHUV Hospital). This database is exposed through the [MIP Knowledge Base](http://193.204.145.80:1313/documentation/#mip_variables) (http://193.204.145.80:1313/documentation/#mip_variables) via Hugo static technology (i.e.: Markdown, JQuery, Tipue-Search).

Progress:

- A. Creation of the KB static sections and database.
- B. List of variables available in hospitals as well as other research cohorts have been created.
- C. List of Hospitals’ contacts.

Contributions:

- CHUV and UNIGE prepared the list conjointly.

3.7 T8.2.3 - Research Initiatives and Community Engagement

3.7.1 Key Personnel

Task Leader: Bogdan DRAGANSKI (CHUV)

Other Researcher: Giovanni FRISONI (UNIGE)

Other Researcher: Jean-François DARTIGUES (UBO)

3.7.2 SGA1 DoA Goals

T8.2.3 will initiate and coordinate interactions with the main European and international research initiatives interested in connecting to the MIP network. Operatively, T8.2.3 will compile a list of research studies and consortia efforts focused on neurodegenerative disease to be contacted and recruited. T8.2.3's priority will be to focus on data sources and initiatives that could become federated partners through the MIP. The highest priority sources include: EMIF-AD, PharmaCOG, ARWIBO, EuroPOND, BrainGIFT, GAIN and EGI-MoBrain.

Other activities are: Organising workshops and user engagement activities 1) For external method developers, via the MIP Knowledge Base (e.g. forum discussions) where they can describe the methods developed, create guidelines. 2) For all users: online chats, online votes on a list of suggested new features or improvements ("e-news"), discussions in focus groups. - Wide broadcasts of videos and tutorials explaining not only the existing functionality, but also the vision, short and long term benefits. This will include progression to discovery of Biological Signatures of Diseases, opening platform usage to different disease spaces other than brain, discovering correlations between diseases and their treatments, wanted and unwanted effects of treatments etc. - Recording and publishing focus groups meetings - Workshops, lecturers at HBP and non-HBP education events - Presentations and information exchange with other similar projects worldwide, and the EC - Ideally "roadshows" in neurology units of big hospitals to specifically attract expert users and joining hospitals.

We believe that the end users need to understand the large goal and long-term benefits of the Platform to be motivated to contribute effectively to its development.

3.7.3 Component Progress

3.7.3.1 SERVICES > Connection to 3rd party application and data

Description: Proof-of-concept for the usability of the MIP in the context of two international multi-centre initiatives - AETIONOMY and Traumatic Brain Injury (TBI). Additionally, we analysed 38 international databases with 362,745 patients.

Progress:

- POC of MIP usage for the TBI initiative is well-advanced; potential co-funding for the MIP on TBI side is in discussion.
- Among the 38 databases, 251,637 (69%) records contain longitudinal information, and 251,917 (69%) of them are rich in seven or more types of data (including clinical, demographic, cognitive abilities, imaging, genetic, epidemiologic and biological).
- Providing access to two of the three population-based cohorts (Paquid and 3C-Bordeaux). For the Paquid cohort, 137 subjects, aged 92 years and older, were interviewed at the 27-year-follow-up. Follow-up at 17 years of the 3C-Bordeaux cohort. At this time, 250 subjects (aged 82 years and older) were interviewed.

Contributions:

- CHUV SP8 team promoted the MIP and monitored the POC phase.



- UNIGE team provided the analyses of the 38 databases.
- UBO team provided raw data and analysis of the Paquid and 3C-Bordeaux databases.

3.7.3.2 SERVICES > MIP Marketing & Promotion

Description: Any actions taken for raising profile and marketing MIP to solve audience and users' needs. E.g. leaflets, booklets, Twitter, LinkedIn, Facebook accounts and news.

Progress:

- Twitter and ResearchGate presence of MIP established.
 - 120 tweets.
 - 20 events participation scheduled for 2017 (see events list) with participation of SP8 people.
 - We produced one MIP Leaflet and one flipbook.
 - We published two videos and two to three more are in the pipeline for SGA1.
 - We are co-organising the 5th HBP School in November 2017.
 - Videos published on Youtube - one HBP official channel, one MIP channel, with 620 and 165 views respectively.
 - We also printed 200 flyers distributed at different exhibitions and made a 3D brain to demonstrate the impact of AD on the Human Brain.

Contributions:

- CHUV SP8 team created the concept, social media accounts and provides the maintenance.

3.7.3.3 SERVICES > User Support, Community Outreach & Communication > Community Outreach & Communication

Description: For direct promotion of the Platform to users, focusing on benefits and functionality.

Progress:

- User's guide for the MIP created.

Contributions:

- CHUV SP8 team.

3.7.3.4 MIP - SERVICES > User Support, Community Outreach & Communication > Knowledge Base Content Development, User Guide

Description: The MIP Knowledge Base is a user-oriented interface of MIP services, tools, and available data. KB is currently developed using the Hugo framework for all the static components (i.e.: user guides, documents, videos etc.) and a LAMP setup for the dynamic sections (i.e.: database queries, interactive questionnaires, etc.). The latest version (still hosted in the Fatebenefratelli/UNIGE development site) will soon be moved to the official GitHub location after replacing its dynamic components in favour of other static technologies such as graphQL database.

T8.5.1 Progress: [UNIGE/CHUV] The MIP Knowledge Base is available at the following link: <http://193.204.145.80:1313/documentation/>. It is a user/hospital oriented documentation of the MIP services, tools, and available data. It provides a combination of both static and dynamic contents. Static content ranges from user guides to media links, while dynamic content provides the user with the means to interrogate databases containing the MIP variables description or the hospitals information system. What was initially realised with the Liferay Portal technology, is now being redesigned by CHUV and implemented by UNIGE



using the slighter Hugo framework coupled to standard LAMP setup for the dynamic sections (database queries, interactive questionnaires, etc.). The stable release is planned for M18.

3.7.3.5 SERVICES > User Support, Community Outreach & Communication > User support & Operational Strategy (user support team, contact, process)

Description: User support team and elaboration of strategies to manage users and meet their expectations.

Progress:

- Stable release of the Liferay Portal technology planned for M18.

Contributions:

- UNIGE team.

4. WP8.3 - Data Features, Tools and Biological Signatures of Disease

4.1 Key Personnel

Work Package Leader: Mira MARCUS-KALISH (TAU)

4.2 WP Leader's Overview

Contribution to clinical neuroscience and clinical benefit: the WP made significant progress towards the development of advance data mining, neuroimaging tools and optimised analytical techniques in line and in parallel to the efforts to provide an accurate definition of "Disease Signatures". The understanding of this concept has been progressing steadily, both from theoretical and empirical points of view, and in the ways to quantify it. We have learned that the "Disease Signature" we are targeting is much more complex than what the regular used term implies - a single genetic marker associated with a known disease - a conclusion based on a comprehensive literature review. An example for implication of this understanding is the work on data driven identification of Parkinson's disease subtypes. Relating different patterns of clinical presentation (i.e. cognitive deterioration, psychiatric symptoms, postural instability dominance etc.) to genetic mutations groups (i.e. LRRK2, GBA) and markers in gait analysis. The groups at TAU and JSI and ULM developed the tools while taking into account this complexity.

Contribution to the deployment: under the fast track plan proposed to the EC review, the Task Leaders of this Work Package contributed towards the goal of accelerating the deployment and use of the MIP by clinicians. The Tasks Leaders participated in the Data Governance and Data Selection (DGDS) Committee to provide, at the early stage, recommendations concerning the hypotheses and the selection criteria for the data (e.g. selecting the modality and the resolution of neuroimaging data to ensure best data quality).

Contribution to the Platform: the WP contributed to the specification of the algorithms developed during this period which are all part the components SOFTWARE > Algorithm Library. The Library is used by the Platform developer (WP8.1 and WP8.5) to integrate the data mining tools in the MIP. One algorithm from each group (3C from TAU, Predictive Clustering from JSI and t-sne from ULM) has been selected and prepared in order to be part of the first release deployed in the hospitals. The UCL brain morphology tools, which are based on functionality from the leading software in neuroimaging, have been already integrated to the MIP/data factory.

Everything went according to plan.

The algorithms and strategies developed in this Work Package are critical for achieving one main Milestone - the translation and adaption of research software for use on hospital data. It demands gaining further knowledge of the data, pre-processing and imputation schemes, as well as simulation and modelling to evaluate the tools, as well as the fine tuning of the algorithms themselves. Several papers have been submitting that describe the results obtained during this period. All the Milestones related to the specification and creation of proof-of-concept prototypes for the platform have been achieved.

4.3 Priorities for the remainder of the phase

Our research is centred on preparing the ground for identifying disease signatures from hospital data. In preparation and while acquiring the hospital data, we developed the tools and tested and verified them on research data from various sources. In addition, we made two major advances in the last 12 months: (i) A Parkinson hospital cohort including 1,500 Parkinson patients and 600 of their relatives from the Tel Aviv Medical Centre; and (ii) Data from Rome's Policlinico Gemelli hospital, are being prepared for analysis in TAU. The long



and laborious pre-processing work on these two cohorts, represents the difficulties and barriers that need be overcome when working with hospital sources from different countries. We are almost ready to start using various modes of analysis on these two sources of data.

The analysis methods will be based on joint efforts of the various groups. This will involve utilising and aligning the various tools, providing feedback and different insights to ensure reliable and replicable results. It includes recently upgraded versions of the 3C strategy and visualisation techniques that have been developed at TAU, mining data streams for multi-target regression and multi-label classification and prototypes implementation developed by JSI, as well as new methods for re-description mining. We will make special efforts to integrate the image feature extraction methodologies into these analyses. The datasets being developed lack wide coverage of genomic information, so there is little potential use for the progress in whole genome feature extraction. By the end of this phase, we envisage that these efforts will yield new medical insights.



4.4 Milestones

Table 3: Milestones for WP8.3 - Data Features, Tools and Biological Signatures of Disease

MS No.	Milestone Name	Leader	Task(s) involved	Expected Month	Achieved Month	Comments
MS10	MS 8.3.1 Initial Implementation Plan for Data Mining Algorithms, Concepts and Applications Formulated - Integration plan and method to the MIP and the BRAIN factories.	TAU	All WP8.3 tasks	M02	M02	WP8.3 contributed to the revised MIP Platform Fast-Track Plan for the work to be carried out by the SP8 team during SGA1.
MS104	MS 8.3.2 Initial Proof-of-concept and Results of the Different Algorithms	TAU	T8.3.2	M12	M12	Work was performed on exploring large research medical dataset of patients diagnosed with Parkinson's disease in collaboration with Tel Aviv Medical Center. We aim to explore all measures collected, without conducting any pre-selection. The purpose of the research is to discriminate phenotypes measures as defining characteristic of the disease among two groups of genetic mutations carriers which are known to have an association with the disease. We managed to do so and found significant results that shows a strong connection of one group with more cognitive and motor difficulties, while the other group exhibits more psychiatric difficulties and worse reaction to drug treatment across multiple measures. The results are statistically valid based on the strict significance level kept due to multiplicity correction conducted all along the different stages of the analysis and due to the avoidance of unadjusted for selection of the data in-hand



MS112	MS 8.3.3 Results and Platform Implementation of the Large-Scale Simulation Analysis	TAU	T8.3.1	M12	M12	The initial results of simulation analysis guided the targeted selection and improvement of methods implemented in the 3C strategy. The fruitful collaboration with Policlinico Gemelli hospital in Rome and The Biostatisticians at Associazione Fatebenefratelli per la Ricerca in Isola Tiberina, (Rome) provided additional opportunity to utilise and further develop, test and improve the simulation and 3C strategy on real hospital data. Work is therefore progressing towards a multi-source results and implementation in accordance with SP8 goals.
MS113	MS 8.3.4 Results and Platform Implementation of Algorithm to Generate Gene Expression Maps	LUMC	T8.3.10	M12	M12	Prototype algorithm for gene heatmap generation ready for Platform implementation
MS118	MS 8.3.5 Prototypes of the distributed stream-based methods for learning predictive clustering trees (e.g., for multi-target regression) developed	JSI	T8.3.5	M12	M12	We have developed new methods for mining data streams for multi-target regression and multi-label classification and implemented their prototypes (Osojnik et al. 2016b, 2016c, 2016d), thus achieving the Milestone MS 8.3.5.
MS119	MS 8.3.6 Prototypes of the new redescription mining algorithms developed.	JSI	T8.3.6	M12	M12	We have developed new methods for redescription mining and implemented their prototypes (Mihelčić et al. 2016, 2017, and one paper submitted to PlosONE), thus achieving the Milestone MS8.3.6. Note: Unfortunately, for the two papers Mihelčić et al. 2016 (DOI: 10.1016/j.eswa.2016.10.012), 2017 (DOI: 10.1007/s10844-017-0448-5) we forgot to include HBP acknowledgements and they are not in the publications database.
MS120	MS 8.3.7 Prototype of the ontology for describing data on patients with neurological diseases developed.	JSI	T8.3.9	M12	M12	We have developed a prototype of an ontology for describing data on neurological diseases in patients (Panov et al. 2016, Soldatova et al. 2016), thus achieving the Milestone MS 8.3.7.
MS126	MS 8.3.8 Initial implementation of image factorisation method without distributed computing	UCL	T8.3.11	M12	M12	There is an initial MATLAB implementation of an algorithm for factorising large datasets of images.



4.5 T8.3.1 - Tools to Mine Replicable Selection and Integration of Hierarchical Features, Inter and Across Domains using FDR

4.5.1 *Key Personnel*

Task Leader: Mira MARCUS-KALISH (TAU)

4.5.2 *SGA1 DoA Goals*

The focus of T8.3.1 is to continue the development of the 3-C approach for mining disease signatures based on the RUP. In this Task we develop tools to mine replicable selection and integration of hierarchical features, inter and across domains, using FDR.

4.5.3 *Component Progress*

4.5.3.1 SOFTWARE > Algorithm Library > Statistical Analytics > Integrating multi-domain data

Description: Introduction of new biomarkers domains.

Progress: Work was done on exploring large research medical and motor dataset of healthy subjects with family members who were diagnosed with Parkinson's disease, in collaboration with Tel Aviv Medical Centre. The purpose of the research is to Incorporate longitudinal information and new biomarkers of healthy relatives of Parkinson's disease patients. Subjects belong to two groups of genetic mutations carriers which are known to have an increased risk for disease and one group with no known mutations. Our results show significant strong connection of one mutation (GBA) with less motor stability among healthy subjects. The statistical validity of the results is ensured by using a stringent significance level. This arises from the multiplicity correction along the different stages of the analysis and a conscious effort to avoid selective inference. These results support the hypothesis of different development routes of Parkinson's disease for different genetic mutations. The result interpretation points that there could be 3 different disease subtypes with the same diagnosis, Parkinson's disease. The differences between healthy mutation carries and healthy subjects without mutations stimulate the continuation of the work to seek for early markers of Parkinson's disease.

The graphs below show different measurements of stability between healthy mutation carriers (3) and healthy subjects without any known mutations (0)

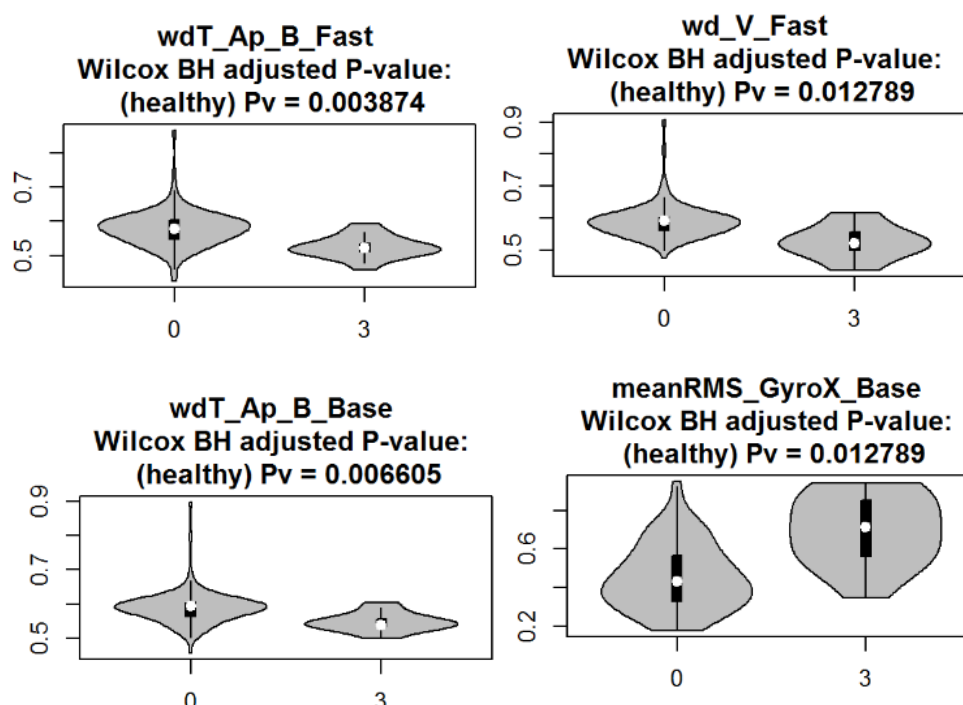


Figure 6: Measurements of stability between healthy mutations carriers and healthy subjects

4.5.3.2 SOFTWARE > Algorithm Library > Statistical Analytics > 3-C Longitudinal Modelling

Description: This Component will be part of the further development of the 3-C strategy, and it refers to incorporating longitudinal information (day-to-day and multiple patient visits).

Progress: ADNI longitudinal dataset designed. Development of procedure for longitudinal dataset in the ADNI data. A total of 3,042 patients, from which 1,738 have more than one visit with a median of three visits per patient. This will serve as the basis for development of temporal disease signatures. These algorithms will serve the analysis of patient data in the MIP once incorporated into the Platform.

4.5.3.3 SOFTWARE > Algorithm Library > Statistical Analytics > 3-C (Categorize, Cluster & Classify)

Description: Methodology for Medical big data analysis and disease sub-type identification.

Progress: The 3C Methodology was further developed, more algorithms for feature selection explored; an article describing methodology "Data and Knowledge driven strategy for disease subtypes identification, Mitelpunkt A, Galili T, Kozlovski T, Markus-Kalish M, Cui J, Shachar N, Benjamini Y" was submitted.

Over the last few months there has been extensive work in order to implement the 3C strategy on another dementia patient dataset: the pre-processing of the data (1,843 different neurological patients across 14 different labs in Rome) and its preparation for analysis; understanding the structure of this new database, as well as the meaning of the variables; all of this while overcoming obstacles of language, domain knowledge and workflows differences. All of these serve to adjust and improve the algorithms and strategy

for incorporation in MIP as we expect to face similar challenges with the MIP hospital data. Data variables were labelled into Clinical Measurements (total of 195) and potential biomarkers (200), including variable types and their logical range, in order to use the semi-automatic non-linear transformations procedures, as part of the preparation process to run the 3C strategy. We used 'MipMap' tools - originally planned for fast preprocessing and datasets integration - collaborating with AUEB partners in SP8.

A procedure for matching between same variables appearing differently across files was needed. The best resemblance between variables (d_{min}) was determined using 'Stringdist' R function and the factor (d_{min}) over variable length has been tested

$$f = \frac{d_{min}}{var\ length}$$

The results were verified manually. This is expected to assist the process of incorporating new clinical datasets (hospitals) into MIP.

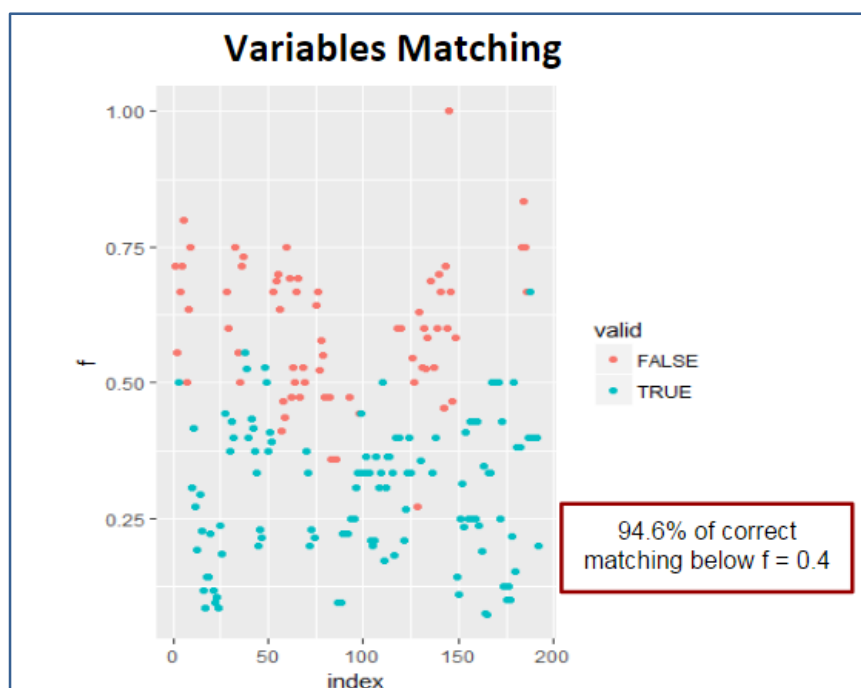


Figure 7: Variables matching

As preparation for the second step of 3C, 19 out of 195 Clinical Measurements (CM) were selected, based on significance regression models with 'diagnosis' as explanatory variable to each CM. Imputation was done in order to complete missing values;

4.6 T8.3.2 - Developing Methods for High-Dimensional Data with Possible Informative Missing Values

4.6.1 Key Personnel

Task Leader: Mira MARCUS-KALISH (TAU)

4.6.2 SGA1 DoA Goals

The goal of this Task is to address the workflow variability among hospital and physicians regarding patients' treatment, as reflected by missing values in hospital data. This task will develop methods for high-dimensional data with possible informative missing values.

4.6.3 Component Progress

4.6.3.1 SOFTWARE > Algorithm Library > Statistical Analytics > Methods for high-dimensional data with possible missing values

Description: Symmetry targeted monotone transformations, and the advantage gained in variance stability, linearity and clustering.

