Project Title:	Human Brain Project
Sub-Project Title:	Medical Informatics Platform (SP8)

Document Title:	MIP Services – User Guideline V1.0 Public release

Summary:	Step-by-step guidelines describing the main services of the Medical Informatics Platform (MIP) Audience: any users of Medical Informatics Platform

Versions, changes and contributors:

Authors	Version	Date	Status	Change Details
Elia Sbeiti (UNIGE / FBF)	V1	29/01/2016	Draft 1	Initial draft
Alberto Redolfi (UNIGE / FBF)	V1.1	29/01/2016	Draft 1.1	To be reviewed before the release

Human Brain Project

CHUV

TABLE OF CONTENTS

Web Portal: **The web portal** is the main user interface for the Medical Informatics Platform. It provides an easy-to-use interface that allows researchers to explore the data coming from research sources and the hospitals, build model, to perform online analytics on that data and to publish and share results.

User Documentation: **the User-Documnetation of the MIP (aka Knowledge Base)** is specially designed for the users with an easy way of accessing information (i.e. guidelines), provide feedbacks, and interact with the MIP developers and team. In the long run, the objective of the KB is to provide a flexible tool, fully integrated in the HBP COLAB platform, allowing the remote attendance of courses, tutorials, hand-on-sessions, forum, discussion boards, and ad-hoc quizzes to test the increase of knowledge of the MIP.

Data factory: The data factory includes **applications and** tools to start, execute, and monitor the feature extraction algorithms from MRI data (spatial registration and atlasing) and genetic data (plink tools).

Algorithm factory: The algorithm factory provides the tooling to execute of the self and custom data analytics algorithms written in a variety of languages and platforms. Algorithms are exposed as on-demand web services. At its core, the application today is composed of woken, an engine for on-demand analytics that launches Docker containers and collects their results, a few scripts in R encapsulated in Docker images and producing PFA models, and a web portal allowing the user to select data, build analytical model from a limited set of algorithms and view the results. The algorithms can be executed anywhere in the HPC and in the hospitals.

Information & Scientific reference Services include the services related to the meta-data or variable (ontology) as well as the scientific workflows, i.e.: the data used, the algorithms, the models created and the results that are saved as Research Objects (RO). The RO contain all the provenance information needed to reproduce the results. RO (http://www.researchobject.org/) are save a publication and shared by defaults. In addition to the RO specifications The standards for description of the results is based on PFA (PFA · Portable Format for Analytics - Data Mining Group dmg.org/pfa/). The semantic language for the description of the meta-data is based on RDF and Json-ld. Ontology services will be connected to the knowledge graph of the HBP Neuroinformatics Platform and other external resources using open-data and open-services.

1. Hospital Bundle: The Hospital Bundle is a software stack that will run at every participating hospital or medical center of the Federated Network of Hospitals and Centers (FNHC) of the Medical Informatics Subproject of the HBP. The MI Hospital Bundle will be installed