Progress: Transformations in medical big data: the article “The importance of non-linear transformations use in medical data analysis” was submitted. RUP work was implemented on the Parkinson TLVMC Hospital cohort, continued and extended, Statistical procedures and workflows for missing values: progress-development of visualisation and statistical methods to understand structure of missing values. Further progress was made on the *heatmaply* R package for presenting cluster heat maps. New features were added, including sidebar annotation, as well as a better detection algorithm of the number of clusters (using silhouette coefficient in the *dendextend* R package). The interactive cluster heat map is an effective method for visualising and exploring patterns of informative missing values. For example:

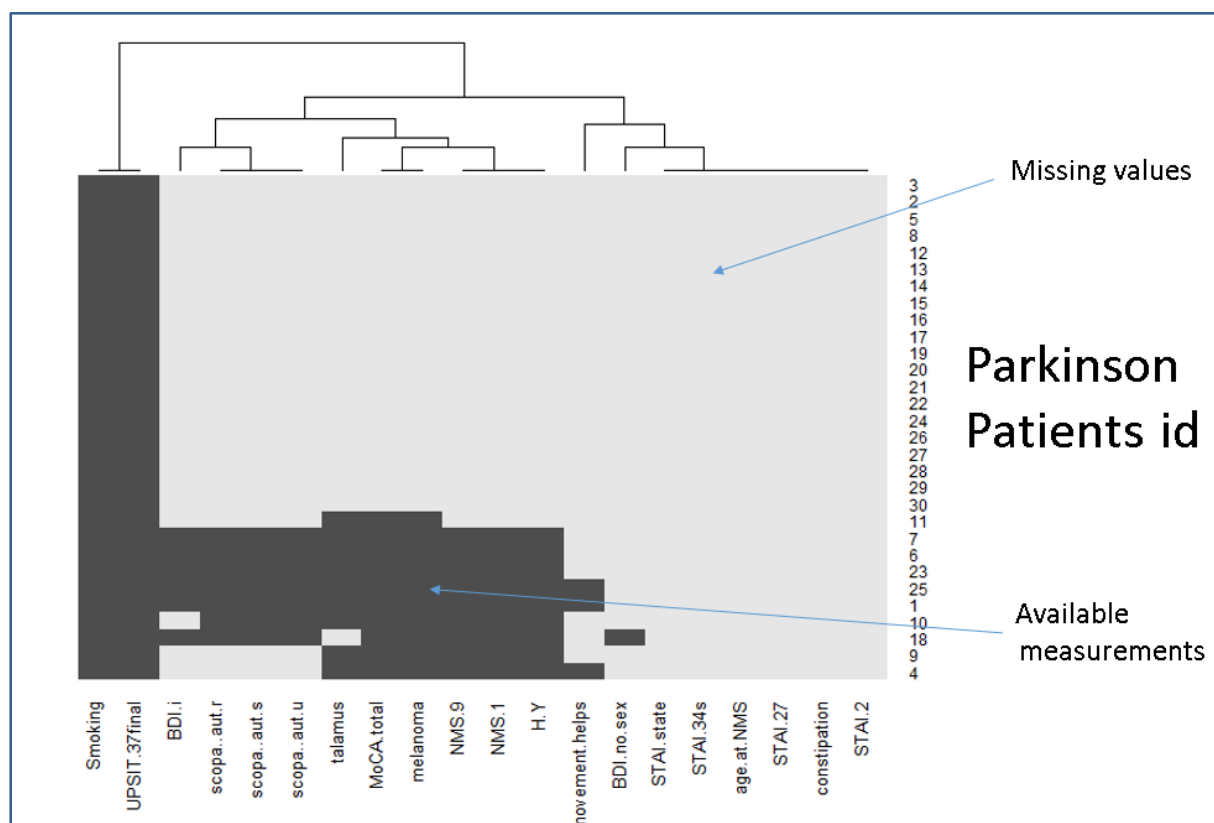


Figure 8: Missing values in hospital's clinical measurements

The *heatmaply* R package is now freely available through CRAN: <http://cran.r-project.org/package=heatmaply>. A paper on the package was submitted. Also, this work will be presented in the upcoming user!2017 conference in Belgium.

Displaying the patterns of missing values (grey to squares) using cluster heat map visualisation in a sample of Parkinson patients' data matrix. Missing values are abundant in hospital data due to: variance in hospitals' protocols regarding which medical tests to use; new medical measuring devices; and variability in diagnosis by the physician. Further theoretical progress was made regarding the testing of an interaction effect of change in NA patterns between two groups between two time points. This could be measured by treating NA presence as a binary yes/no variable and comparing based on RD (Risk Difference) (also for RR (Relative Risk), and OR). An initial R code was written for implementing the methods developed thus far.



4.7 T8.3.3 - Introducing Selective Inference into Dimensionality Reduction and Clustering Methods

4.7.1 *Key Personnel*

Task Leader: Yoav BENJAMINI (TAU)

4.7.2 *SGA1 DoA Goals*

The goal of this Task is introducing Selective Inference into Dimensionality Reduction and Clustering Methods.

4.7.3 *Component Progress*

4.7.3.1 SOFTWARE > Algorithm Library > Statistical Analytics > Clustering: incorporating Knowledge into the process

Description: Developing advanced methods for selective inference, use of Selective inference on multiple families of hypotheses.

Progress: Work was performed done on exploring large research medical datasets of patients diagnosed with Parkinson's disease in collaboration with Tel Aviv Medical Center. We aim to screen, explore and infer from all measures collected, without conducting any pre-selection while controlling the error rate during the process in order to guarantee the statistical validity of the results. The purpose of the research is to discriminate phenotype measures as defining characteristics of the disease among two groups of genetic mutations carriers which are known to have an association with the disease. We managed to do so and found significant results that show a strong connection of one group with more cognitive and motor difficulties, while the other group exhibits more psychiatric difficulties and a worse reaction to drug treatment across multiple measures. The statistical validity of the results is ensured by using a stringent significance level that arises from the multiplicity correction along the different stages of the analysis and a conscious effort to avoid selection. The results are statistically valid based on the strict significance level kept due to multiplicity correction conducted all along the different stages of the analysis and due to the avoidance of unadjusted for selection of the data in-hand. The suggested method and results were presented in: 1. MAESTRA summer school in Ohrid, Macedonia - student's poster session 2. Fall IAB 2016 at UMPBC and won the 1st place as best student poster presented. The paper summarising the medical results is in advanced stages before submission. Moreover, during June 2017 the work also will be presented at the International congress of Parkinson's disease and movement disorder Vancouver Canada.

Visualisation from the upcoming paper (work in progress):

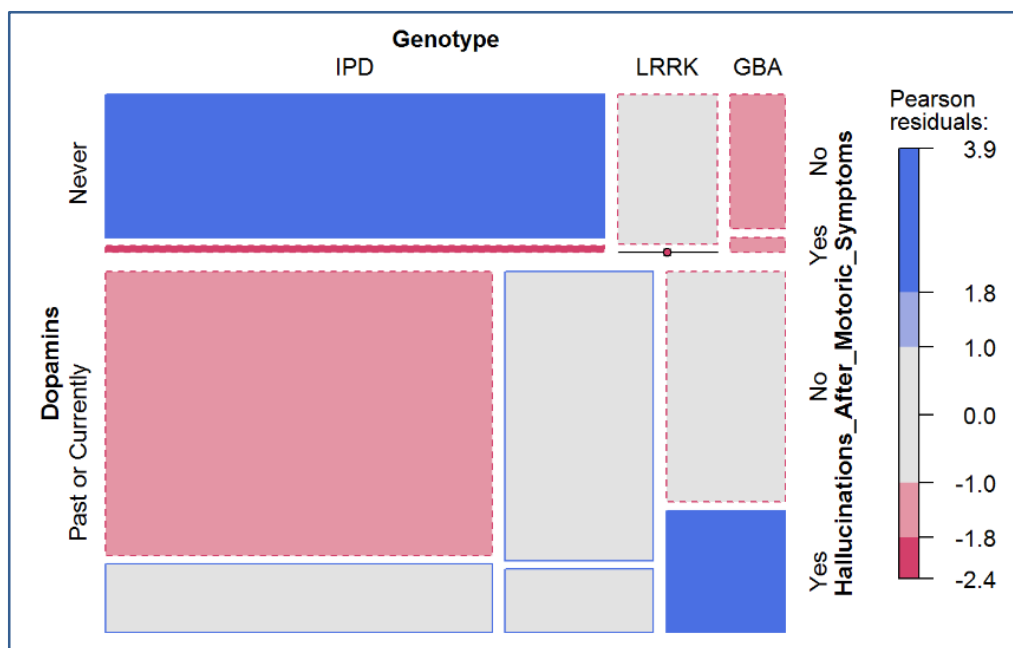


Figure 9: Genotype vs symptoms vs dopamines

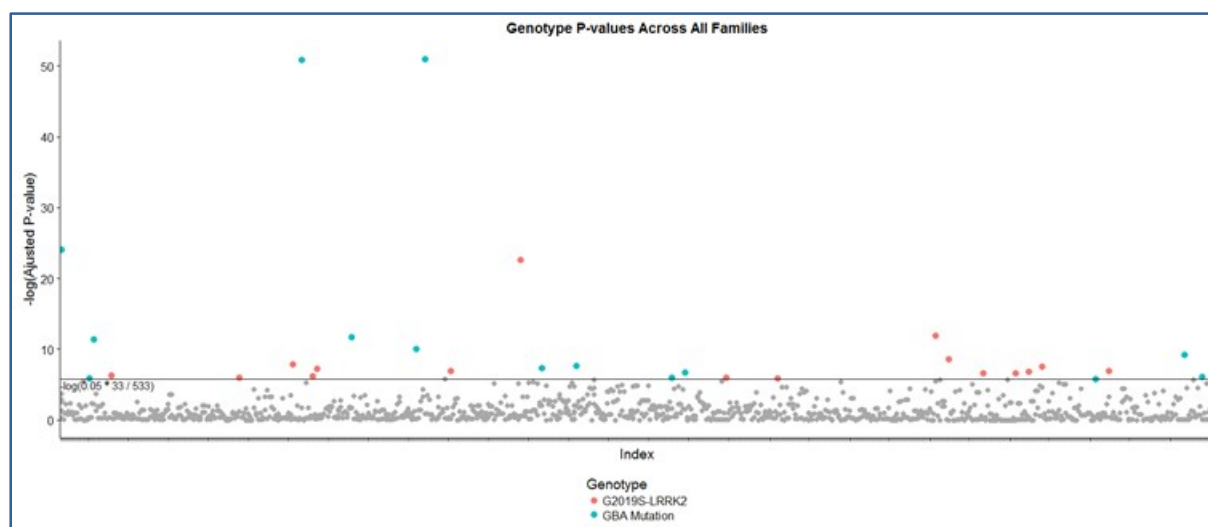


Figure 10: Genotype P-values across all families

4.8 T8.3.4 - Statistical Methods for "Disease Signature" Confidence Assessment

4.8.1 Key Personnel

Task Leader: Yoav BENJAMINI (TAU)

4.8.2 SGA1 DoA Goals

The goal of this Task is to develop a statistical approach for assessing the level of confidence of a data-mined signature, i.e. statistical methods for "Disease Signature" confidence assessment.



4.8.3 *Component Progress*

4.8.3.1 SOFTWARE > Algorithm Library > Statistical Analytics > Disease Signatures - concept and methodology

Description: Define and propose a model for disease signature.

Progress: A systematic literature review was performed; a poster summarising initial results was presented at the HBP Summit in Florence. Implementation and demonstration of the concepts on different diseases is completed. Initial conceptual models for disease signature representation have been developed. We have concluded that a model for “Disease Signature” should consist of a concept of a marker or markers that will point to the disease cause and of an undesired effect as a manifestation of the disease. A unique combination of those markers and effects will define the signature. In addition, for a standard and clear way to define a disease this might help better identify disease sub-type and facilitate data mining approaches.

4.9 T8.3.5 - Methods for Distributed Rule-Based Disease Signature Discovery

4.9.1 *Key Personnel*

Task Leader: Sašo DŽEROSKI (JSI)

4.9.2 *SGA1 DoA Goals*

The goal of the Task is to provide methods for distributed rule-based modelling for disease signature discovery.

4.9.3 *Component Progress*

4.9.3.1 SOFTWARE > Algorithm Library > Machine Learning Library > Disease signature: Distributed rule-based methods

Description: In the RUP, we developed rule-based clustering methods for finding disease signatures that are based on predictive clustering trees and predictive clustering rules for solving different tasks of predicting structured outputs (e.g., multi-target regression). We considered both the batch learning setting and the streaming setting, but not a distributed setting. Within this Task, distributed versions of the tree- and rule-based methods for predictive clustering developed during the RUP are being developed and evaluated. The methods are implemented in an environment that integrates stream-based processing and distributed processing. Namely, they work on distributed stream processing engines, which allow us to express parallel computation on streams and combine the scalability of distributed processing with the efficiency of streaming algorithms.

Progress: The Task is being executed as planned. We explored how different local and global tree-based approaches for multi-target regression compare in the streaming setting. Namely, recent studies in the batch setting have shown that global approaches, predicting all of the targets at once, tend to outperform local approaches, predicting each target separately. We applied a local method based on the FIMT-DD algorithm and proposed a novel global method, named iSOUP-Tree-MTR. We also performed an experimental evaluation that explored the differences between the local and the global approach. We observed that the local FIMT-DD and global iSOUP-Tree-MTR have similar predictive performance, however, the global method is much more scalable in terms of time and memory consumption.

Multi-label classification (MLC) tasks are encountered more and more frequently in machine learning applications. While MLC methods exist for the classical batch setting, only a few methods are available for the streaming setting. We developed a new methodology for MLC via multi-target regression in a streaming setting that uses the streaming multi-target



regressor iSOUP-Tree-MTR. We experimentally compared the two variants of the iSOUP-Tree method (regression and model trees), as well as ensembles of iSOUP-Trees with state-of-the-art tree and ensemble methods for MLC on data streams. The results show that iSOUP model trees perform better than iSOUP regression trees for a large set of evaluation measures for multi-label classification.

4.10 T8.3.6 - Methods for Redescription Mining

4.10.1 Key Personnel

Task Leader: Nada LAVRAČ (JSI)

4.10.2 SGA1 DoA Goals

The goal of this Task is to provide redescription methods for disease signature discovery.

4.10.3 Component Progress

4.10.3.1 SOFTWARE > Algorithm Library > Machine Learning Library > Integrating multi-domain data: Methods for redescription mining

Description: Redescription mining is a relatively novel data mining and knowledge discovery approach that aims to find multiple rule-based descriptions of subsets of examples (e.g. patients), where each of the descriptions is based on a different set of descriptive variables (called a view). It is related to the so-called multi-view learning. The rule-based models generated in this way tend to be more reliable (described examples must be homogeneous with respect of two or more independent views) and can also be used to find interesting properties and connections between variables from different views. Initial experiments in this direction already started in the RUP and the results suggest that redescription mining is a promising research direction well suited for rule-based modelling needed within the MIP with the purpose of disease signature discovery. Our current work extends the above-mentioned work with additional approaches for redescription generation and with additional data views used; we consider all the data layers stored in MIP, such as images, genetics, proteomics, clinical scores, etc.

Progress: The Task is being executed as planned. We have applied redescription mining to the data on patients with different degrees of cognitive impairment (including the Alzheimer's disease - AD). Our goal was to relate a set of biological indicators (containing different biomarkers and genetic markers) with a set of clinical indicators (containing various cognition tests and other clinically measured symptoms such as headache, nausea, etc.). The preliminary results, demonstrating several obtained redescrptions, were presented at the International Conference on Brain Informatics and Health (BIH2015).

Further study lead to the development of three different datasets containing different indicators connected to AD and to the extension of the redescription mining algorithm used, namely the CLUS-RM, to the fully automated, general constraint-based redescription mining setting. The extended algorithm allowed us to find many interesting associations, relevant for the detection and understanding of AD. Many of the discovered associations have been already reported in the literature, however we also found several associations that are disputed - still not independently confirmed by multiple studies and one completely unexplored discovery (currently not reported in the literature). These findings and the extension made to the CLUS-RM algorithm are presented in the manuscript currently under review at the PlosOne journal.

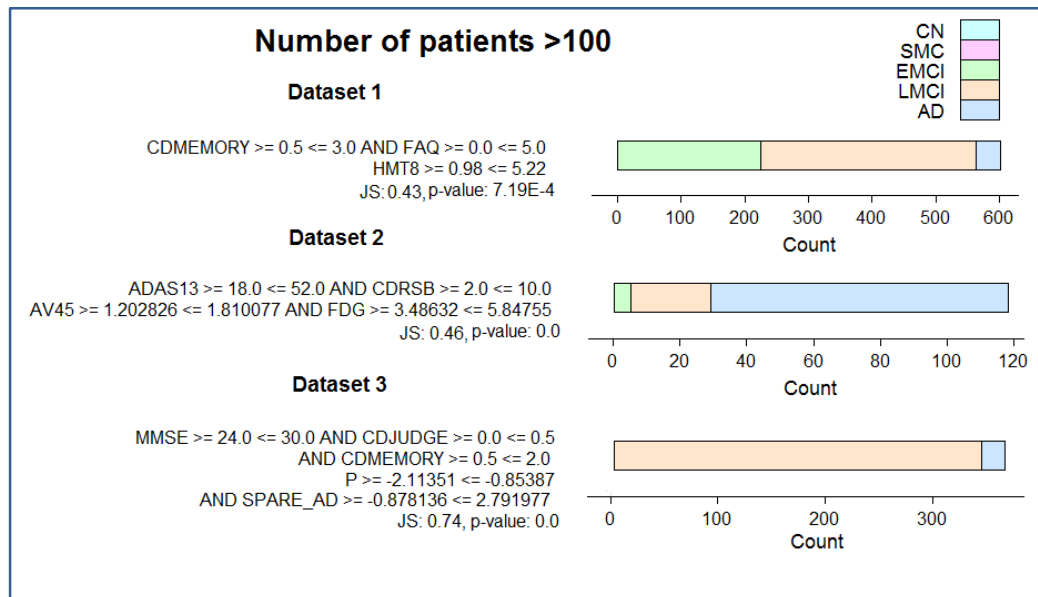


Figure 11: Redescriptions (derived from different parts of the ADNI dataset), each describing at least 100 patients.

In the figure above (Redescriptions (derived from different parts of the ADNI dataset), each describing at least 100 patients) all patients described are diagnosed with EMCI, LMCI or AD. (Preprint: <https://arxiv.org/abs/1702.06831>).

4.11 T8.3.7 - Methods for Heterogeneous Networks

4.11.1 Key Personnel

Task Leader: Nada LAVRAČ (JSI)

4.11.2 SGA1 DoA Goals

The goal of the Task is to provide methods mining text-enriched heterogeneous information networks.

4.11.3 Component Progress

4.11.3.1 SOFTWARE > Algorithm Library > Machine Learning Library > Integrating multi-domain data: Methods for heterogeneous networks

Description: In recent years, analysis of heterogeneous information networks has gained momentum. In contrast to homogeneous networks, heterogeneous information networks describe heterogeneous types of entities and different types of relations. Moreover, in enriched heterogeneous information networks, nodes of certain type contain additional information, for example in the form of experimental results or documents. We further developed the wordification technique to transform relational databases into text documents, and improved our techniques developed during the RUP for mining text-enriched heterogeneous information networks to deal with other types of node enrichment.

Progress: The Task is being executed as planned. We developed an approach for mining heterogeneous information networks by decomposing them into homogeneous networks. The proposed HINMINE methodology is based on previous work that classifies nodes in a heterogeneous network in two steps. In the first step the heterogeneous network is decomposed into one or more homogeneous networks using different connecting nodes. We improve this step by using new methods inspired by weighting of bag-of-words vectors mostly used in information retrieval. The methods assign larger weights to nodes which are more informative and characteristic for a specific class of nodes. In the second step, the resulting

homogeneous networks are used to classify data either by network propositionalisation or label propagation. We propose an adaptation of the label propagation algorithm to handle imbalanced data and test several classification algorithms in propositionalisation. We tested the new methodology on three data sets with different properties. For each data set, we performed a series of experiments and compared different heuristics used in the first step of the methodology. We also investigated different classifiers which can be used in the second step of the methodology when performing network propositionalisation. Our results show that HINMINE, using different network decomposition methods, can significantly improve the performance of the resulting classifiers, and also that using a modified label propagation algorithm is beneficial when the data set is imbalanced.

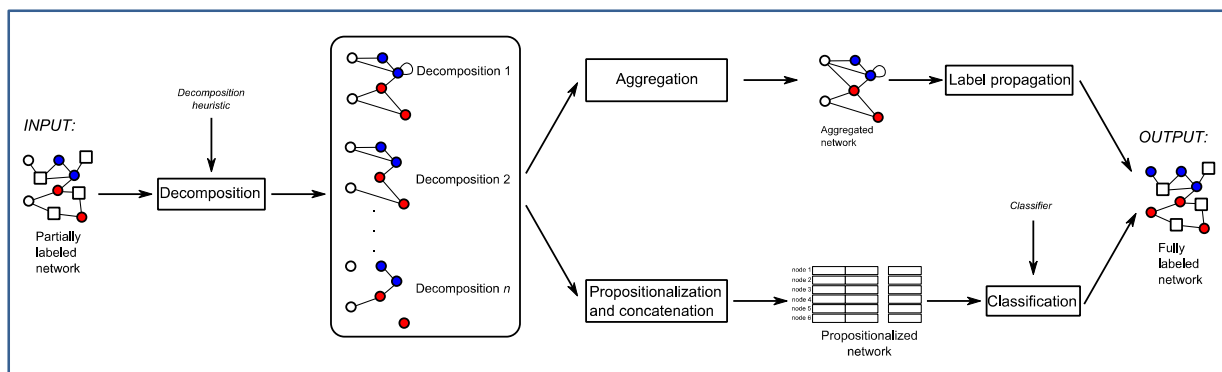


Figure 12: Overview of the proposed methodology for heterogeneous network

In the figure above you can see the overview of the proposed methodology: an input partially labelled heterogeneous network is first decomposed into one or more homogeneous networks. In the second step, these decompositions are merged and used to classify the data using label propagation (top branch) or each decomposition is used to calculate feature vectors for each base node in the network (bottom branch).

4.12 T8.3.8 - Methods for Disease Progression Modelling

4.12.1 Key Personnel

Task Leader: Sašo DŽEROSKI (JSI)

4.12.2 SGA1 DoA Goals

The goal of the Task is to provide methods for disease progression modelling.

4.12.3 Component Progress

4.12.3.1 SOFTWARE > Algorithm Library > Machine Learning Library > Longitudinal modeling: Tree-based and equation-based methods

Description: Recently, the task of modelling the progression of neurodegenerative diseases has received increasing attention. Ordinary and stochastic differential equations have been used to model the dynamics of biomarkers in the context of progression of Alzheimer's Disease. Both protein-based and image-derived biomarkers have been considered. In this Task, novel machine learning methods for describing and modelling the temporal dynamics of disease and its clinical and biological markers are developed. On the one hand, this includes predictive clustering approaches for dealing with time-changing features. On the other hand, this includes methods for automated modelling of dynamics with ordinary and stochastic differential equations. Methods for hierarchical tree-based predictive clustering of time series can relate the values of independent variables (e.g. biomarkers) to a time-series of values of a target variable (e.g. the temporal trend of a clinical score). The Task

extends these methods to consider multiple time series (e.g., clinical scores) as targets. It also extends them towards learning predictive clustering rules, which are easier to understand. The Task develops methods for learning stochastic models of dynamic systems from data and domain knowledge. As a starting point, we take the ProBMoT system, which learns deterministic process-based models that correspond to ordinary differential equations. We extend both the representational formalism of process-based modelling as well as the learning methods to deal with stochastic models (that will correspond to stochastic reaction equations or stochastic differential equations). The first part of the Task is related to the work in the RUP, while the second brings a completely new class of methods to HBP.

Progress: The Task is being executed as planned. We developed an approach for modelling dynamical systems in discrete time using regression trees, model trees and option trees for on-line regression. Some challenges that modelling dynamical systems pose to data mining approaches motivated the use of methods for mining data streams: The algorithm FIMT-DD for mining data streams with regression or model trees is described, and the FIMT-DD based algorithm ORTO, which learns option trees for regression. The experimental evaluation showed that tree-based approaches to on-line regression are appropriate for solving the task of identification of dynamical systems in discrete time. Among the considered approaches, the ORTO-A algorithm clearly stands out as the best-performing one. ORTO-A learns option trees for regression, where an example may be sorted down multiple branches of the tree in option nodes, and averages the predictions obtained from each of the branches in the option tree that an example follows.

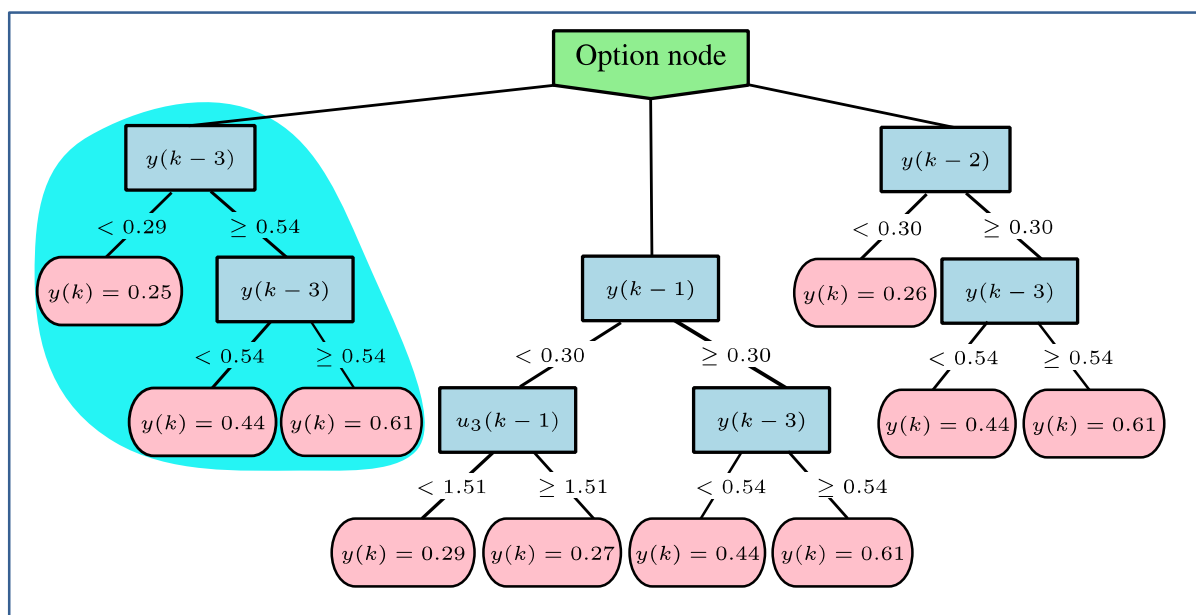


Figure 13: Example option tree learned on test data. The highlighted area represents the best tree.

We have developed a new method that represents a unified approach to modelling dynamical systems and allows for flexible formalisation of the space of candidate model structures, deterministic and stochastic interpretation of model dynamics, and automated induction of model structure and parameters from data. The method follows the paradigm of process-based modelling and learns stochastic reaction equations together with an understandable explanation of system dynamics in terms of entities and processes present in the system. The method is able to reconstruct models of dynamical systems from synthetic and real data.

We also developed a new, process-based, design methodology for computational design of dynamical systems. The new methodology combines a flexible process-based formalism for specifying the space of candidate designs with multi-objective optimisation approaches for selecting the most appropriate among these candidates. At input, it takes domain-specific



knowledge about modelling systems in the study domain and expected properties of the desired behaviour. The knowledge is formalised as a taxonomy of modelling templates and a specification of an incomplete model. The incomplete model uses inner nodes in the taxonomy to specify a set of alternative design choices; during the enumeration of the candidate model designs, these are instantiated with leaf nodes in the taxonomy that correspond to specific design choices. The parameters of each candidate model are estimated using multi-objective optimisation with objectives corresponding to the expected properties of the desired system behaviour, yielding a Pareto front of solutions for each candidate. Finally, at output, the candidate designs (model structures) are ranked according to the hyper-volumes under their Pareto fronts obtained with multi-objective optimisation. For each design, the output contains its structure, parameters and simulated behaviour.

4.13 T8.3.9 - Ontologies for Describing Data on Neurological Diseases in Patients

4.13.1 Key Personnel

Task Leader: Sašo DŽEROSKI (JSI)

Other Researcher: Bogdan DRAGANSKI (CHUV)

4.13.2 SGA1 DoA Goals

The goal of the Task is to provide ontologies for describing data on neurological diseases in patients.

4.13.3 Component Progress

4.13.3.1 SOFTWARE > Algorithm Library > Machine Learning Library > Ontologies for describing data on neurological diseases, patients

Description: We have recently developed OntoDM, a suite of ontologies for describing data and data mining. OntoDM includes the OntoDT ontology of datatypes, the OntoDM-core ontology of essential data mining entities and OntoDM-KDD, used for describing knowledge discovery processes. These ontologies include the information artefacts necessary to describe data, algorithms and models (learned by the algorithms from the data). They are designed according to the latest/best practices in ontology engineering and are interoperable with other ontologies. Based on OntoDT, in this task we develop a mid-level ontology for describing various types of data on patients with neurological diseases. On the one hand, this should include specific data types, such as neuro-images, genetics data, proteomics data, and clinical scores. It should also include features derived from these types of data. Taken together with OntoDT and OntoDM-core/KDD, the new ontology facilitates the creation of data analysis workflows for the datasets described with its terminology.

Progress: The Task is being executed as planned. We have developed a prototype of an ontology for describing data on neurological diseases in patients. This is a mid-level ontology, which takes as an upper-level ontology OntoDT, the ontology of data types.

The ontology was developed in a hybrid manner, combining bottom-up and top-down approaches. Following the bottom-up approach, we started from two instances of datasets used previously/currently for data analysis, i.e. the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset and the Parkinson's Progression Markers Initiative (PPMI) dataset. Following the top-down approach, we used the documentation of ADNI and PPMI studies (study objectives, study protocol, study procedures, schedule of activities). We mapped terms appearing there to terms from bio-medical ontologies and vocabularies (from <http://biportal.bioontology.org/>).

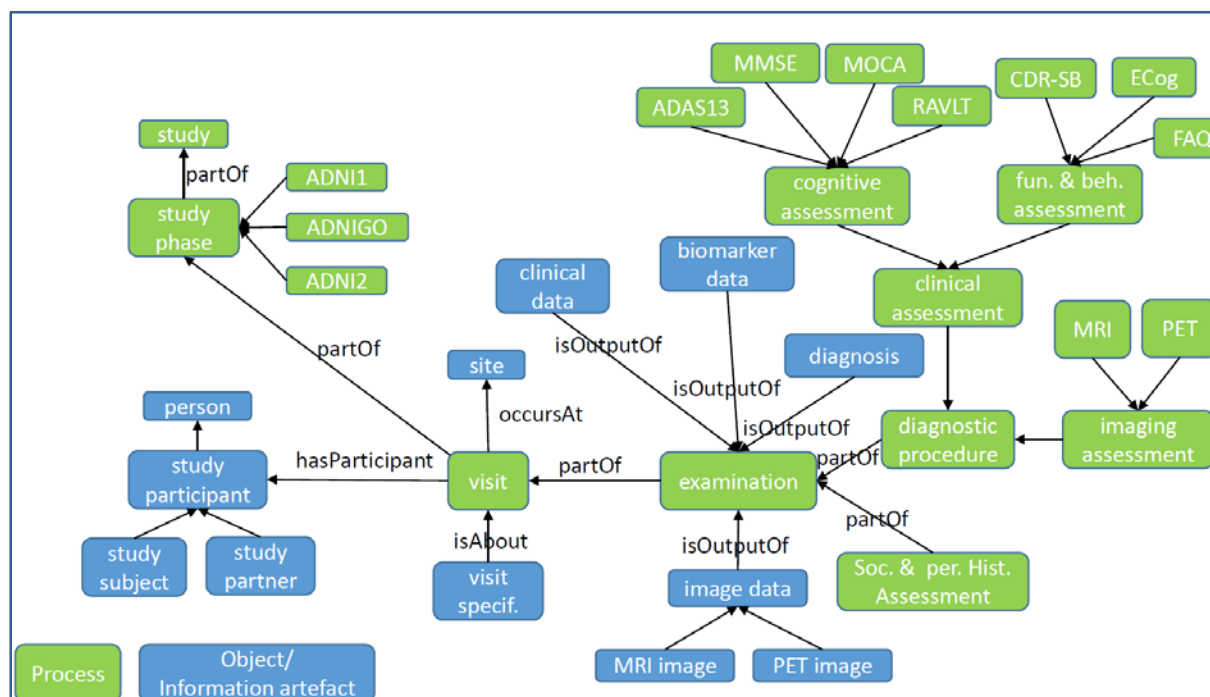


Figure 14: Excerpt from the mid-level ontology for describing data on neurological diseases in patients

Contribution:

- The ontologies were developed in collaboration between JSI and CHUV: JSI provided the expertise on ontology building and their existing ontologies OntoDM and OntoDT, while CHUV provided their expertise on neurological data in order to be included in the ontologies.

4.14 T8.3.10 - Methods for Linkage of Local SNP Data (Individual SNPs) to Imaging Data through SNP

4.14.1 Key Personnel

Task Leader: Boudewijn LELIEVELDT (LUMC)

4.14.2 SGA1 DoA Goals

This Task will develop technologies to link locally acquired SNP data of individuals (local layer) to heterogeneous imaging and non-imaging data in the context of regional gene expression. We therefore aim to develop mapping techniques to map SNPs to single genes, then to express genes as spatial expression heat maps, and then to register these heat maps to clinical imaging data. Through correlation with aggregate features that can be computed in the local layers (population deformation maps), the Task aims to enable an integrated mining of imaging and genetic features. To this end, T8.3.10 will develop a novel technique to generate gene expression maps of the brain from the Allen Brain Atlas, and register these to imaging data. These methods can serve as a bridge between locally stored gene expression and imaging data, and the aggregated hospital data in the MIP. This work package links to WP8.2 and WP8.5.

4.14.3 Component Progress

4.14.3.1 SOFTWARE > Algorithm Library > Brain Anatomy > GeneHeatMapper

Description: Algorithm that generates a 3D expression heat map of an SNP name, gene name or co-expression module.

Progress: In M1-M12, LUMC developed a Matlab prototype code for generation of 3D gene expression heatmaps for single genes and clusters of co-expressed genes (see Figure 1).

This builds upon the recently released novel dimensionality reduction technique dual-tSNE was introduced for visualising gene-gene coexpression throughout the brain (www.brainscope.nl). In year one, we concentrated on a module to generalise this concept towards generating a 3D brain heatmaps of gene-gene co-expression of multiple genes. See the figure below for an example. This first version prototype module is ready for delivery to the other Partners in SP8 for testing as of April 2017. After testing and refinement, we expect that integration into the MIP will be possible as of October 2017. Proof of concept studies in clinical use cases are ongoing and reported in 8.4.4, some of which have been published / submitted in a number of journal papers.

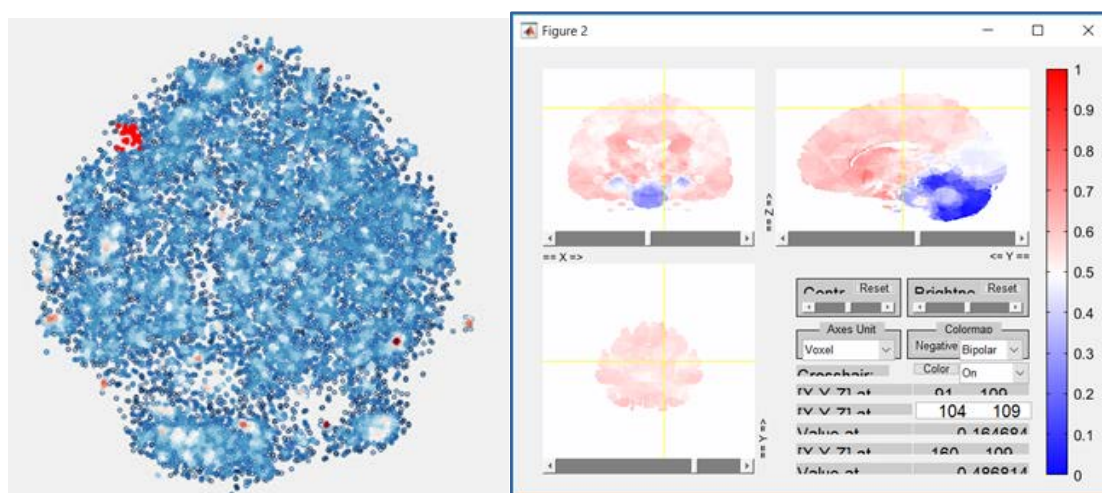


Figure 15: Screenshot from the developed matlab heatmap generation module

The figure above: Screenshot from the developed matlab heatmap generation module for gene co-expression. The tool builds on the dual-tSNE framework recently launched by Huisman et al.¹. The left shows a t-SNE map of the whole genome, where densities represent concentrations of genes with a similar transcriptional profile throughout the brain (from the Adult Allen Human). The right map shows the corresponding 3D average gene expression heatmap of this cluster of co-expressing genes, in nifti format.

In addition, we have been working on a module for the reverse route: given a heatmap from the structural or functional imaging data in the form of a voxel mask, this module identifies which genes are differentially expressed genes for these voxels.²

4.15 T8.3.11 - Brain Morphological Features

4.15.1 Key Personnel

Task Leader: John ASHBURNER (UCL)

¹ Huisman SMH, van Lew B, Mahfouz A, Pezzotti N, Höllt T, Michielsen L, Vilanova A, Reinders MIJT, Lelieveldt BPF. "BrainScope: interactive visual exploration of the spatial and temporal human brain transcriptome", doi: 10.1093/nar/gkx046, Nucleic Acids Research, February 2017

² Doorenweerd N, Mahfouz A, van Putten M, Kaliyaperumal R, t Hoen P, Hendriksen J, Aartsma-Rus A, Verschuuren J, Niks E, Reinders, Kan H, Lelieveldt BPF. "Timing and localization of human dystrophin isoform expression provide insights into the cognitive phenotype of Duchenne muscular dystrophy", Scientific Reports, pending final acceptance.

4.15.2 SGA1 DoA Goals

This Task aims to develop a privacy preserving approach for the generalised principal component analysis of large image datasets. The goal is to maximise the amount of anatomical variability captured, using as few components as possible. Anatomical variability will be encoded by principal modes of both shape and appearance. These component data can be shared across sites, as they do not reveal anything about individuals. Reconstructing each individual's anatomy also involves a set of subject-specific coefficients. These coefficients will serve as features for data mining, and will remain secure within each hospital site. This work will involve close collaboration with SP2.

4.15.3 Component Progress

4.15.3.1 DATA > Hospital Data > Shape and appearance models for human brain variability

Description: Develop a privacy preserving approach for a generalised principal Component analysis of large image datasets, which maximises the amount of anatomical variability captured using as few Components as possible. Part of this work is shared with SP2.

Progress: There is an initial MATLAB implementation of an algorithm for factorising large datasets of images. Development involved using a mixture of MATLAB code, and C code in the form of mex files for parts that are difficult to vectorise.

A version of the purely shape-based model (using principal geodesic analysis) has been developed for 3D data and implemented in such a way as to demonstrate that it can function within a privacy-preserving framework (using MATLAB's distributed computing toolbox). This has been tested with a dataset of 580 publicly available scans (IXI).

Ongoing work is on refining the combined shape and appearance modelling aspects. Until recently, this work has been applied to only 2D images, with both a Gaussian noise model of appearance (for regular images, as in the illustrations below), as well as a binomial noise model (for binary images of tissue classes). The binomial model has been tested on the MNIST dataset, which is an established dataset of hand-written digits used by the machine learning community. Model accuracy was assessed in terms of how accurately the digits can be recognised. For the full training set (~6,000 examples per digit), accuracy of around 99.2% was achieved, which is not quite as good as the best deep learning approaches. For smaller training sets (100 examples per digit), the accuracy was around 98.6%, which is better than achieved by deep learning.

The Gaussian noise model was extended so that multiple image contrasts can be handled simultaneously (cf. grayscale versus colour images). A multinomial noise model has recently been developed, which would allow combinations of tissue classes to be fitted (e.g. grey and white matter in the brain, as well as CSF and other tissues). Example fits may be seen at <http://www.fil.ion.ucl.ac.uk/~john/pga/ImageFactorisation.html>.

Current work is on implementing the combined shape and appearance model in 3D, which would be applicable within a privacy-preserving framework using compiled MATLAB code. There is an emphasis on handling missing data, as the field of view of hospital scans does not always cover the entire brain.



A sample of the face images from the Olivetti Research Laboratory dataset of 400 images.



The shape and appearance model automatically fit to the Olivetti dataset. Each reconstruction is encoded by 40 latent variables that describe both shape and appearance (see below). Those 40 variables per image provide a parsimonious encoding of the original data, which are suitable for applying multivariate privacy-preserving data mining techniques to.



The shape model part of the fit to the Olivetti dataset. This shows the mean image warped to match the original data. Note that no manually defined landmarks were involved, and the shape model was learned automatically.



The appearance model part of the fit to the Olivetti dataset.

Figure 16: Set and variability modelling on a dataset example.

4.16 T8.3.12 - Genetic, Proteins and Neurological Features

- Task name to be changed to "T8.3.12 - Disease progression model" (SGA1 amendment).
- Task goal to be changed to description below (SGA1 amendment).
- Task Leader to be shifted from Alexis BRICE (ICM) to Olivier COLLIOT (ICM).

4.16.1 Key Personnel

Task Leader: Olivier COLLIOT (ICM)



4.16.2 SGA1 DoA Goals

It is important to note that the Task and Task Leader have changed since the SGA1 proposal, as was explained in the SGA1 amendment. This Task now aims to develop and implement a method to construct disease progression models from longitudinal biomarkers. The method will use a statistical learning technique to infer a long-term disease progression model from multiple short term data from a series of individuals. This Task will contribute to brain disease research by providing tools to model disease progression from longitudinal sets of biomarkers. Digital models of disease progression will be used to identify disease signatures.

4.16.3 Component Progress

4.16.3.1 SOFTWARE > Machine Learning Library > Machine Learning > Tool to build disease progression models from scalar measurement

Description: The Component developed is a tool to build disease progression models from scalar longitudinal measurements (cognitive, image-derived biomarkers, clinical...). A built model accounts for variability in age at disease onset, pace of disease progression and trajectories of biomarkers changes across individuals in the observed population. The modelling is based on a generic approach proposed by ICM and described in Schiratti's 2015 article.³

The code is stored in the private GitLab, used by the ICM team: <https://gitlab.icm-institute.org/aramislab/longitudina/> .

Progress: Until now, work has been focused on developing a C++ version of the tool that is to be implemented in the MIP. The current version is functional and has been successfully tested on Alzheimer's disease data from the ADNI database. The code is currently being refined and optimised in order to be more robust and stable, and has yet to be tested with larger datasets. Subsequent steps are: creation of a docker image in the federation, adaptation of inputs and outputs, testing with MIP data.

³ Schiratti J-B, Allasonniere S, Colliot O, Durrleman S. Learning spatiotemporal trajectories from manifold-valued longitudinal data. In Advances in Neural Information Processing Systems pp. 2395-2403, 2015.



5. WP8.4 - Theory, Disease Models & Big Data Engineering

5.1 Key Personnel

Work Package Leader: Ferath KHERIF (CHUV)

5.2 WP Leader's Overview

Contribution to clinical neuroscience and clinical benefit: This WP made significant progress in developing neuroimaging and bioinformatics analytical methods to produce "Disease Signatures" in AD, PD and Autism. The Task Leaders identified the main Components that explain the organisation of the neural circuits underlying normal and adaptive behaviours in a genetically homogenous population. They identified latent variables that can be used in clinical practice to score disease severity in AD patients. The scoring uses MRI data only, however the latent variables correlated with all other measures of severity used in clinics. The Task Leaders demonstrated that latent variables identified in one population can be used in other population.

Contribution to the deployment: Under the Fast Track plan, the Task Leaders contributed towards accelerating the deployment and use of the MIP by clinicians working in the Data Governance and Data Selection (DGDS) Committee to provide, at an early stage, recommendations concerning the hypotheses and selection criteria for the data (e.g. selecting the modality and resolution of neuroimaging data to ensure best data quality).

The Fast Track plan impacted the Tasks in this Work Package, as the significant work had to be done on the deployment. However, the impact on Task progression was low.

Contribution to the platform: The WP contributed to the specification of the algorithms developed during this period which are part of the components SOFTWARE > Algorithm Library. The Algorithm Library is used by the platform developers (WP8.1 and WP8.5) to integrate data mining tools in the MIP. The latent variables for scoring disease progression will be used in the MIP local in the first three hospitals.

This Work Package is critical in providing the first demonstration for the deep brain scale disease modelling. The Task Leaders led the "Disease Related Neuroscience" meeting with other SPs and discussions led to the identification of use cases that can be implemented directly in SGA1, and others into SGA2.

5.3 Priorities for the remainder of the phase

The priority for the next phase is centred on preparing the ground for disease signature identification based on hospital data using neuroscience priors. The Work Package will develop the strategy and roadmap for disease modelling in collaboration with the other SPs but also medical initiatives and pharma.

5.4 Milestones

Table 4: Milestones for WP8.4 - Theory, Disease Models & Big Data Engineering

MS No.	Milestone Name	Leader	Task(s) involved	Expected Month	Achieved Month	Comments
MS12	MS 8.4.1 Initial Implementation Plan for Model Base Algorithms Concepts and Application Formulated - Integration plan to the BRAIN software	CHUV	All tasks	M02	M02	WP8.4 contributed to the revised MIP Platform Fast-Track Plan for the work to be carried out by the SP8 team during SGA1.
MS114	MS 8.4.3 Case Study 1 integrated into MIP	CHUV	T8.4.3	M12	M10	Papers demonstrating the results have been submitted
MS115	MS 8.4.2 Initial proof-of-concept and results of the different algorithms	CHUV	T8.4.4	M12	M12	Papers demonstrating the results have been submitted

5.5 T8.4.1 - Brain Scale High Performance Deep Phenotyping

5.5.1 *Key Personnel*

Task Leader: Andrew POCKLINGTON (CF)

Other Researcher: Bogdan DRAGANSKI (CHUV)

5.5.2 *SGA1 DoA Goals*

The goal of this Task is to understand the organisation of the neural circuits underlying normal and adaptive behaviours in a genetically homogenous population with high risk for brain disorders to enable differentiation from normal behaviour. Specifically, we will:

- 1) Define typical trajectories for development-related brain structure changes (anatomical connectivity, cortical thickness, grey matter volume).
- 2) Identify the main components of inter-individual variability in brain anatomy features to accurately estimate differences in dementia, Parkinson's disease and epilepsy;
- 3) Investigate the influence of disease-linked mutations on brain function and organisation.

5.5.3 *Component Progress*

5.5.3.1 MIP - DATA > Reference Data > Functional gene annotation

Description: Captures biological properties assigned to genes. Each annotation consists of a name plus a list of genes (specified by human NCBI/Entrez gene ID and gene symbol).

Progress: Scripts have been written to process the MGI (Mammalian Phenotype ontology) annotation data (re-construct an ontology tree; extract and filter gene phenotype terms; populate parent phenotype terms; and map genes from mouse to human). Once the performance of these scripts is optimised, they will be adapted to process the GO annotation ontologies. The final bundle of scripts will be set up to automatically generate updated MGI & GO annotation gene-sets on a monthly basis.

Contribution:

- CF team - The research associate who will perform the work in this component was appointed early January 2017 (~6 months were lost due to the SGA1 start date being backdated). Work is progressing as rapidly as possible - and we expect the planned work for both components to be completed by the end of SGA1.

5.5.3.2 MODELS > Biological Signature of Diseases > *Disease associated gene sets*

Description: Functional gene sets enriched for genetic variants associated with complex brain disorders.

Progress:

- Definition of a set of copy number variations (CNVs - e.g. 16p11.2, 1q22.1) associated with schizophrenia and autism.
- Provided MRI and behavioural data from open databases (SFARI - Simon's VIP) and hospital data.
- Awaiting the annotation gene-sets to be generated by the previous component.

Contributions:

- CHUV SP8 team - see above.
- CF team - The research associate who will perform the work covered by this Component was appointed early January 2017 (~6 months were lost due to the SGA1 start date being

backdated). Work is now progressing as rapidly as possible - the planned work for both of these Components should still be completed by the end of SGA1.

5.5.3.3 MODELS > Biological Signature of Diseases > Alzheimer's Disease

Description: Definition of typical trajectories in Alzheimer's disease - related brain structure changes (anatomical connectivity, cortical thickness, grey matter volume) and differentiation of these from trajectories seen in healthy ageing. Identification of the main components of inter-individual variability in brain anatomy features to more accurately estimate differences in Alzheimer's disease.

Progress: First biological signatures of AD created.

Releases: Atlas of AD brain features created.

Contributions:

- CHUV SP8 team

5.5.3.4 MODELS > Biological Signature of Diseases > Epilepsy

Description: Definition of typical trajectories for epilepsy-related brain structure changes (anatomical connectivity, cortical thickness, grey matter volume) and differentiation from trajectories of healthy ageing. Identification of the main components of inter-individual variability in brain anatomy features to accurately estimate differences in epilepsy.

Progress: First biological signatures of temporal lobe epilepsy (TLE) created.

Releases: Atlas of TLE brain features created.

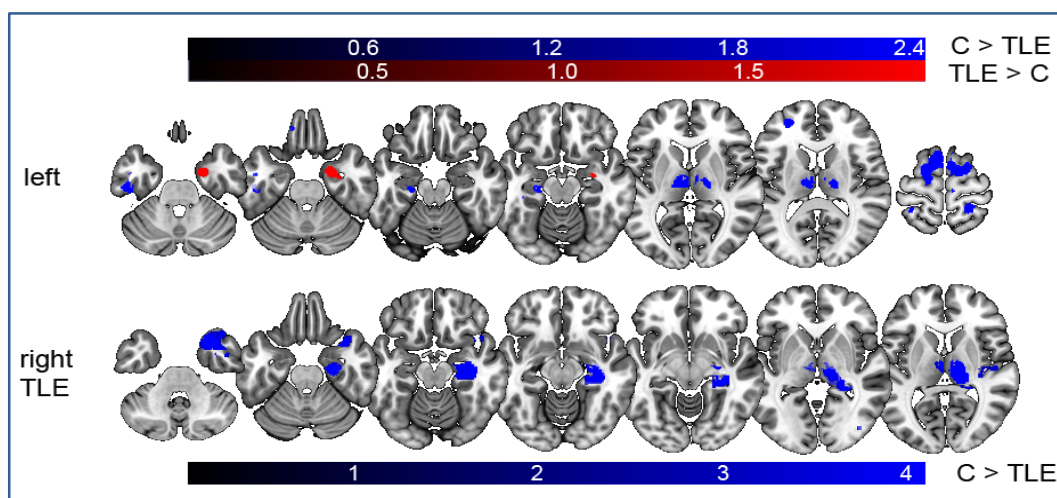


Figure 17: Atlas of TLE brain features

In the figure above : Atlas of left lateralised TLE [top] and right lateralised TLE [bottom]. RED - increases in grey matter volume, BLUE - decreases of grey matter volume.

Contributions:

- CHUV SP8 team

5.5.3.5 MODELS > Biological Signature of Diseases > Parkinson Disease

Description: Definition of abnormal trajectories for Parkinson disease-related brain structure changes (anatomical connectivity, cortical thickness, grey matter volume, tissue properties) and differentiation from trajectories of healthy ageing.

Progress: First biological signatures of idiopathic Parkinson's disease (PD) created

Releases: Atlas of PD brain features created

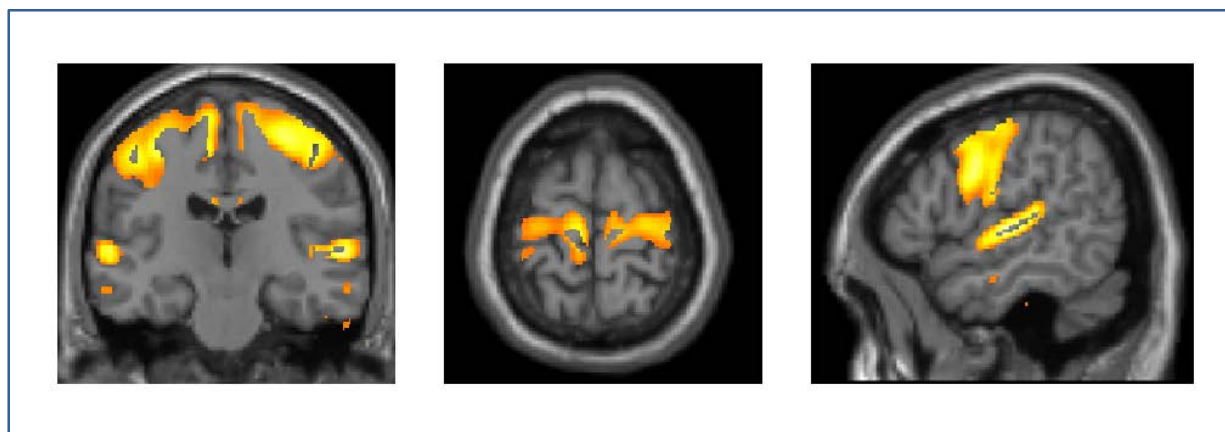


Figure 18: Atlas of PD brain features

Atlas of grey matter volume loss in early PD (based on analysis of >400PD patients compared with 300 healthy controls)

Contributions:

- CHUV SP8 team

5.5.3.6 MODELS > Biological Signature of Diseases > Healthy Aging

Description: Definition of typical trajectories for development-related brain structure changes (anatomical connectivity, cortical thickness, grey matter volume, brain tissue properties and differentiation of these from abnormal trajectories Identification of the main components of inter-individual variability in brain anatomy features to accurately estimate differences in brain disorders.

Progress: We provide a model of healthy brain ageing based on high-resolution MRI data obtained from >500 healthy individuals following linear and non-linear trajectories of healthy ageing.

Releases: Atlas of healthy ageing brain features created

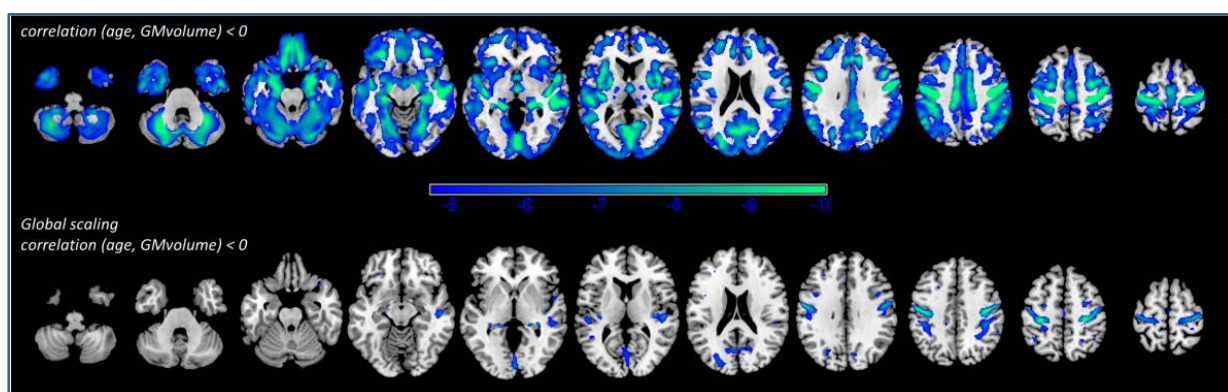


Figure 19: Atlas of healthy ageing

Atlas of grey matter volume loss across the lifespan (based on analysis of >460 healthy controls)

Contributions:

- CHUV SP8 team

5.6 T8.4.2 - Brain Scale Disease Bayes Modelling

5.6.1 Key Personnel

Task Leader: Ferath KHERIF (CHUV)

5.6.2 SGA1 DoA Goals

The goals of this Task are

- 1) To develop, implement and deploy mathematical methods for predicting multi-level features of diseases.
- 2) To develop tools for identification of homogeneous disease using the Biological signatures and construct unified models of brain diseases.

5.6.3 Component Progress

5.6.3.1 MODELS > Biological Signature of Diseases > Construct unified models of brain diseases neurological and psychiatric

Description: Build, test and validate Bayesian Model that deals with the problem of mapping thousands of multivariate variables to a small dimensional space. The proposed model has a direct impact in identifying anatomical signatures of brain in both normal and pathological states using big multimodal data from different databases. The predictive ability of the model can be tested in an unbiased way by applying the model to a new independent data sets.

Progress: During the SGA1 phase until M06 the CHUV team focused on the development of Bayesian and multivariate models that take into account multi-modal data from both an open data source (ADNI initiative) and clinical data (CHUV). The model is based on two Components described below. It allows to predict disease severity using novel datasets. In the next step is to assess the generalisability of the model to new types of diseases e.g. Parkinson's disease.

We applied the developed method to test whether morbid personality traits and neuropsychological/psychiatric assessments (A) relate to individual differences within the MTL (B) independent of cognitive state. To further improve the discrimination between two groups (Mild cognitive impairment and healthy) in term of brain anatomical changes. Our results supported the hypothesis that personality traits can alter the vulnerability and pathoplasticity of disease and therefore modulate related biomarker expression.

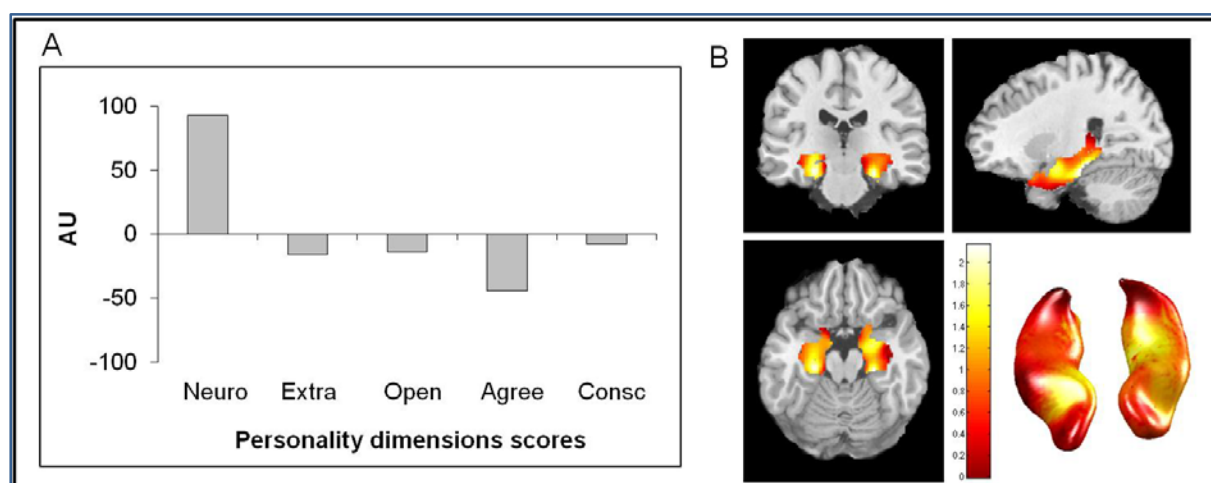


Figure 20: Association between cognitive decline, personality traits and biological changes in the Medial Temporal Lobe (MTL).

We found a specific association between cognitive decline, personality traits and biological changes in the Medial Temporal Lobe (MTL). Interestingly from disease modelling point of view, the association was explained by a low dimensional factor which corresponded to a personality profile dominated by neuroticism trait associated with a spatial gradient along the MTL.

5.6.3.2 SOFTWARE > Algorithm Library > Bayesian methods and deep learning tools for identification of homogeneous disease using the Biological signatures

Description: In this work, T8.4.2 will build Bayesian Models to identify the optimal subset of biological signatures that would help clinicians diagnose and categorise stages of brain diseases.

Progress: The latent variable algorithm using Maximum likelihood estimation on brain GM volume estimates was created to predict disease severity. The extracted latent trait was compared to baseline clinical diagnosis and disease conversion as well as to proteomic and metabolic biomarkers. The estimated latent trait showed significant difference between the clinical groups and correlation with memory and cognitive performance tests. The estimated latent trait significantly associated with clinical conversion: cognitive normal (CN) to mild cognitive impairment (MCI) and MCI to AD. In addition, the latent trait significantly correlated with Tau, A β , pTau and A β /Tau proteins levels in the CSF, cortical A β burden metabolic glucose reductions. Our results showed evidence in two independent cohorts that the MRI-derived latent trait is significantly correlated to baseline clinical diagnosis, disease conversion and disease-related biomarkers. These characteristics suggest that the latent MRI trait could serve as a surrogate to quantify disease staging in the clinical practice.

5.6.3.3 SOFTWARE > Algorithm Library > mathematical methods for predicting multi-level features of diseases

Description: The motivation for this methodological development comes from the intention to integrate computational modelling into the discovery of disease signatures. A Bayesian Model is proposed as framework to decode the organisational principles of the brain anatomy.

Progress: We developed a Bayesian methods based Factor analyses to identify latent variable. In this study, we propose to extract a latent variable based on MRI-measured atrophy patterns to quantify the disease severity for each subject by applying item response theory (IRT) The main advantage of this method is that it does not completely rely on the supervision of clinical diagnosis to identify disease-related regions. The software was made available for integration into the Algorithm Factory.

5.7 T8.4.3 - Tools for Macro- to Micro-Scale Data Analysis and Atlasing

5.7.1 Key Personnel

Task Leader: Antoine LUTTI (CHUV)

5.7.2 SGA1 DoA Goals

The goal of this Task is to develop models of MRI data for specific characterisation of tissue microstructure *in vivo*. Particular emphasis will be placed on implementation, allowing direct use of the models on *in vivo* data.

5.7.3 Component Progress

5.7.3.1 MIP - DATA > Reference Data > Normative values from qMRI data

Description: This Component will deliver normative values of reference qMRI data.

Progress: This Component follows from SOFTWARE>Algorithm library>Brain Anatomy>Quantification of tissue properties from qMRI (see below). Using the developed



software and provided by the latter component, it will compute qMRI data from raw MRI images from a cohort of >1,000 datasets to provide normative reference qMRI values. This is currently being implemented in the data factory. The normative values will be made available to the MIP users for reference.

5.7.3.2 SOFTWARE > Algorithm library > Brain Anatomy > Quantification of tissue properties from qMRI

Description: This Component will deliver methods for specific characterisation of tissue microstructure *in vivo* from qMRI data (e.g. biomarkers of myelin and iron concentrations and axonal g-ratio).

Progress: Over the 12 months, extensive developments was made by a senior scientist to the software that extracts qMRI biomarkers of the brain from raw MRI data. Work on gratio estimates has been halted due to the specificities of the processing involved. Instead, the work has focused on relaxometry data that provides *in vivo* biomarkers of water, myelin and iron concentration. The improvements include optimisation of the accuracy of the qMRI maps and minimisation of systematic bias between the qMRI maps (see figure below). Objective measures of data quality were developed (homogeneity of the qMRI values within brain tissue, patient motion parameters). The new features of the software were developed by visual and quantitative assessment on a restricted cohort of ~30 datasets. The new software features were tested quantitatively on a cohort of ~1,000 datasets. Normalisation of the qMRI data was extensively examined to preserve the accuracy of the qMRI data in group space, with particular attention to the accuracy of local Grey Matter regional values. The new software version is currently being implemented in the Data Factory.

- An international workshop of experts in the field was organised by the Task Leader to disseminate the outcome of this Task (13-14 March 2017).
- The developed software is available on the following github repository: <https://github.com/LREN-CHUV/qMRI>

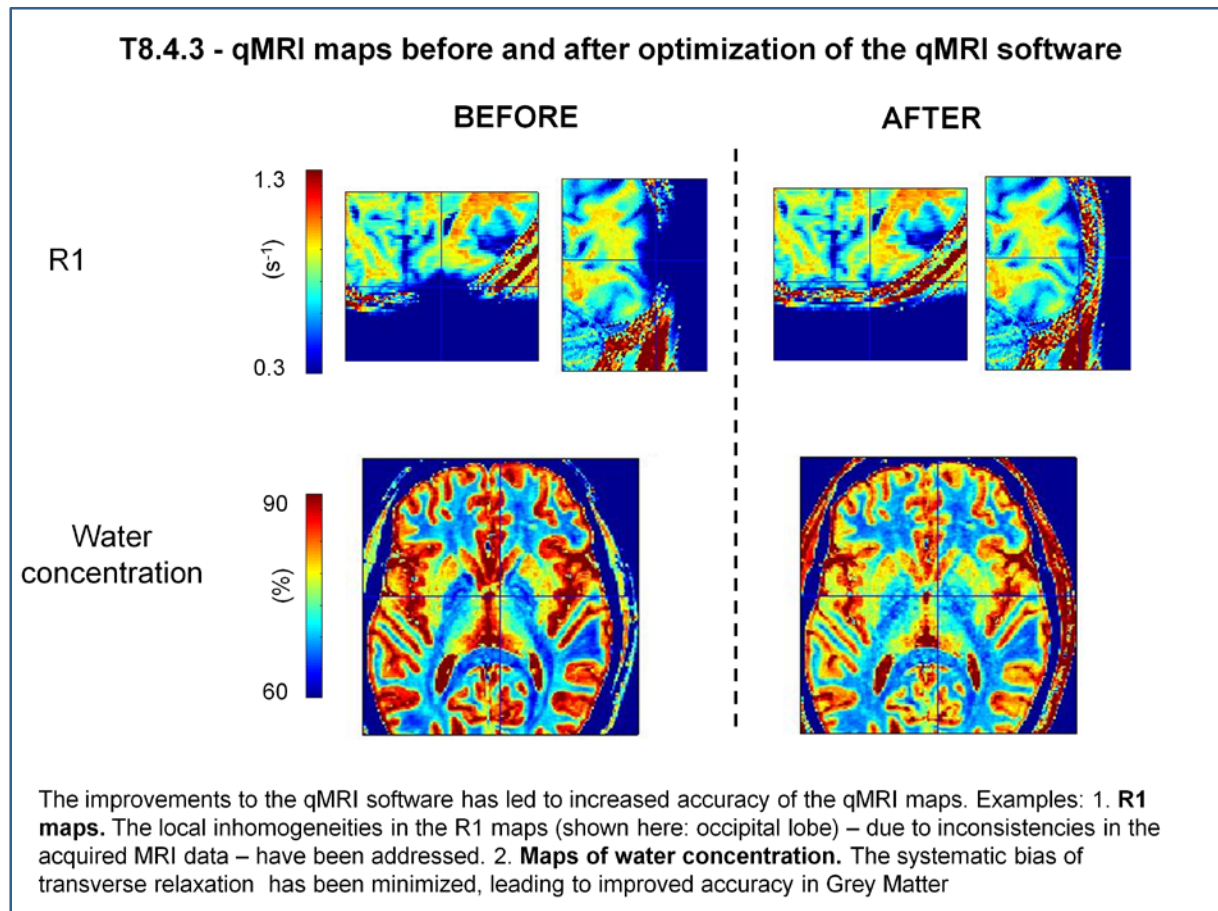


Figure 21: qMRI maps before and after optimisation of qMRI software

5.8 T8.4.4 - Case Studies in Discovering Disease Signatures and Modelling Disease Progression

5.8.1 Key Personnel

Task Leader: Bogdan DRAGANSKI (CHUV)

Other researcher: Saso DZEROSKI (JSI)

Other researcher: Boudewijn LELIEVELDT (LUMC)

Other researcher: Mira MARCUS-KALISH (TAU)

5.8.2 SGA1 DoA Goals

The goals of this Task are

- 1) To replicate results from the literature making full use of the available multi-modal and multi- scale data.
- 2) To enhance diagnostic accuracy through improved brain anatomy feature extraction (interactions with WP8.3);
- 3) To improve the accuracy of an individual's clinical outcome prediction through implementation of the framework for disease modelling (WP8.4) and developments in methods for modelling the temporal dynamics of disease (WP8.3).



5.8.3 Component Progress

5.8.3.1 SERVICES > SP8 Coordination & Project Management Services > Clinical demonstrators

Description: Build a series of applications and user stories with clinicians from different medical domains.

Progress: Applications and uses stories with clinicians focused on the two most frequent neurodegenerative disorders - Alzheimer's and Parkinson's disease.

Contribution:

- JSI - Application of rule-based clustering methods to discover disease signatures for Alzheimer's and Parkinson's diseases
- JSI & TAU - Application of multi-layer clustering in Alzheimer's, two journal papers published.
- JSI & CHUV - Application of predictive clustering to find signatures for Parkinson's disease.
- JSI - Identification of patterns of brain atrophy and their explanation in clinical scores.
- CHUV - Clinical interpretation of first JSI results in Parkinson's disease models (manuscript in preparation).
- CHUV - Detection of cognitive decline in Parkinson's disease (Gee et al., 2017)
- CHUV - Detection of patterns of brain remodelling in TLE (manuscript in preparation)
- CHUV - Detection of patterns of brain remodelling in healthy ageing (two manuscripts in preparation)
- LUMC - Executed a number of clinical case studies on paroxysmal disease (Migraine⁴), Alzheimer's disease⁵ (see publication list[2]) and the cognitive aspects of Duchenne's muscular dystrophy⁶. These studies were done on public imaging cohorts (ADNI) as well as partner-specific cohort studies. In addition, a comprehensive review paper was written and published on how brain transcriptome atlases can be applied to disease signature discovery.⁷ These pilot studies demonstrate that:
 - 1) Early detection of disease signatures from longitudinal MRI data mixed with clinical data is feasible.
 - 2) Spatial gene expression heatmaps as developed in task 8.3.10 are a valuable augmentation to clinical use cases for disease signature discovery.

⁴ Eising E, Huisman SMH, Mahfouz A, Vijfhuizen L, [the International Headache Genetics Consortium], Nyholt D, de Vries B, Lelieveldt BPF, van den Maagdenberg AMJM, Reinders MJT. "Gene co-expression analysis identifies brain regions and cell types involved in migraine pathophysiology: a GWAS-based study using the Allen Human Brain Atlas", Human Genetics, 135 (4), 425-439, 10.1007/s00439-016-1638-x

⁵ Sun Z, van de Giessen M, Lelieveldt BPF, Staring M. "Detection of Conversion from Mild Cognitive Impairment to Alzheimer's Disease Using Longitudinal Brain MRI", Frontiers in Neuroinformatics, <https://doi.org/10.3389/fninf.2017.00016>

⁶ Doorenweerd N, Mahfouz A, van Putten M, Kaliyaperumal R, t Hoen P, Hendriksen J, Aartsma-Rus A, Verschuuren J, Niks E, Reinders, Kan H, Lelieveldt BPF. "Timing and localization of human dystrophin isoform expression provide insights into the cognitive phenotype of Duchenne muscular dystrophy", Scientific Reports, pending final acceptance.

⁷ Mahfouz A, Huisman SMH, Lelieveldt BPF, Reinders MJT. Brain transcriptome atlases: a computational perspective, (2017) 222: 1557, doi:10.1007/s00429-016-1338-2



5.9 T8.4.5 - Large-Scale Data Analytics on Massively Parallel Architecture

5.9.1 *Key Personnel*

Task Leader: Thomas HEINIS (ICL)

Other Researcher: Ferath KHERIF (CHUV)

5.9.2 *SGA1 DoA Goals*

The goals of this Task are to develop and deploy an array of analytics algorithms for large amounts of data on HPC infrastructure provided by SP7.

5.9.3 *Component Progress*

Note regarding ICL contribution: The hired PostDoc should start work soon. The Task Leader has kept Task implementation on track.

5.9.3.1 *Data uploader*

Description: Facility to upload data to HPC resources.

Progress: Progress on this Component was made by ICL and will result in a first release in M16 on infrastructure by ICL and on HPC infrastructure provided by SP7 by M24. CHUV provided representative clinical data.

5.9.3.2 *Data cleaning & formatting*

Description: Tool to translate (and clean) data into format expected by analytics library. The tool will translate data from the most common formats.

Progress: The variety of analytics algorithms increases the workload for this Component but development is on track.

5.9.3.3 *Analytics library*

Description: Library set up on HPC resources for distribute analytics.

Progress: After an initial survey of available implementations for analytics on MPI, we discovered that little work has been done. There are a number of algorithms available but effort to harmonise implementation & documentation is bigger than expected. We are also investigating porting Spark on HPC infrastructure to make a wide variety of analytics algorithms available instantly. Still, the Component is on track to manage release on time.

5.9.3.4 *Data download*

Description: Facility to download data from HPC resources.

Progress: Progress on this component was made and will result in a release soon.



6. WP8.5 - The Medical Informatics Platform

6.1 Key Personnel

Work Package Leader: Ferath KHERIF (CHUV)

6.2 WP Leader's Overview

The CHUV team has been strongly mobilised to deliver key components of the platform on time, in particular the new Data Factory and the deployment scripts for hospitals.

Our open-source contribution metrics really show the amount of effort: 2,000 commits, 16 new projects, a total of 32 projects actively maintained for the period and 50 tasks completed on Trello boards. The metrics for Github commits by CHUV are summarised at <https://dashboard.cauldron.io/goto/3a8908ce0822936025289aecaf8014d7>.

Several tools have been setup to address the communication challenges between developers and contributors to SP8 working in different institutions and geographical locations: including a reorganised [Github repository](#) with a [web site](#) showing the technical sides of the project; instant discussions on Slack for [platform development](#) and [deployment to hospitals](#); Trello boards for [planning platform development](#) and [planning deployment to hospitals](#) (switch to YouTrack planned but not yet complete) as well as continuous integration and quality assurance. A Software Development Committee was put in place to steer tools and communication effort and to increase professionalism in SP8 development. As a result, communication and quality within SP8 has greatly improved and we are now able to demonstrate productive collaborations, such as integration of MipMap from AUEB and PostgresRAW from EPFL into the Platform. Full integration of Exareme from UoA and several more algorithms from TAU, JSI, ICM, LUMC are all Tasks in progress.

A Platform update was released for production in October 2016, with new functionality including cross-validation added to the Algorithm Factory and many usability improvements for Web-based Modelling and Visualisation.

A new Data Factory Component was released in March. It brings new data pipelines for pre-processing clinical MRIs and integration of patient data. Data pipelines for research datasets ADNI, PPMI, EDSD (images+EHR) and clinical datasets from CHUV CLM (images+EHR) reached TRL5 (first data delivery - prototype data) and TRL6 for PPMI (second data delivery - candidate final quality).

The CHUV Platform development team is down to three engineers after the departure of one developer. Recruitment of a web developer is on-going and until the person arrives we can only correct big user issues at the web front-end.

- Major interest and collaboration with TBI/CREACTIVE with a proof-of-concept validated by CREATICE at the International TBI meeting in Washington, October 2016.
- The improved architecture based on building blocks allows to reduce dependencies between elements of the MIP stack. Including additional analysis methods is also simplified.
- The improved architecture allows also for more robust and automated deployment activities in various environments. The development and deployment activities between Local and Federated parts are decoupled.
- The deployment possible thanks to the MIP-Local allows to be closer to the real-world patient data and thus better ensure the important mission of collecting data from various sources.

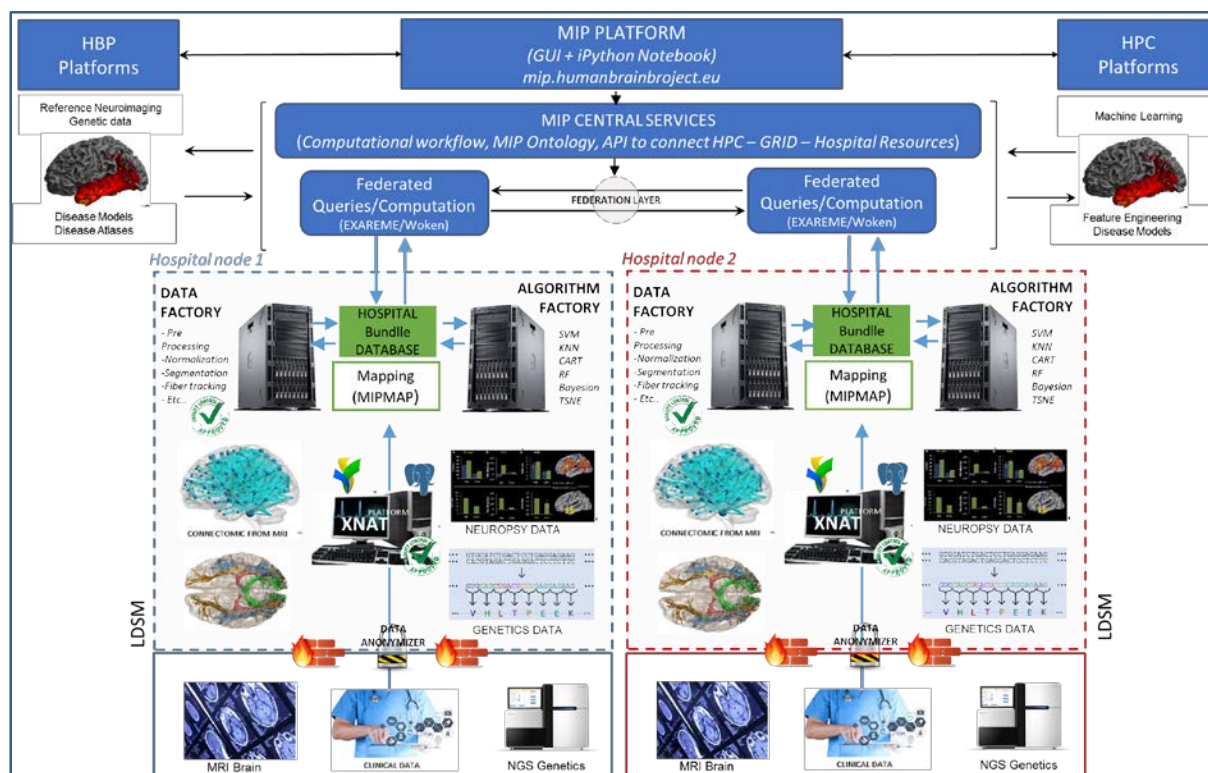


Figure 22: MIP overall architecture

6.3 Priorities for the remainder of the phase

We need to complete integration of the ETL tool MipMap in the Data Factory in collaboration with the T8.1.4.

After MIP-Local deployment the development efforts will focus on Federating hospitals. This includes:

- 1) Completing integration of Exareme into the platform with help from T8.1.5 to enable distributed queries.
- 2) Work on providing and securing network communications between the hospitals and the Federation with help from T8.1.1
- 3) Improve the Algorithm Factory to support distributed data analytics or machine learning workflows, provide a catalogue of algorithms for distributed datasets to the Algorithm Library with help from T8.3, T8.4

Further on, the functionality of the Platform needs to be improved in the following areas:

- 1) User interface and Web portal: the usability of the Platform and the functionality needs to better highlight the hard work of clinical data collection that has been performed so far by the team and hence the rich data available for clinical research questions.
- 2) Deeper integration with Collaboratory and SP5 Knowledge Based services.
- 3) Industrialise packaging of the Platform to prepare for wider deployment and usage in SGA2.



6.4 Milestones

Table 5: Milestones for WP8.5 - The Medical Informatics Platform

MS No.	Milestone Name	Leader	Task(s) involved	Expected Month	Achieved Month	Comments
MS13	MS 8.5.1 Project Implementation & Delivery Plan, WP level - Integration plan to the BRAIN software	CHUV	T8.5.3	M02	M02	Contribution to the revised MIP Platform and the Fast-Track Plan for the work to be carried out by the SP8 team during SGA1. Define new Components for each Task of the WP. Details can be found in D8.6.1.
	MS 8.5.2 Web Portal v2, Software Foundation & Microservices v1, Algorithm Factory v1, Insight Factory v1, Web Portal v2, Algorithm Factory v1 and Insight Factory v1	CHUV	T8.5.31, T8.5.2	M06	M08	<p>All functionality & Components planned have been 100% successfully completed.</p> <p>Software Foundation & Microservices v1: Released in October. Development of V1 complete, unit and functional testing complete, UAT complete.</p> <p>Algorithm Factory v1 and Algorithm Library v1: Released in October. Development of V1 complete, unit and functional testing complete, UAT delayed by one month to match the release of Web Portal v2.</p> <p>Insight Factory has been integrated into Web Portal and Algorithm Factory.</p> <p>Web Portal v2: Development was almost complete in October, but due to increased workload, its release has been effected in November 2016.</p> <p>This release completed the release v2 of the Medical Informatics Platform, including usability improvements and new experiments for brain imaging.</p> <p>Link : http://mip.humanbrainproject.eu/</p>
MS116	MS 8.5.3 Release platform usage tool (database, data, reports) Data Factory v1 released	CHUV	T8.5.3	M12	M12	Release platform usage tool - Tool has been created using Google Analytics and is operational.



						<p>Data Factory v1 - Data Factory v1 has been released. Data Factory data pipelines have produced first results on 4 datasets containing MRI scans.</p> <p>Links:</p> <ul style="list-style-type: none">- https://hbpmedical.github.io/specifications/data-factory/- https://github.com/HBPMedical/mip-microservices-infrastructure/tree/master/demo/data-factory/airflow
--	--	--	--	--	--	---

6.5 T8.5.1 - Web-Based Medical Data Analyses Foundation

6.5.1 Key Personnel

Task Leader: Ferath KHERIF (CHUV)

Other Researcher: Giovanni FRISONI (UNIGE)

6.5.2 SGA1 DoA Goals

The goal of this task is to provide the front-end environment for operating the Algorithm Factory, Data Factory and Web Analytics:

- 1) MIP Web Portal, where users can interrogate and analyse MIP data. (CHUV)
- 2) Further development of the MIP Knowledge Base. (UNIGE)

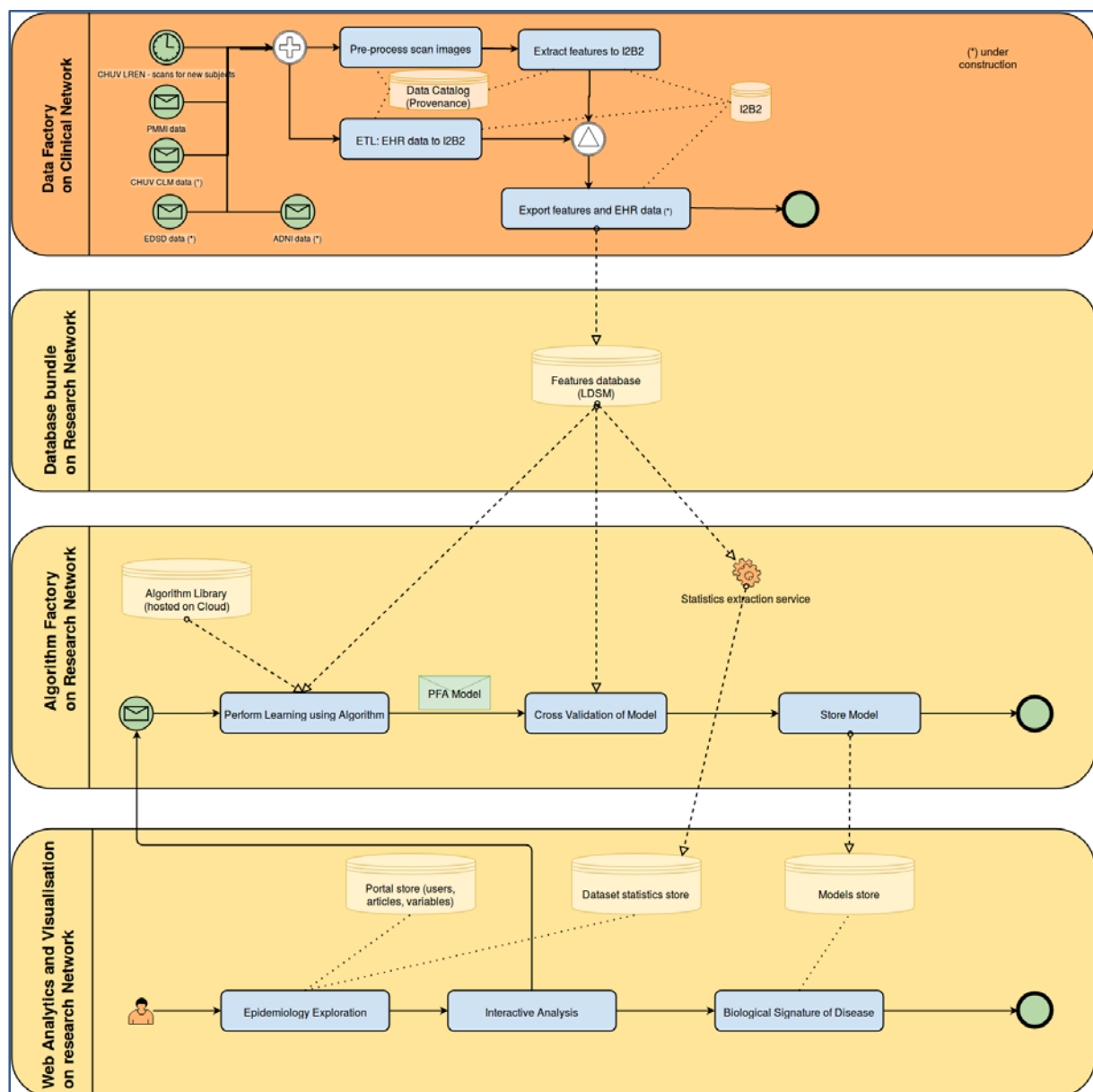


Figure 23: MIP-Local functional architecture showing Data Factory, Database bundle, Algorithm Factory and Web Analytics.

6.5.3 Component Progress

6.5.3.1 MIP - SOFTWARE > Web-based Modeling and Visualisation

Description: A 'building block' of MIP. The Web-based Modelling and Visualisation provides a web portal for end-users of the MIP, integration of the web portal with the Collaboratory, access to Jupiter notebooks from the Collaboratory providing functions from MIP and allowing users to explore (statistics only for privacy reasons) and analyse data coming from hospitals and research centres connected to MIP.

Progress: From the initial Platform release at the end of the RUP, CHUV has developed and internally tested multiple versions of the web portal which currently represent the main entry-point for the end-user of the Platform. After several iterations, version 2.0.0, was publicly released on 14 November 2016. This version includes many additional features which are listed here by their corresponding sub-components:

- EE > Variable Exploration, Categorization & Viewer
 - User interface has been enhanced based on internal user acceptance testing feedback to provide better user experience.
 - Decoupling and refactoring to allow quick updates of the variable groups hierarchy and prepare end-user customisation capability.
- EE > Descriptive Stats generator & Viewer
 - Complete rewriting of the generator as a container included in the Algorithm Library and launched by the Algorithm Factory.
 - On-the-fly generation of the descriptive statistics and caching mechanism to handle any future data import in an automated way.
 - Integration with the new meta-data bank and display of new fields related to variables such as methodology, units, type...
- Model Configuration > Variable selection
 - Added support for batch selection and removal using variable's groups.
 - Clarification of model's co-variable types as either continuous or nominal. Grouping property is not relevant anymore at this stage.
 - Various general usability improvements.
- Interactive Analysis
 - Various general performance and usability improvements.
- Experiment Builder
 - User interface that allows interactive experiment launching consisting of supervised/unsupervised training of models and cross-validation (in supervised settings), which represents a new feature of the Algorithm Factory.
 - Background user notification system for running experiments.
 - Display of experiment cross-validation results as bar charts and confusion matrices.
 - Persistency of experiment results.
- Information and Scientific references > Research Object wrapper
 - Display of human friendly PFA strings produced by experiments on a result page.
- Image & Genetic Viewer



- Added 2D biological rules viewer.
- Updated 2D DICOM/NIFTI viewer.
- Refactoring to prepare transition towards new version of the external apps toolkit.
- External Apps
 - Refactoring to prepare transition towards new version of the toolkit.

Planned features:

- EE > Descriptive Stats generator & Viewer
 - Display of provenance information on the variables.
- Model Configuration > Variable selection
 - Allow selection of multiple variables to reflect Algorithm Factory's new capability.
- Model Configuration > Variable filtering
 - Development, integration, testing to be finished.
 - Add support for temporal axis.
- Experiment Builder
 - Add new validation schemes based on data provenance information.
 - Add support for longitudinal experiments.
- External Apps
 - Transition to a more powerful toolkit allowing integration of web technologies (both backend and frontend) based on any environment and using containers in a similar manner to how Algorithm Factory allows algorithms.
- Collaboratory Integration
 - Finalise the integration of the Platform with the Collaboratory that has been initiated and will allow advanced users to call web portal backend and Algorithm Factory's services through Jupyter notebooks.

Links:

- The latest web portal version is accessible publicly at this address:
<https://mip.humanbrainproject.eu>
- Sources are publicly available in the two following Github repositories:
<https://github.com/HBPMedical/portal-backend>
- <https://github.com/HBPMedical/portal-frontend>

Quality Control:

Upstream:

- MIP - SOFTWARE > Algorithm Factory (software) - T8.5.2
 - Intermediate release
 - Integration successfully achieved
- MIP - DATA > MDR (Meta Data Register) (data) - T8.5.2
 - Intermediate release
 - Enough meta-data has been provided to allow further development



6.5.3.2 SERVICES > Security & Monitoring > User Management service

Description: Management of users, integration with Collab user authentication service, ACL and logging of user activities (EE, IA, Model building, article sharing).

Progress: CHUV delivered the Management of users and integration with Collab user authentication service. ACL and logging of user activities are pending.

Quality Control:

Upstream

- Collaboratory Service (service) - T11.3.2
 - Finished component
 - Working well, good documentation, responsive and helpful support

Downstream:

- *SOFTWARE > HDB > Access Right Module* - T8.1.4
 - Only a PDF containing some specifications has been provided
 - No feedback received

6.5.3.3 SERVICES > Security & Monitoring > Platform Usage Monitoring

Description: Monitor usage of the Platform and its services. Keep an audit log of all user actions and take preventive measures in case abusive usage of the Platform is detected. Provide web analytics dashboard about users coming to the site.

Progress: CHUV has setup a Google Analytics dashboard to track users of MIP and their behaviour. CHUV also started work on a solution to log backend calls using an Elasticsearch/Logstash/Kibana stack.

6.5.3.4 SOFTWARE > Web Exploration and Analytics > Knowledge Base Application

Description: Knowledge base for the Medical Informatics Platform

Progress: UNIGE has decommissioned the old Knowledge Base instance based on LifeRay and deployed a new static website created by a static website engine (HUGO) augmented by dynamic features served by a classic PHP + MySQL solution.

Missing/planned features: Simplify the dynamic part of the website by replacing PHP/MySQL with Javascript/GraphQL technologies.

6.5.3.5 SERVICES > Upgrade - Deploy - Release > CI (Continuous Integration) workflow

Description: Tools and workflows for Continuous Integration of the software.

Progress: CHUV setup continuous integration for 11 projects on CircleCI.com, 1 project on Wercker.com, and continuous code quality checks for 10 projects on Codacy.com. CHUV also deployed a complete QA environment for testing purposes.

Links:

- For a live dashboard: <https://hbpmedical.github.io/software-catalog/>
- QA version of the Platform can be accessed using standard HBP login at <https://hbps1.chuv.ch/>.

6.6 T8.5.2 - Web API and Microservice Architecture for Community-Driven Data Analyses and Workflows



6.6.1 *Key Personnel*

Task Leader: Ferath KHERIF (CHUV)

6.6.2 *SGA1 DoA Goals*

The goal of this Task is to deliver a generic and modular Foundation & Microservices Platform that will enable the construction of the Algorithm Factory and Data Factory.

6.6.3 *Component Progress*

6.6.3.1 MIP - DATA > Reference Data > Features

Description: Database of features extracted from a Reference dataset using the Data Factory.

Progress: CHUV provided a prototype database containing reference data from ADNI. Work has started to include features from other datasets (PPMI, ESDS) but is not yet complete.

Quality Control:

Upstream:

- SOFTWARE > Data Factory: T8.5.2, intermediate release, missing extraction of features from I2B2 database

Downstream:

- MIP - SOFTWARE > Algorithm Factory: T8.5.2, provided the prototype database

6.6.3.2 SOFTWARE > Data Factory

Description: A 'building block' of MIP. The main role of the Data Factory is to perform offline pre-processing, feature extraction and transformation (ETL) of data captured by the 'Data Capture' building block. Features are normalised and exported to the 'Hospital Database' building block. Several image pre-processing pipelines for -omics data are provided and a flexible workflow engine that can use Docker containers allows customisation of the factory. ETL tools and processes perform the transformation and normalisation of the diverse data captured from hospitals or research datasets. Finally, various storage databases track all processing to build provenance information, provide new integration points with I2B2 and provide safe storage for neuroimaging data (XNAT).

Progress:

CHUV provided version 1.0 of the Data Factory with the following features:

- A generic workflow engine (Apache Airflow) maintained by a large open-source community, where CHUV contributed by providing automated deployment scripts and monitoring of the service;
- A set of pipelines (DAGs in Airflow terminology) for image pre-processing and ETL tasks. The pipelines are configurable and can process various datasets with very different organisation of data, for example research and clinical datasets, structural MRI data in DICOM or Nifti formats;
- Functionality to track provenance and transformations of datasets.
- Features computed by the image pre-processing pipelines and variables harvested from EHR data are imported to an I2B2 database, making further data sharing and import of other I2B2 datasets easier.

CHUV has achieved the following level of maturity for the data pipelines:

- ADNI pre-processing: TRL5, image pre-processing performed on a small subset of ADNI

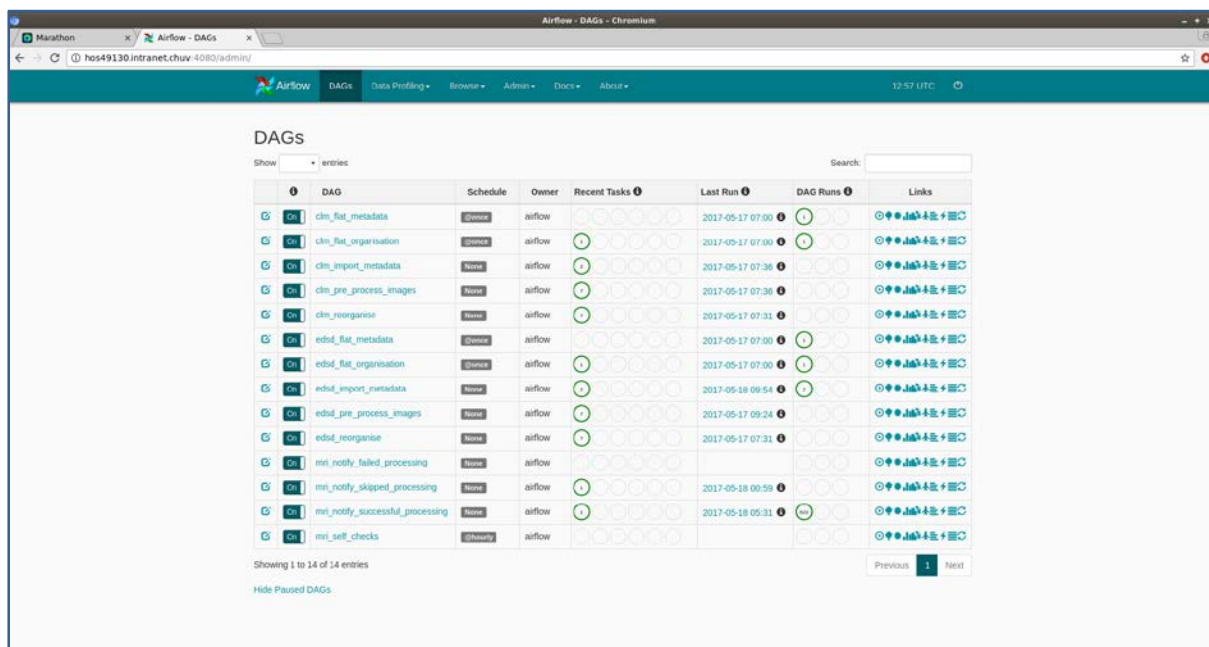
- PPMI pre-processing: TRL5, image pre-processing performed on full PPMI imaging data (720 subjects).
- EDSD pre-processing: TRL5, image pre-processing performed on a small subset of EDSD
- CHUV CLM pre-processing: TRL5, image pre-processing performed on full CLM imaging data (185 patients).

Missing/planned features:

- Export of provenance information to HBP-PROV,
- Registration of datasets and provenance in SP5 KnowledgeGraph
- Using Docker containers with the Matlab runtime and compiled Matlab code to remove the need to install Matlab with a license for hospitals.
- Better utilisation of computing resources with Mesos to run computations over a small and cheap cluster of machines
- Support for BIDS format to be able to process a wider set of research datasets, such as those provided by OpenfMRI.org
- Leverage I2B2 database to integrate software from other initiatives (IMI, Transmart) and integrate their datasets into the platform.
- Integration of automated PACS access still ongoing due to interfacing issues.

Links:

- <https://hbpmedical.github.io/specifications/data-factory/>
- <https://github.com/HBPMedical/mip-microservices-infrastructure/tree/master/demo/data-factory/airflow>



The screenshot shows the Airflow web interface with a table of DAGs (Directed Acyclic Graphs). The table lists various workflows such as 'clm_flat_metadata', 'clm_flat_organisation', 'clm_import_metadata', 'clm_pre_process_images', 'clm_reorganise', 'edsd_flat_metadata', 'edsd_flat_organisation', 'edsd_import_metadata', 'edsd_pre_process_images', 'edsd_reorganise', 'mri_notify_failed_processing', 'mri_notify_skipped_processing', 'mri_notify_successful_processing', and 'mri_self_checks'. Each row includes columns for DAG name, Schedule, Owner, Recent Tasks, Last Run, DAG Runs, and Links.

DAG	Schedule	Owner	Recent Tasks	Last Run	DAG Runs	Links
clm_flat_metadata	once	airflow		2017-05-17 07:00	1	View Refresh Delete
clm_flat_organisation	once	airflow		2017-05-17 07:00	1	View Refresh Delete
clm_import_metadata	once	airflow		2017-05-17 07:30	1	View Refresh Delete
clm_pre_process_images	once	airflow		2017-05-17 07:30	1	View Refresh Delete
clm_reorganise	once	airflow		2017-05-17 07:31	1	View Refresh Delete
edsd_flat_metadata	once	airflow		2017-05-17 07:00	1	View Refresh Delete
edsd_flat_organisation	once	airflow		2017-05-17 07:00	1	View Refresh Delete
edsd_import_metadata	once	airflow		2017-05-18 09:54	1	View Refresh Delete
edsd_pre_process_images	once	airflow		2017-05-17 09:24	1	View Refresh Delete
edsd_reorganise	once	airflow		2017-05-17 07:31	1	View Refresh Delete
mri_notify_failed_processing	once	airflow				View Refresh Delete
mri_notify_skipped_processing	once	airflow		2017-05-18 00:59	1	View Refresh Delete
mri_notify_successful_processing	once	airflow		2017-05-18 05:31	1	View Refresh Delete
mri_self_checks	once	airflow				View Refresh Delete

Figure 24: Data Factory using Airflow engine workflows for processing EDSD, PPMI and CLM datasets

Quality Control:

Upstream:



- *SOFTWARE > Data Factory > Data Anonymisation*: T8.1.1 - the anonymisation software prototype provided cannot connect to PACS systems and retrieve images. This is a Component needed for full automated collection and anonymisation of MRI scans and patient medical records. T8.1.1 developed a temporary solution by requesting a one-time export of anonymised data performed by the Hospital IT teams.
- *SOFTWARE > Algorithm Factory (AF) > MIP microservice infrastructure*: T8.5.2 - adapted to deploy the Data Factory

Downstream:

- *SOFTWARE > Data Factory > Workflow engine*: T8.5.2 - fully integrated and tested
- *SOFTWARE > Data Factory > Workflow tools*: T8.5.2 - fully integrated, functional tests ok, scientific tests pending
- *SOFTWARE > Data Factory > Omics Pipeline for feature engineering for Airflow*: T8.4.3 - fully integrated, functional tests ok, scientific tests pending
- *SERVICES > Upgrade - Deploy - Release > MIP Integrated Release > Algorithm Factory, Data Factory and Web Analytics Integration, Collaboratory integration*: T8.5.2 - this new component was taken into account in the release schedule

Other components have been partially integrated:

- *SOFTWARE > Data Factory > Data Storage* - the following databases have been integrated: Data Catalog, I2B2 capture, I2B2 CDE
- *MipMap*: T8.1.4 - EPFL provided the Docker image for MipMap, AUEB their expertise on this component plus bug fixes, CHUV provided the integration between MipMap and Airflow. Remaining tasks: having a fully functional ETL using MipMap to process EHR data from hospitals and research datasets and leveraging the I2B2 capture and I2B2 CDE databases.

6.6.3.3 *SOFTWARE > Data Factory > Omics Pipeline for feature engineering for Airflow*

Description: Automated processing of raw neuroimaging data using existing Neuroimaging tools, with standardised methods covering volume-based structural imaging.

Progress: CHUV provided version 1.2.6 of MRI-pre-processing-pipeline and worked with several partners on the VBQ toolbox for SPM.

The current version of the pipelines has been improved to facilitate its integration with Airflow and it has been demonstrated to work to produce feature extraction on the following research datasets: PPMI (727 subjects), EDSD (200 SUBJECTS), ADNI (partial results, 5000 subjects total); and clinical datasets: CHUV CLM (185 patients). Scientific confirmation of the results is pending.

Link: <https://github.com/HBPMedical/mri-preprocessing-pipeline>: 18 releases, current version 1.2.6.

Quality Control:

Upstream:

- *SOFTWARE > Data Factory > Workflow tools*: T8.5.2 - intermediate release

Downstream:

- *SOFTWARE > Data Factory* - Component delivered and integrated

6.6.3.4 SOFTWARE > Data Factory > Omics Pipeline for feature engineering for CBRAIN

Description: Automated processing of raw neuroimaging data using existing Neuroimaging tools, with standardised methods covering volume-based structural imaging.

Progress: Prototype integration of CBRAIN running inside Docker container realised with the help of McGill Centre for Integrative Neuroscience.

We expect to use CBRAIN in order to launch imaging pipelines on a cluster if needed. Further integration is planned with the developers of CBRAIN.

6.6.3.5 SOFTWARE > Data Factory > Data Storage

Description: Storage of data (EHR and imaging) in managed repositories for long-term storage, versioning and querying data. This storage is mostly intended for internal use in the Data Factory but it should also facilitate import of data from existing datasets found in hospitals or from research databases.

Progress: We have selected a Postgres relational database with the I2B2 database schema to efficiently capture all variables coming from hospitals or research, in line with how data import is done in Transmart / IMI project.

A second I2B2 schema is used on the same database to store the normalised Common Data Elements of the MIP (I2B2 CDE) containing the variables selected by the Data Governance and Data Selection (DGDS) Committee. Data in this second schema is then exported to PostgresRAW representing the LDSM (Local Data Store Mirror) that provides algorithms and queries with data containing usable features.

Finally, still on the same database, we have created a Data Catalog schema to keep track of the provenance information for all files and software used to transform data.

We have also evaluated several third-party software components:

- [LabKey](#): CHUV has created a Docker image to execute LabKey in a managed environment. After extensive testing, we found that LabKey did not have a good import module and we had concerns with the rigid organisation of data in LabKey, plus the lack of support for imaging data. We will keep monitoring development on LabKey as this tool could be used to build a stronger 'Data Capture' building block.
- [XNAT](#): CHUV has created a Docker image to execute XNAT in a managed environment (MIP microservices). This work has attracted external contribution from Phillips Research. XNAT is an interesting solution as it can manage imaging data as well as some metadata, but we found the product complex and we are developing collaboration from labs using XNAT.
- [LORIS](#): CHUV has tested LORIS for its image storage and quality control features. We also established contact with the developers of LORIS, from McGill University who have kindly offered their contribution in order to install and teach LORIS, but we need more time to proceed further (install and training)
- **RAW**: this component from T8.1 could be used to provide better queries over data stored into DICOM or Nifti files, but it could not be tested it as the functionality is not yet delivered. In addition, the only version of RAW that supports the query plug-ins for reading DICOM, Nifti and BIDS data is the proprietary RAW software provided from RAW Labs. The PostgresRAW software which contains upstream Postgres database plus a plug-in for read data stored in CSV files will need to support these plug-ins.

Links:

The following projects have been created for this Component:

- <https://github.com/HBPMedical/i2b2-setup> (Creation of the I2B2 database schema): 7 releases, current version 1.4.5



- <https://github.com/HBPMedical/data-catalog-setup> (Creation of the Data Catalog database used to capture provenance information on all files) : 10 releases, current version 1.4.5
- <https://github.com/HBPMedical/labkey-docker> (Docker image for LabKey): one release on Docker hub, LabKey software evaluated but not seen as suitable for our needs
- <https://github.com/HBPMedical/xnat-docker> (Docker image for XNAT): one release on Docker Hub, XNAT software under evaluation, no immediate need as the number of datasets that we manage is still small.

6.6.3.6 SOFTWARE > Data Factory > Workflow tools

Description: Software used by the Workflow engine: i.e. Neuroimaging software (Matlab, SPM), Genetic software.

Progress: CHUV updated Matlab to version 2016b and SPM 12 to version r6906. The use of Matlab is justified because of the strong community and trust in the scientific results produced by Matlab algorithms. SPM 12 requires Matlab to function. CHUV is investigating the use of compiled Matlab scripts as these scripts require only a Matlab runtime without license from Matlab's company. Discussion with Mathworks are ongoing, to provide the best solution at no additional cost for the data provider.

CHUV provided a new version of VBQ toolbox (svn166).

CHUV also provided automated deployment scripts for those tools which track the version of the software installed on the system.

Links:

- <http://www.fil.ion.ucl.ac.uk/spm>
- https://bitbucket.org/hbpmip_private/vbq-toolbox (open-source but no public release yet, pending approval from other institutions)

6.6.3.7 SOFTWARE > Data Factory > Workflow engine

Description: Adapt and configure an existing workflow engine (Apache Airflow) to provide automated processing of neuroimaging data using existing Neuroimaging tools, to keep track of provenance and to schedule ETL tasks.

Progress: CHUV integrated Apache Airflow 1.8.0 into the Data Factory. Deployment scripts have been provided (ansible-airflow) and we have built additional libraries to add some missing features to Airflow and customise the software to our needs:

- data-tracking to track data transformations and build provenance information
- airflow-imaging-plugins to facilitate the integration of ETL and neuroimaging pipelines into Airflow, with new Airflow operators providing provenance tracking and error reports in addition to the upstream functionality.
- hierarchiser to reorganise incoming imaging datasets and adapt their organisation to fit the structure expected by downstream neuroimaging methods. Current datasets supported are ADNI, PPMI, EDS, CHUV CLM.
- i2b2-import to import imaging features and provenance information into the I2B2 capture database.
- data-factory-airflow-dags to create the DAGs (pipelines) in Airflow leveraging the tools above and enabling image pre-processing and ETL data importation and transformation from all upstream datasets currently supported by the Data Factory (ADNI, PPMI...)

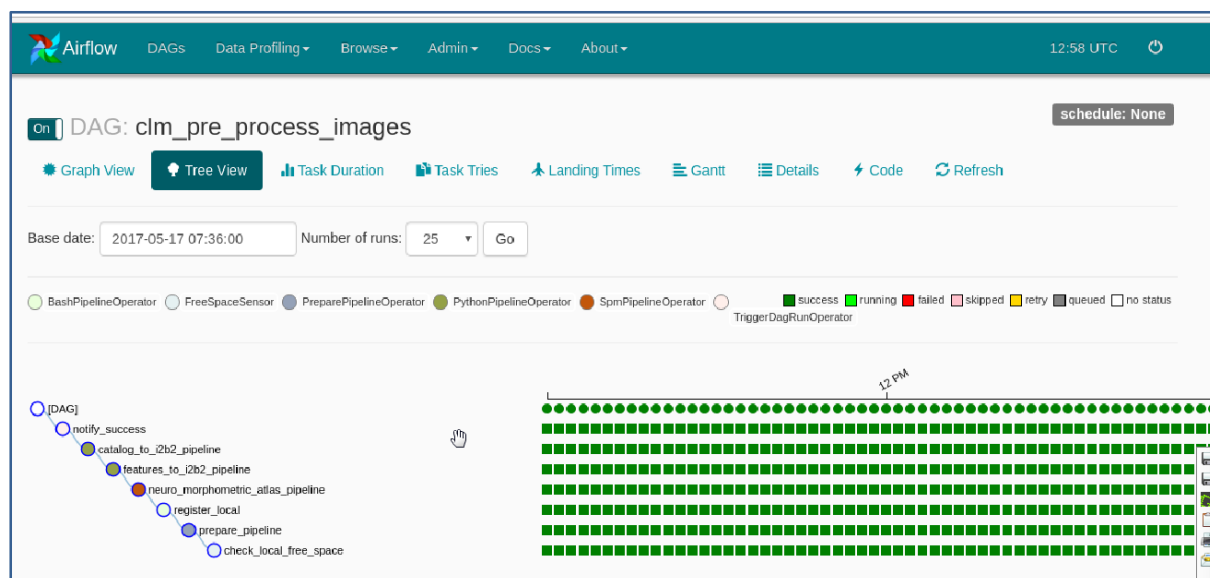


Figure 25: Image processing in Airflow engine

Links:

The following projects have been created for this component:

- <https://github.com/HBPMedical/data-factory-airflow-dags> : 40 releases, current version 0.9.0
- <https://github.com/HBPMedical/airflow-imaging-plugins> : 29 releases, current version 2.2.9
- <https://github.com/HBPMedical/data-tracking> : 27 releases, current version 1.5.4
- <https://github.com/HBPMedical/hierarchizer> : 17 releases, current version 1.2.0
- <https://github.com/HBPMedical/i2b2-import> (import of imaging and data catalog data to I2B2): 25 releases, current version 1.5.4
- <https://github.com/HBPMedical/ansible-airflow> : continuous releases

Quality Control:

Upstream:

- *SOFTWARE > Data Factory > Workflow tools*: T8.5.2 - fully integrated, functional tests ok, scientific tests pending
- *SOFTWARE > Data Factory > Data Storage* - the following databases have been integrated: Data Catalog, I2B2 capture, I2B2 CDE
- T8.1.4, EPFL provided the Docker image for MipMap, AUEB their expertise on this component plus bug fixes, CHUV provided the integration between MipMap and Airflow. Remaining tasks: having a fully functional ETL using MipMap to process EHR data from hospitals and research datasets, and leveraging the I2B2 capture and I2B2 CDE databases.

Downstream:

- *SOFTWARE > Data Factory > Omics Pipeline for feature engineering for Airflow* - a test environment has been provided (Vagrant airflow demo) and the pipeline has been tested in a production environment (Vertex machine installed in CHUV clinical network)

6.6.3.8 SOFTWARE > Algorithm Factory

Description: A 'building block' of MIP. The Algorithm Factory provides the runtime environment to execute data mining algorithms or machine learning algorithms, store the results of learning as predictive models, classifications or Biological Signatures of Diseases. It interacts with the 'Web-Based Medical Data Modelling and Visualisation' building block that provides the user interaction and launches the algorithms, and 'Hospital Database' or research database that provide the data with the features to learn from.

Progress: The Algorithm Factory has seen some consolidation of its main Component, Woken, which contains new features such as experiments. Those features have improved the user experience on the downstream *Web-based modelling and Visualisation*.

CHUV has tested [Cromwell](#) from the Broad Institute to provide complex workflow and machine learning tools capable of working on images. This looks like a good candidate and it could complement the functionality provided by Woken without additional development.

Link:

A demonstration of the main component of the Algorithm Factory, Woken, is available for testing by any developer at <https://github.com/HBPMedical/mip-microservices-infrastructure/tree/master/demo/algorithm-factory/woken>.

Quality Control:

Upstream:

- *SOFTWARE > Data Factory:* T8.5.2 - in progress - The Data Factory should provide the features extracted from imaging data and EHR data (patient records).
- *SOFTWARE > HDB > Query Engine:* T8.1.1 - in progress - The Query Engine provides the storage for the features. As the query engine delivered is a fork of a Postgres database, no additional work needs to be done. We still need to complete end-to-end testing of the integration between Data Factory, HDB and Algorithm Factory.
- *SOFTWARE > Algorithm Factory (AF) > MIP microservice infrastructure:* T8.5.2 - adapted to deploy the Algorithm Factory.

Downstream:

- *SOFTWARE > Web-based modelling and Visualisation:* T8.5.1 - working version of the Algorithm Factory provided and adapted to the requirements of this component.

6.6.3.9 SOFTWARE > Algorithm factory > Scientific computations > Woken

Description: An orchestration platform for Docker containers running data mining algorithms. Woken provides a web service and an API to execute on demand data analytics and machine learning algorithms encapsulated in Docker containers. Algorithms and their runtime are fetched from the Algorithm Repository, a Docker registry containing approved and compatible algorithms and their runtimes. Woken provides the algorithms with data loaded from a database, monitors the execution of the algorithm on one machine of a cluster, then it collects the result formatted as a PFA document and returns a response to the client. Woken tracks provenance information, runs cross-validation of the models produced by the ML algorithms.

Progress: CHUV setup a basic repository on public Docker Hub and containing a first selection of standard algorithms: linear-regression, summary-statistics, KNN, naive-bayes. Implementation and integration: The algorithms are executable from Woken and their results can be displayed on the Web Exploration and Analytics application. Testing: Unit tests provided for each algorithm integrated in the repository. Algorithm developers, statisticians and neuroscientists are evaluating the algorithms for correctness and usefulness.

Links:

- <https://github.com/HBPMedical/woken> : 19 releases on Docker hub, current version 2.0.
- <https://github.com/HBPMedical/woken-messages>
- <https://hub.docker.com/r/hbpmip/woken/> : List of Docker images containing the Woken program

Quality Control:

Upstream:

- *SOFTWARE > Algorithm factory > Scientific computations > Woken > Algorithm repository*: T8.5.2 - in progress - A first set of algorithms has been delivered, but the algorithm repository itself is only a simple list and should mature into a proper repository.

Downstream:

- *SOFTWARE > Algorithm factory* - current version of Woken has been delivered and packaged in a Docker image.
- *SOFTWARE > Algorithm Factory (AF) > MIP microservice infrastructure* - Deployment scripts for Woken has been provided.
- *SOFTWARE > Web-based modelling and Visualisation* - specifications for the new web services have been provided and work to integrate Experiments in the Web UI completed.

6.6.3.10 SOFTWARE > Algorithm factory > Scientific computations > Woken > Algorithm repository

Description: Repository of Docker images that can be used in Woken and workflow that allows contributors to provide new algorithms in a secured and principled manner.

Progress: CHUV provided a basic store implemented in a Postgres database. Implementation and integration: The store is used to keep track of results created by the execution of an algorithm. The MIP microservice infrastructure is used to deploy this store on the servers. Remaining work: search existing models from Web Exploration and Analytics application.

Links:

- <https://hub.docker.com/u/hbpmip/> (basic algorithm repository)
- <https://hub.docker.com/r/hbpmip/r-linear-regression/> (Docker image for the Linear Regression algorithm)
- <https://hub.docker.com/r/hbpmip/r-summary-stats/> (Docker image for the Summary statistics algorithm)
- <https://hub.docker.com/r/hbpmip/java-rapidminer-knn/> (Docker image for the KNN algorithm)
- <https://hub.docker.com/r/hbpmip/java-rapidminer-naivebayes/> (Docker image for the Naïve Bayes algorithm)
- <https://hub.docker.com/r/hbpmip/python-histograms/> (Docker image for the Histograms algorithm)

Quality Control:

Upstream:

- *SOFTWARE > Algorithm Library > Statistical Analytics > 3-C (Categorize, Cluster & Classify)*: T8.3.1 - in progress - Code delivered, integration pending.



- *SOFTWARE > Algorithm Library > Machine Learning Library > Disease signature:* Distributed rule-based methods: T8.3.5 - in progress - Integration work with Woken has started.
- *SOFTWARE > Algorithm Library > Machine Learning Library > Disease progression model from longitudinal biomarkers:* T8.3.12 - in progress - Integration work with Woken has started.

Downstream:

- *MIP - SOFTWARE > Algorithm factory > Scientific computations > Woken* - repository is usable from Woken, the list of algorithms provided will be formalised to simplify configuration.

6.6.3.11 SOFTWARE > Algorithm factory > Scientific computations > Woken > PFA model store

Description: Storage and search service for PFA models.

Progress: CHUV provided cross validation using random K-Fold Sampling methods.

Implementation and integration: The method is implemented in a separate application that can scale up to work around performance and memory issues. This application integrates with Woken and its startup and scaling is controlled by the MIP microservice infrastructure (via Marathon). **Testing:** Algorithm developers, statisticians and neuroscientists are evaluating the cross-validation functionality for correctness and usefulness.

Links:

- <https://github.com/HBPMedical/woken> : 8 releases on Docker hub, current version 2.0.
- <https://github.com/HBPMedical/woken-messages>
- <https://hub.docker.com/r/hbpmip/woken-validation/>: List of Docker images containing the Woken program

Quality Control:

Upstream: none

Downstream:

- *MIP - SOFTWARE > Algorithm factory > Scientific computations > Woken* - the cross-validation module is integrated with Woken and the REST API has been extended to provide support for experiments with cross-validation.

6.6.3.12 SOFTWARE > Algorithm factory > Scientific computations > Woken > Cross-validation module

Description: Cross validation module enables comparisons of models produced by different algorithms by evaluating the models on the same set of test data.

Progress: CHUV provided cross validation using random K-Fold Sampling methods.

Implementation and integration: The method is implemented in a separate application that can scale up to work around performance and memory issues. This application integrates with Woken and its start-up and scaling is controlled by the MIP microservice infrastructure (via Marathon).

Testing: Algorithm developers, statisticians and neuroscientists are evaluating the cross-validation functionality for correctness and usefulness.

Links:

- <https://github.com/HBPMedical/woken> : 8 releases on Docker hub, current version 2.0.
- <https://github.com/HBPMedical/woken-messages>



- <https://hub.docker.com/r/hbpmip/woken-validation/>: List of Docker images containing the Woken program

Quality Control:

Upstream: none

Downstream:

- *MIP - SOFTWARE > Algorithm factory > Scientific computations > Woken* - the cross-validation module is integrated with Woken and the REST API has been extended to provide support for experiments with cross-validation.

6.6.3.13 SOFTWARE > Web Exploration and Analytics > Information and Scientific references > Research Object wrapper

Description: Viewers, search and export for full Research Objects including OpenBEL.

Progress: CHUV selected PFA (Portable Format for Analytics) to provide most of the information for a Research Object. CHUV also participates in the work group that defines and promotes PFA as a standard for representing the results of predictive algorithms. This working group includes people from IBM, NIST, Amazon, SAS, University of Chicago, HPE, MITRE. We will build on PFA specifications to include provenance information, a solution for automated visualisation of the results, integration with other research object standards such as OpenBEL. Meta-data in the form of PFA documents are stored in a Postgres database.

Implementation and integration: PFA is used end-to-end, from the result of a machine-learning algorithm to the visualisation of the results in the Web Exploration and Analytics application.

Testing: Web developers and algorithm developers use PFA to represent the results of statistical and machine learning algorithms.

Link: <http://dmq.org/pfa/index.html>

Quality Control:

Upstream: none

Downstream:

- *MIP - SOFTWARE > Algorithm factory > Scientific computations > Woken > Cross-validation module* - the cross-validation module uses the predictive algorithms serialised in PFA to compute the cross validation of a model.
- *MIP - SOFTWARE > Web-based Modeling and Visualisation* - The web interface provides several viewers to show the models and the results of cross-validation serialised into PFA documents.

6.6.3.14 SERVICES > Connection to 3rd party application and data > POC with CREATIVE & Traumatic Brain Injury data (TBI)

Description: In collaboration with CREATIVE team in Bergamo, Italy

Progress: CHUV provided a prototype that demonstrates the usefulness of MIP for analysis the traumatic brain injury data captured from one site. CREATIVE approved our demonstration and are requesting MIP to help them deliver distributed analyses over several sites. Next step: multi-site demonstration of the platform with data from CREATIVE and other partners

6.6.3.15 SERVICES > Connection to 3rd party application and data > POC with Aetionomy

Description: Integration of BINE (Genes and Rules). Integration as App into MIP.



Progress: CHUV provided a prototype that integrates BINE as an app into MIP.

6.6.3.16 SERVICES > Upgrade - Deploy - Release > CI (Continuous Integration) workflow

Description: Tools and workflows for Continuous Integration of the software

Progress: CHUV has setup continuous integration for 11 projects on CircleCI.com, 1 project on Wercker.com, and continuous code quality checks for 10 projects on Codacy.com

Link: <https://hbpmedical.github.io/software-catalog/> for a live dashboard.

6.6.3.17 SERVICES > Upgrade - Deploy - Release > QA (Quality Assurance) Tools

Description: Tools for quality assurance of the project

Progress: CHUV is organising the 'SP8 Software Development committee' and has created a public website to inform all SP8 members.

CHUV has setup code quality checks for 10 projects on Codacy.com, and setup 12 continuous integration workflows on CircleCI.com and Wercker.com.

Links:

- <https://hbpmedical.github.io/development-guidelines/intro/> for the development guidelines prepared by the 'SP8 Software Development committee' (work in progress)
- <https://hbpmedical.github.io/software-catalog/> for a live dashboard

6.6.3.18 SOFTWARE > Algorithm Factory (AF) > MIP microservice infrastructure

Description: Platform for rapidly deploying globally distributed services. It supports clustering, security, monitoring and more out of the box. It is based on Cisco's Mantl cloud project (<https://github.com/CiscoCloud/mantl>). The aim of this Component is to support the deployment of many services in the MIP and provide an environment for big data software such as Apache Spark.

Progress: CHUV stabilised the deployment of Mesos stack on Ubuntu 16.04, added support for automated deployment or upgrade of services managed by Mesos and Marathon, added support for new services and databases, added security hardening of an Ubuntu system, provided demonstrations of several building blocks (Web Portal, Algorithm Factory, Data Factory) using Vagrant and Ansible scripts. CHUV maintained and updated all software managed by this project on the production servers. EPFL contributed support for PostgresRAW and an integration test between Data Factory and PostgresRAW.

Implementation and integration: Twelve different software services are managed by MIP microservice infrastructure plus all databases, security and monitoring tools. The scope of the microservice infrastructure has greatly expanded to cover the deployment and monitoring of all building blocks of MIP, deployment of different configurations of the Platform (MIP Local deployed on a single computer, MIP Federated deployed onto several clusters inside hospitals).

Testing: Vagrant projects are provided to test the deployment of some of the services and building blocks (Portal, Algorithm Factory/Woken, and Data Factory/Airflow).

Release: MIP microservice infrastructure is used to manage the following environments: QA in CHUV academic datacenter, Federation/production in CHUV academic datacenter, Data Factory/production in CHUV clinical network.

Quality Control:

Upstream:

- SOFTWARE > Encrypted Overlay Network, T8.1.3 - development still ongoing.

Downstream:

- SERVICE > MIP Local > Deployment, T8.1.3 - CHUV provided working deployment scripts on Ubuntu system with numerous examples (demos in Vagrant containers and fully configured environments - QA, production, Vertex) under our management. Documentation may be sparse but it is based on well documented Ansible tool.
- SERVICE > MIP Federated > Deployment, T8.1.3 - waiting on SOFTWARE > Encrypted Overlay Network and a solution to deploy Exareme in MIP infrastructure.
- SERVICE > Data Capture > Data Access > Deployment: deployment scripts for the Data Capture and anonymisation software pending.
- SERVICE > HDB > Query Engine > Integration - EPFL has provided an integration of the Query Engine PostgreRAW
- SOFTWARE > Remote Starting of Services - this service is provided by the Marathon tool contained in MIP Microservice Infrastructure

6.6.3.19 SERVICES > Upgrade - Deploy - Release > MIP Integrated Release > Algorithm Factory, Data Factory and Web Analytics Integration, Collaboratory integration

Description: Platform for rapidly deploying globally distributed services. It supports clustering, security, monitoring and more out of the box. It is based on Cisco's Mantl cloud project (<https://github.com/CiscoCloud/mantl>). The aim of this Component is to support the deployment of many services in the MIP and provide an environment for big data software such as Apache Spark.

Progress: CHUV released the Algorithm Factory, Web Analytics Integration, Collaboratory integration version 2.0 on the production Federation in M08 SGA1.

CHUV released Data Factory version 1.0 on the production clinical server at CHUV at M12 (March 2017).

6.6.3.20 SERVICES > Security & Monitoring > Security, Load balancing, Clustering and Recovery Services

Description: Services for managing applications running in a cluster provided by the Mesos stack.

Progress: Those services are part of the MIP Microservice infrastructure. Health monitoring and load balancing has been improved. Security hardening is provided only on SSH servers and web servers.

Missing/planned features: Security at all levels, including for remote administration of MIP Infrastructure / Mesos.

6.6.3.21 SERVICES > Security & Monitoring > Security, Load balancing, Clustering and Recovery Services

Description: Services for managing applications running in a cluster provided by the Mesos stack.

Progress: These services are part of the MIP Microservice infrastructure. Health monitoring and load balancing has been improved. Security hardening is provided only on SSH servers and web servers.

Missing/planned features: Security at all levels, including remote administration of MIP Infrastructure / Mesos.

6.6.3.22 SERVICES > Security & Monitoring > User Record Management Services

Description: User management from the Collaboratory.



Progress: CHUV provided the integration with Collaboratory User Management services. **Implementation and integration:** Use more data from user profile, fix security hole on logout. **Testing:** Basic checks. **Release:** General availability.

6.7 T8.5.3 - Integration and Technical Coordination of WP8.5, Integration into Collaboratory

6.7.1 Key Personnel

Task Leader: Ferath KHERIF (CHUV)

6.7.2 SGA1 DoA Goals

The goal of this Task is to ensure that the work of T8.5.1 and T8.5.2 (Web Portal, backend micro-services, Collaboratory) is well integrated into the MIP. Perform functional tests and approve releases, plan UAT tests. Coordinate the project team.

6.7.3 Component Progress

6.7.3.1 SERVICES > Upgrade - Deploy - Release > MIP Integrated Releases > Platform Integration, Release Planning & Coordination

Description: Platform integration work and related projects, planning and coordination of Platform releases.

Progress: Version v1 of the MIP was delivered at M30 RUP (1st public release).

Version v2 of the MIP was delivered at M08 SGA1. V2 was a major release with important new functionalities delivered to the end-user. This functionality is currently in production. The integration of the different components, the system testing, user acceptance tests and deployment was coordinated and in majority done by the technical team at CHUV. The work was in collaboration with the other Partners whose Components were included: UoA.

User functionality in v2 (live): Improved usability of Exploration of variables (EE), added description of variables, improved descriptive statistics view, improved usability of Interactive Analysis (IA), NEW functionality: the Experiment builder (deep learning algorithm configuration, creation of disease models, X-validation results), improved Apps section (GUI), integrated experiments into article writing and user personal dashboards).

Current Component-blocks in production: Algorithm Factory, Algorithm Library (deep learning algorithms developed at CHUV), Web Analytics & Visualisation, Meta-data (MIP variables).

Current data in MIP (demo): PPMI, ADNI, ESDI.

Other work & accomplishments:

CHUV: Set up the different software development environments for the integrated MIP solution, the testing strategy, the QA-tools (Quality Assurance, Docker-based) and CI-process (Continuous Integration). The CI process allows an automation and versioning of deployment in-between stages, and so minimising the errors and the effort significantly. The QA-tools & CI process are in place and have been presented to the HBP Software Infrastructure committee on 9 September 2016.

6.7.3.2 SERVICES > Upgrade - Deploy - Release > MIP Integrated Release > Algorithm Factory, Data Factory and Web Analytics Integration, Collaboratory integration

Description: This service coordinates the release of the different Components of the MIP and uses the services from MIP Microservice infrastructure.



Progress: CHUV released the Algorithm Factory, Web Analytics Integration, Collaboratory integration version 2.0 on the production Federation in M08 SGA1. UoA released Exareme, EPFL released PostgresRAW and AUEB released WebMipMap on the CHUV node at the same date.

CHUV released Data Factory version 1.0 on the production clinical server at CHUV in M12 SGA1.

6.7.3.3 SERVICES > SP8 Coordination & Project Management Services > Quality Controls > Testing Strategy > User involvement in platform implementation (design, tests)

Description: Involve users in the platform development by setting up feedback forms, recruiting beta testers, involving users in the Platform plans

Progress: The MIP Testing strategy was defined, delivered as part of the D8.6.1, and approved by the External Review in October 2016. The MIP users (current and future) have been involved in all stages of the project lifecycle, from the design of the user interface, to functional testing and later in providing post-release feedback. User types used: statisticians, clinicians, data scientists, neuroscientists.

Partners: the MIP testing strategy was delivered by CHUV, and the coordination of the user involvement, the collation of feedback and its analysis, also done by CHUV.

In addition, a Software and Infrastructure Governance group was created with members from all SP8 software developing groups. The group meets regularly and agree on common standards within SP8 software, project tracking tools to use etc. The group is also cross-cutting with the homologue parent, the HBP Infrastructure and Software group, to ensure overall governance direction is applied. The committee is led by CHUV, is open to all SP8 members, but with required attendance from the software developing groups (EPFL, UoA, AUEB).



7. WP8.6 - Scientific Coordination

7.1 Key Personnel

Work Package Leader: Ferath KHERIF (EPFL)

7.2 WP Leader's Overview

The WP8.6 has performed activities defined in the Definition of Actions - coordinating the work of Work Packages in SP8 and their interaction with HBP management, other SPs and the larger community, the scientific and technical coordination within SP8 with other SPs.

This work includes also: coordination of Platform testing; standardisation of data formats, terminology and development processes; coordination of documentation and dissemination of standards; coordination of software development, release management and deployment; coordination of user engagement; and representation of SP8 on the Scientific and Technical Coordinators Committee.

To answer concerns raised by the reviewers', WP8.6 (CHUV) took, according to the Fast-Track plan, new roles and responsibilities to ensure the successful deployment of the platform to hospitals. A Release Manager was recruited and a deployment plan was proposed to and accepted by the European Commission.

Therefore, the coordination includes the additional workload related to reviews as well as that defined within the Fast-track plan. These are:

- Review preparation and coordination;
- Coordination with the Data Governance & Data Selection committee (DGDS) which includes data scientists from SP8 and clinicians from the five hospitals. This group works with the Deployment Coordinator to ensure smooth adoption of the system to the hospitals (MIP-Local & MIP-Fed solutions);
- Coordination of MIP-Local deployment to the 5 hospitals. The hospitals were provided with a complete *MIP Deployment plan* (deployment plan in detail, software to be installed, hardware and security needs, collaboration efforts expected, ethics) and *Evaluation Agreement* (legal agreement between MIP and Hospitals for the deployment process and MIP evaluation period by the local clinicians);
- Deployment to hospitals is underway and is supervised by CHUV.

A CHUV Project Manager (M. DAMIAN) reduced her activity to 40% since M12. A full-time Project Manager started on 1 April 2017.

7.3 Priorities for the remainder of the phase

Preparation and submission of an amendment for the integration of the hospitals as 3rd party partners.

Preparation of the next technical review.

Preparation of SGA2 proposal.

Building close interaction with the other SPs.

Continue development and deployment of the MIP.



7.4 Milestones

Table 6: Milestones for WP8.6 - Scientific Coordination

MS No.	Milestone Name	Leader	Task(s) involved	Expected Month	Achieved Month	Comments
MS17	MS 8.6.1 Project Implementation and Delivery Plan, SP level	CHUV	T8.6.1 / T8.6.2	M03	M06	During the first 6 months of SGA1, SP8 (T8.6.1/T8.6.2) prepared the revised MIP Platform Fast-Track Plan for the work to be carried out by the SP8 team during SGA1. It includes the new MIP uses cases, new scientific requirements, new functional and non-functional requirements that follow from the use cases, and scientific needs
MS117	MS 8.6.2 Progress Update Report	CHUV	All tasks	M12	M12	The D8.6.2 Deliverable (M12) includes the information on the progress made in the period SGA1 M0-M12.



7.5 T8.6.1 - Scientific Coordination and SP Coordination

7.5.1 *Key Personnel*

Task Leader: Ferath KHERIF (CHUV)

7.5.2 *SGA1 DoA Goals*

This Task will:

- Coordinate SP8 reporting and writing of Deliverables.
- Monitor scientific progress within the Subproject.
- Act as the SP8 point of contact for the HBP Administration.
- Coordinate work and communication between SP8's Workpackages and Tasks.
- Organise SP8-wide meetings.
- Organise one SIB meeting.
- Coordinate with the Management team on issues related to innovation.
- Coordinate with the Ethics Manager and with SP12 on issues related to ethics.
- Provide support to partners on issues related to administration, innovation and ethics.

7.5.3 *Specific work effected*

The information below describes work effected in addition to the on-going standard coordination activities.

Project management follows standard PM processes. The development projects and components are monitored using the Trello tool.

Concerning user support and strategies to manage users and meet their expectations, several workshops and webinars (see chapter Education) have been conducted. The MIP has generated interest from other potential data providers (including other hospitals and research cohorts), pharma (via IMI or directly) and the machine learning industry (Bearingpoint, Google Deepmind) among others.

With respect to data governance, the Data Governance and Data Selection (DGDS) Committee enforces the policies for data handling in the HBP and contributes to the HBP Data Governance policies document. The Data DGDS Committee ensures in concerted action with SP12 the conformity to ethical principles of data handling and addresses ethical concerns for all data used by the Medical Informatics Platform. On the technical side, the DGDS Committee selected variables of interest. None of these variables contains individually identifiable patient information. Work progressed in coordination with T8.1.4, Exareme to further protect patient anonymity using aggregated results and query filtering.

7.6 T8.6.2 - Medical Informatics Platform Strategy and Business Model

- Task name to be changed to Medical Informatics Platform Software deployment, Release management and strategy.

7.6.1 *Key Personnel*

Task Leader: Ferath KHERIF (CHUV)



7.6.2 SGA1 DoA Goals

The Task aims to enforce the two-stage deployment strategy (MIP-Local, and MIP-Federated). More specifically it should:

- Prepare the formal deployment and evaluation agreement documents for the hospitals
- Formalise precise technical requirements for hospitals relative to the deployment of both stages
- Provide a main point of contact for administrative, technical, scientific requests from the hospitals and relay them if necessary to the relevant person within the SP
- Ensure that every software Component produced by the different SP's partners has been properly tested and integrated to the platform before being released
- Coordinate technical teams, both from SP8 and hospitals, during the installation and configuration of the product at hospitals
- Keep report status and deployment plan up-to-date
- Ensure that the architecture and software are documented

7.6.3 Specific work affected

The information below describes work effected in addition to the on-going standard coordination activities.

7.6.3.1 Fast-track - Hospital deployment

Following the definition of the Fast-Track plan, this Task is also in charge of deployment of MIP locally ("MIP-Local") and federatively ("MIP-Federation") to the five collaborating hospitals.

CHUV formalised a general strategy and a concrete plan describing administrative, technical, and scientific steps and their dependencies to be achieved to deploy both MIP-Local and MIP-Federated.

The hospitals have been provided with the complete plan (timeline, software description, hardware and security needs, collaboration efforts expected, ethical agreement needs) and the Evaluation Agreement (legal agreement between SP8 and Hospitals for the deployment process and the evaluation period of the MIP by the local clinicians). MIP deployment is on track.

Deliverables. Following the Expert Reviews' comments, "Document-1" was delivered and submitted. Based on the Fast-Track plan it details the MIP-Local and MIP-Federated deployment strategy and process, as well as the Agreement sent to Hospitals. Before submission. "Document-1" was approved by the Medical Platform Advisory Team. Furthermore, it also addresses foreseeable risks associated with this plan and methods enabling risk control, ensuring staying on track.

CHUV also introduced a concrete task repartition schema concerning the technical steps within the SP for deploying the Platform at hospitals. The RASCI matrix as well as the detailed work plan was the result of a consultative work across the whole SP8 team.

7.6.3.2 Platform Integration, Release Planning & Coordination

The plans for the release of Components and the elaboration of Milestones, the Platform integration work and related projects, planning and coordination of Platform release are discussed and monitored in the SP-wide meetings. The advancement of Components is monitored and integrated into the advancement of the building blocks.



For MIP-Local, a set of software components has been selected and approved to be part of MIP-Local release. They are listed below by building blocks:

- Data Capture
 - *SOFTWARE > Data Factory > Data Anonymisation* - T8.1.1
- Data Factory v1.0.0
 - All related Components from T8.5.2
 - *SOFTWARE > HDB > Online Data Integration Module (software)* - T8.1.4
- Algorithm Factory v2.0.0 T8.5.2
 - All related Components from T8.5.2
- Web portal v2.0.0
 - All related Components from T8.5.1

Automated deployment scripts were provided by CHUV. They are still being adapted to support operating systems packaged on the machines provided by the hospitals.

For MIP-Federated, the main features have been identified and development and integration of components are in progress.

7.6.3.3 Software Development coordination

The Software Development and Integration Committee was created with members from all SP8 software developing groups, and a public website to inform all SP8 members. It provides quality standards, strategy of validation of solutions and tools and coordinates development projects.

The delivery of contributions in code (open source, on Github only) amounts to 1,644 contributions in the last year, with 13 new projects, for a total of 35 projects. There are public projects actively maintained and seven private/confidential projects. The number of activities completed (development, documentation, infrastructure, management amounts to 60.

For code Quality Assurance tools, CHUV has setup code quality checks for 10 projects on Codacy.com, and setup 12 continuous integration workflows on CircleCI.com and Wercker.com.

Links:

- <https://hbpmmedical.github.io/development-guidelines/intro/> for the development guidelines prepared by the 'SP8 Software Development Committee' (work in progress)
- <https://hbpmmedical.github.io/software-catalog/> for a live dashboard.

7.6.3.4 Platform Testing Strategy

The Software Development and Integration Committee is defining the baseline requirements to meet for each of the participating projects. The strategy includes unit tests, integration tests and user acceptance tests.

Unit testing is mandatory for all projects and so far we have 32 projects published on Github or internal with some form of unit testing. We are working towards improving and monitoring test coverage for all projects. A detailed plan for unit testing of the algorithms is also under preparation.

We are using Continuous integration to execute the unit tests regularly on each commit and monitor the health of the code for the Platform.



CHUV and EPFL are also working on defining integration tests for the various components of the platform. Our strategy is to use Vagrant to build virtual machines containing parts of the components, for example Data factory + LDSM from Hospital Database where Vagrant is used to build the runtime environment and test scripts verify that processed by the Data Factory goes into the LDSM database and can be used by downstream algorithms and queries.

Link:

- <https://github.com/HBPMedical/mip-microservices-infrastructure/tree/master/integration-tests> (Integration tests using Vagrant to recreate parts of the infrastructure)

8. Publications

Venetis T, Stoilos G, Vassalos V. (2016). Rewriting Minimisations for Efficient Ontology-Based Query Answering. *ICTAI*;1095-1102. [T8.1.4]

Mallios X, Vassalos V, Venetis T, Vlachou A. (2016). A Framework for Clustering and Classification of Big Data Using Spark. *OTM Conferences 2016*:344-362. [T8.1.4]

Mahfouz A, Huisman SMH, Lelieveldt BPF, Reinders MJT. (2017). "Brain Transcriptome Atlases: a Computational Approach", *Brain Structure and Function*; DOI: 10.1007/s00429-016-1338-2, 2017. [T8.3.10]

Sun Z, van de Giessen M, Lelieveldt BPF, Staring M. (2017). "Detection of Conversion from Mild Cognitive Impairment to Alzheimer's Disease Using Longitudinal Brain MRI", *Frontiers in Neuroinformatics*, <https://doi.org/10.3389/fninf.2017.00016>. [T8.4.4]

Eising E, Huisman SMH, Mahfouz A, Vijfhuizen L, [the International Headache Genetics Consortium], Nyholt D, de Vries B, Lelieveldt BPF, van den Maagdenberg AMJM, Reinders MJT. (2016) "Gene co-expression analysis identifies brain regions and cell types involved in migraine pathophysiology: a GWAS-based study using the Allen Human Brain Atlas", *Human Genetics*; doi: 10.1007/s00439-016-1638-x. [T8.4.4]

Li T, Heinis T, Luk W. (2016). Hashing-Based Approximate DBSCAN. *ADBIS*; 31-45. [T8.4.5]

Li T, Heinis T, Luk W. (2017). ADvaNCE - Efficient and Scalable Approximate Density-Based Clustering Based on Hashing. *Informatics Journal*. [T8.4.5]

Kralj J, Robnik-Šikonja M, Lavrač N. (2017). HINMINE: heterogeneous information network mining with information retrieval heuristics, *Journal of Intelligent Information Systems*; 1-33, doi:10.1007/s10844-017-0444-9. [T8.3.7]

Osojnik A, Panov P, Džeroski S. (2016). Modeling dynamical systems with data stream mining. *Computer Science and Information Systems*;13(2):453-473, doi: 10.2298/CSIS1505180090. [T8.3.8]

Osojnik A, Džeroski S, Kocev D. (2016). Option predictive clustering trees for multi-target regression. In: *Discovery Science: 19th International Conference, DS 2016 Bari, Italy, 19-21 October 2016: proceedings*, (Lecture Notes in Computer Science, ISSN 0302-9743, Lecture Notes in Artificial Intelligence, LNCS 9956, Springer, vol. 9956, pp. 118-133, doi: 10.1007/978-3-319-46307-0_8. [T8.3.8]

Simidjievski N, Todorovski L, Džeroski S. (2016). Learning ensembles of process-based models by bagging of random library samples. In: *Discovery Science : 19th International Conference, DS 2016 Bari, Italy, 19-21 October 2016: proceedings*, (Lecture Notes in Computer Science, ISSN 0302-9743, Lecture Notes in Artificial Intelligence, LNCS 9956), Springer, vol. 9956, pp. 245-260, doi: 10.1007/978-3-319-46307-0_16. [T8.3.8]

Petković M, Panov P, Džeroski S. (2016). A comparison of different data transformation approaches in the feature ranking context. In: *Discovery Science : 19th International Conference, DS 2016 Bari, Italy, 19-21 October 2016 : proceedings*, (Lecture Notes in Computer Science, ISSN 0302-9743, Lecture Notes in Artificial Intelligence, LNCS 9956), Springer, vol. 9956, pp. 310-324, doi: 10.1007/978-3-319-46307-0_20. [T8.3.5]

Novak M, Zalar P, Ženko B, Džeroski S, Gunde-Cimerman N. (2016). Yeasts and yeast-like fungi in tap water and groundwater, and their transmission to household appliances. *Fungal Ecology*;20:30-39, doi: 10.1016/j.funeco.2015.10.001. [T8.3.8]

Osojnik A, Panov P, Džeroski S. (2016). Multi-label classification via multi-target regression on data streams. *Machine Learning*; 1-26, doi: 10.1007/s10994-016-5613-5. [T8.3.5]

Osojnik A, Panov P, Džeroski S. (2016). Comparison of tree-based methods for multi-target regression on data streams. In: *New Frontiers in Mining Complex Patterns: 4th International*



Workshop, NFMCP 2015 held in conjunction with ECML-PKDD 2015 Porto, Portugal, 7 September 2015 : revised selected papers, (Lecture Notes in Computer Science, ISSN 0302-9743, Lecture Notes in Artificial Intelligence, 9607), Springer, vol. 9607, pp. 17-31, doi: 10.1007/978-3-319-39315-5_2. [T8.3.5]

Soldatova LN, Panov P, Džeroski S. (2016). Ontology engineering: from an art to a craft. In: *Ontology Engineering: 12th International Experiences and Directions Workshop on OW, OWLED*, 2015, Co-located with ISWC 2015, Bethlehem, Pa, USA, 9-10 October 2015 : revised selected papers, (Lecture Notes in Computer Science, ISSN 0302-9743, vol. 9557), Springer, vol. 9557, pp. 174-181, doi: 10.1007/978-3-319-33245-1_18. [T8.3.9]

Tanevski J, Todorovski L, Džeroski S. (2016). Process-based design of dynamical biological systems. *Scientific Reports*;6:34107-1-34107-13, doi: 10.1038/srep34107. [T8.3.8]

Levatić J, Ceci M, Kocev D, Džeroski S. (2017). Semi-supervised classification trees. *Journal of Intelligent Information Systems*; doi: 10.1007/s10844-017-0457-4. [T8.3.8]

Levatić J, Ceci M, Kocev D, Džeroski S. (2017). Self-training for multi-target regression with tree ensembles. *Knowledge-Based Systems*;123:41-60, doi: 10.1016/j.knosys.2017.02.014. [T8.3.8]

Kranjc J, Orač R, Podpečan V, Lavrač N, Robnik-Šikonja M. (2017). ClowdFlows: Online workflows for distributed big data mining. *Future Generation Computer Systems*;68:38-58, doi: 10.1016/j.future.2016.07.018. [T8.3.5]

Gamberger D, Ženko B, Mitelpunkt A, Lavrač N. (2016). Homogeneous clusters of Alzheimer's disease patient population. *Biomedical Engineering Online Journal*;15(Suppl 1):78, doi: 10.1186/s12938-016-0183-0. [T8.3.6]

Valmarska A, Lavrač N, Fuernkranz J, Robnik-Šikonja M. (2017). Refinement and selection heuristics in subgroup discovery and classification rule learning. *Expert Systems with Applications*;81:147-162, doi: 10.1016/j.eswa.2017.03.041. [T8.3.6]

Lorio S, Kherif F, Ruef A, Melie-Garcia L, Frackowiak R, Ashburner J, Helms G, Lutti A, Draganski B. (2016). Neurobiological origin of spurious brain morphological changes: A quantitative MRI study. *Human Brain Mapping*;37:1801-1815. doi:10.1002/hbm.23137. [T8.4.3, T8.3.11]

9. Dissemination

IEEE 12th International Conference on eScience (October 2016) - slides of the talk can be found at <http://escience-2016.idies.jhu.edu/wp-content/uploads/2016/11/AILAMAKI-slides-red.pdf> [Keynote, A. AILAMAKI, WP8.1].

IPMI 2016 - Information Processing in Medical Imaging (June 2016) [Keynote, B. LELIEVELDT, T8.3.10].

MAESTRA Summer school Ochrid, Macedonia (July 2016) - Controlling false discovery rates in multiple testing and selective inference and analysis of neuro data [Keynote, Y. BENJAMINI, WP8.3]

Women in Data Science Conference Rutgers (February 2016) - Targeting "Disease Signatures" Towards Precision Brain Healthcare" [Keynote, M. MARCUS-KALISH, WP8.3]

Israeli Psychiatric and Geriatric Association Annual Meeting (March 2016) - "Disease Signatures" The Promise & Challenges - Big Data Vs Small Data Analysis Towards Personalized Medicine Practice" [lecture, M. Marcus-Kalish, WP8.3], [lecture, A. MITELPUNKT., WP8.3].

3rd Beyond Watson Workshop - Knowledge from Data and CHMP - NSF board meeting in UMBC Maryland USA. (May 2016) - "Health & Medical Data Science - The HBP Flagship" [keynote, M. MARCUS-KALISH, WP8.3].



CLINAM - Clinical Nano-Medicine International meeting in Basel (June 2016) - "Micro & Macro Environment Features towards Brain Disease Sub-Type Identification" [Keynote, M. MARCUS-KALISH, WP8.3].

CHMPR NSF Center Board meeting, including 8 USA universities and 20 industries. (December 2016) - "Converging Knowledge & Data Science for Health - the HBP Medical Informatics example" [Keynote, M. MARCUS-KALISH, WP8.3].

FENS 2016 - HBP booth with demo of the MIP [Demo, J. CUI, T8.4.4].

BACN 2016 (September 2016) - Estimation of Alzheimer's disease severity in vivo with MRI-based measures of atrophy [J. CUI, T8.4.2, T8.4.4].

Junior Doctors International Meeting 2016 (November 2016) - Human Brain Project [Keynote, B. DRAGANSKI, T8.2.3].

SfN 2016 - HBP booth with demo of the MIP [Demo, F. KHERIF, T8.4.4, WP8.5].

STOA exhibition (November 2016) - The Human Brain Project - Building a 21st Century European Infrastructure for Advanced Brain Science [Keynote and demo, F. KHERIF, T8.4.4, WP8.5, WP8.6].

13th International Conference on Alzheimer's & Parkinson's Diseases (March 2017) [Talk, G. SANABRIA DIAZ, T8.4.4].

IMI - "Alzheimer's disease: advancing research through collaboration", European Parliament, Brussels (November 2016). [Roundtable discussion, F. KHERIF, WP8.2 & WP8.6].

Epilepsy Alliance Europe "Towards a Global Alliance on Epilepsy Research", European Parliament, Brussels (February 2017). [Keynote, Demo, Talk, F. KHERIF, WP8.2 & WP8.6]

IMI Event - "Collaboration in Alzheimer's disease & beyond: present and future of IMI initiatives in neurodegeneration", European Parliament, Brussels (March 2017). [Keynote, Demo, Talk, F. KHERIF, WP8.2 & WP8.6]

Presentation of the Medical Informatics Platform to the Centre Leenards de la Mémoire, Centre Hospitalier Universitaire Vaudois, Switzerland (June and July 2016). [Talk, F. KHERIF, WP8.2]

SPM Course, LREN - Demo of the Medical Informatics Platform, Centre Hospitalier Universitaire Vaudois, Switzerland (June 2016). [Demo, F. KHERIF, WP8.2]

OHBM 2016 - Presentation on HBP-related work done at CHUV, Geneva, Switzerland (June 2016). [Posters, G. SANABRIA-DIAZ, L.M ELIE-GARCIA, F. KHERIF, WP8.2]

MIP presentation to the heads of Geriatric service at CHUV: C. Bula, P. Chassagne, M.Humbert, Centre Hospitalier Universitaire Vaudois, Switzerland (July 2016). [Demo, F. KHERIF, WP8.2 and WP8.6]

Symposium of the Neuroscience Research Center at CHUV, Update on the Human Brain Project, Centre Hospitalier Universitaire Vaudois, Switzerland (August 2016). [Talk, F. KHERIF, WP8.2 and WP8.6]

Brain meeting of Functional Imaging Laboratory (FIL), UCL, London, UK (September 2016). [Talk, R. FRACKOWIAK]

Big Data for Neuroscience - Medical Informatics and Federated Data Analysis in Neuroscience Networks, Pavia (October 2016) [Talk, F. KHERIF]

10. Education

- Thomas HEINIS recorded two lectures with and for the education program on databases that will be available to PhD students in the HBP.

- Ferath KHERIF and members of SP8 gave both theoretical lectures and hands-on sessions at the 3rd HBP school: Future Neuroscience - the Multiscale Brain: From Genes to Behaviour. 28 November - 4 December 2016. Obergurgl, Austria.
- Preparation of the next HBP education workshop “Brain Disease for non-specialists”.
- Preparation of the next HBP School on “brain disease”. Created a program committee with large contributions from all SPs and external experts (from clinical research to pharma) including theoretical lectures and hands-on sessions.
- Several training workshop were organised at CHUV.

11. Ethics

- During the first 12 months of SGA1, SP8’s Ethics Rapporteur attended meetings with the Ethics Management Compliance group including regular updates of the “One-pagers” with the EAB members (see also SGA1_M12_201609MS1241Firstreportonethicsmanagementactivities.pdf by SP12).
- In the Data Governance and Data Selection (DGDS) Committee the discussion on relevant ethical questions resulted in adopting SP12 guidelines for Informed Consent (see SGA1_UPDATED_D71-1-Informed Consent.pdf by SP12) and adapted to country- and hospital-specific requirements.
- Together with SP12 (Delivery D71.2), SP8’s CHUV team finished the privacy assessment preparatory phase for the Hospital Deployment and created a hospital specific agreement subject to signature by the HBP’s MIP and the hospitals.

12. Innovation

12.1 MIP in the data sharing world

Following a proof-of-concept (PoC) that SP8 delivered to InTBIR Initiative, by which InTBIR data was integrated with the MIP platform for analysis, InTBIR decided to formalise the SP8/MIP presence in their project (the MoU awaits approval of HBP SIB) and strengthen the collaboration on data integration/share from Europe, US and Canada. The collaboration will be with the CReACTIVE⁸ consortium in collaboration with platforms such as BrainCode, FITBIR and OneMind, sharing data from studies like CIHR, NIH, 5P, ADAPT, TRACK-TBI, LABIC, TEAM-TBI, DoD etc.

The project is planning to have CReACTIVE and CENTER-TBI data integrated into the MIP by 1 September 2017, metadata by mid-September, and status of work (summary data, data pipeline description for exploration, analysis and collaboration) presented at end of October 2017.

12.1.1 Further collaboration initiatives, Pharma.

During SGA1 period, SP8 has made contact with and demonstrated the MIP to leaders from other similar platforms, such as DPUK and IMI. Discussions have also taken place with pharmaceutical companies, from which we got a very encouraging interest: Sanofi (France), C4X (UK) and Pharnext (France).

⁸ CReACTIVE (“Collaborative REsearch on ACute Traumatic brain Injury in intensive care medicine in Europe”) is a key partner in InTBIR, coordinated by GiViTI (the Italian Group for Intensive Care Evaluation).



13. Open Research Data

13.1 Open Research data delivered by the project

The vast amounts of software code, modules and libraries generated by the project team are available in public repositories (Github) and are described together with their links in their respective context throughout the document.

13.2 Open Research data to be made accessible by the project

The more data available in the Platform, the more accurate the models of brain diseases. Combining biological information from epidemiological, research and biobank cohorts ensures a high-level of precision for modelling and minimises existing selection bias, thus providing a better representation of the human population. The use of multiple datasets on multiple diseases is essential for assessing the generalisability of the disease models.

For the past few months, we have been discussing with several actors from the scientific and medical community for them to provide data to the feed into the MIP. So far, we have gathered information on 38 datasets from research cohort to be included in the MIP. The datasets from 13 European countries will be used in the Platform for research and clinical purposes.

The targeted datasets are rich, detailed and containing clinical data for a total of 362,745 patients. Each one focuses on different clinical and epidemiological characteristics relevant to the understanding of Ageing, Neurodegenerative diseases, Parkinson's disease, vascular brain damage and psychiatric disorders. Sixty-three percent of the records (n=251,637) contain longitudinal information which is important for modelling disease progression. The vast majority of records also have seven or more types of data (clinical, demographic, cognitive abilities, imaging, genetic, epidemiologic, other biological measurements) which is important for generating multi-modal biomarkers.

The full list of data sources is available online (excel file: http://hbps1.chuv.ch/files/Research_Datasets_of_interest_4_MIP_FBF_20032017.xlsx).

Here below is a description of the different fields.

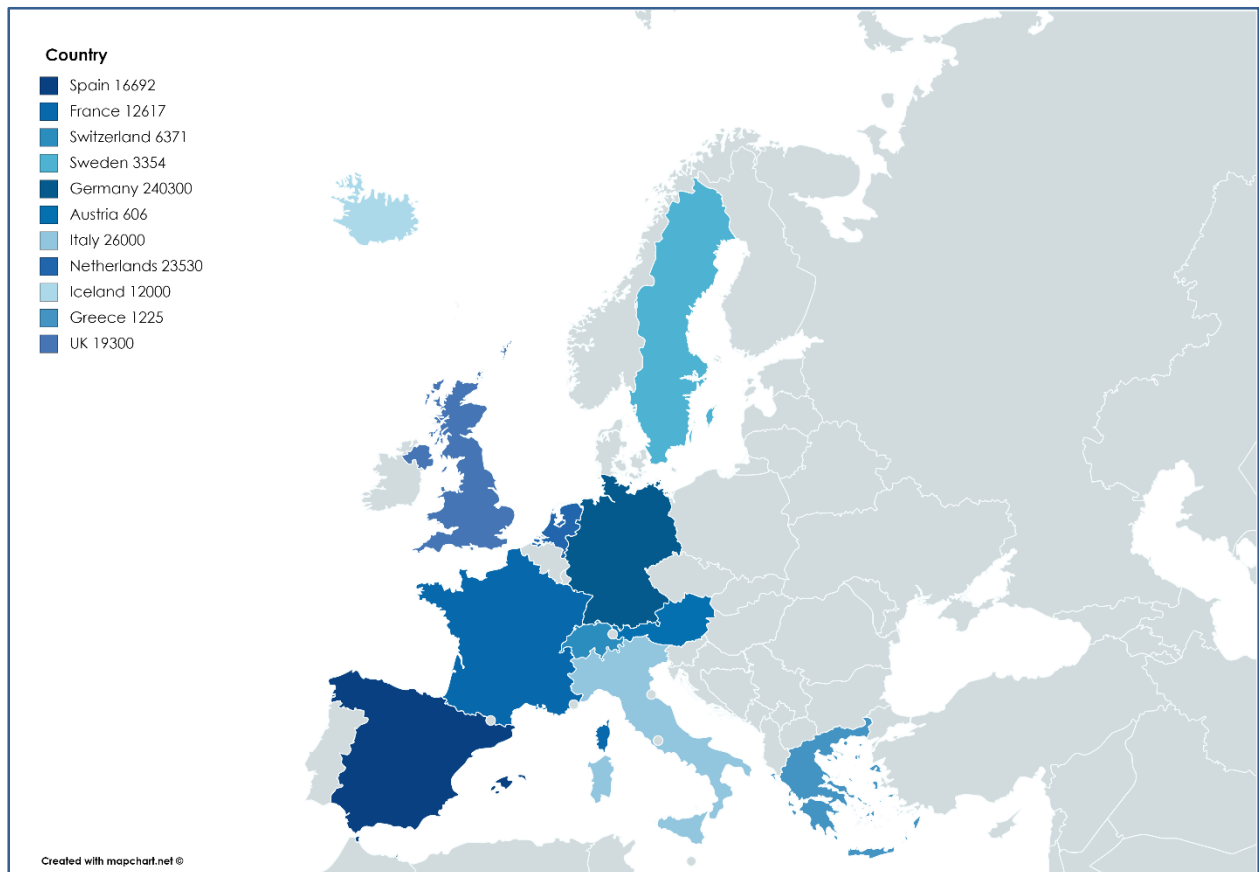


Figure 26: Number of datasets per country

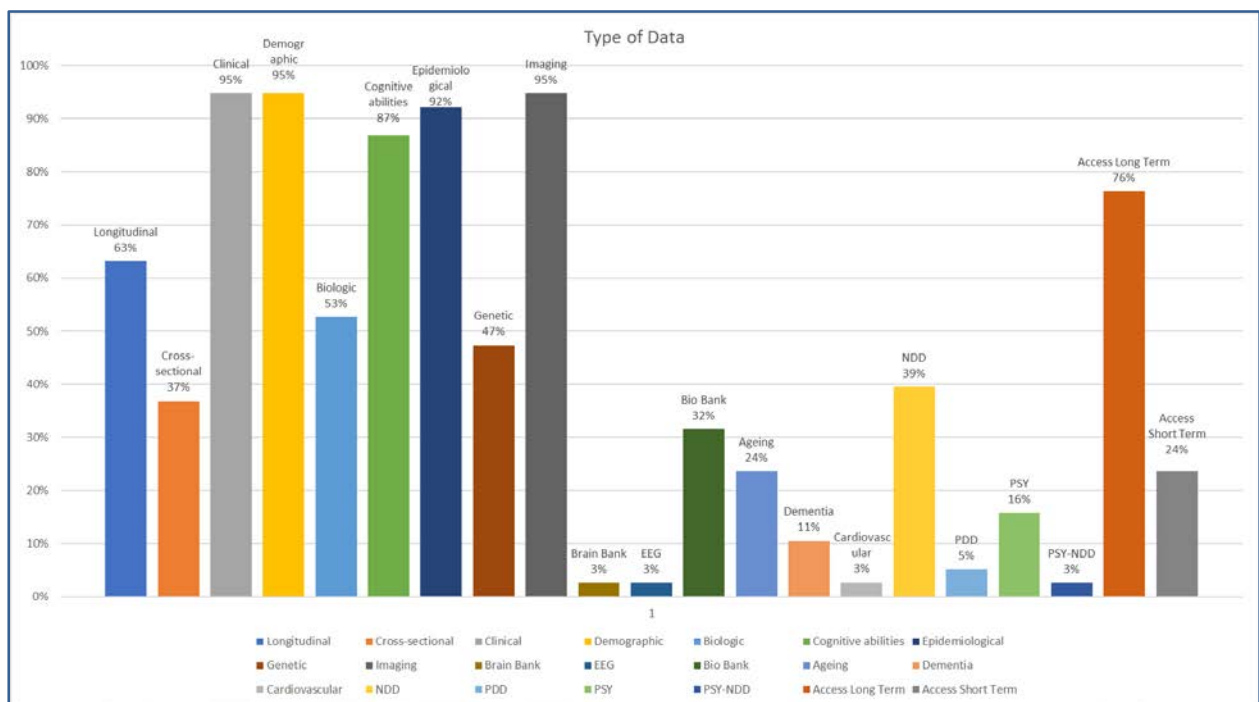


Figure 27: Types of data in percentage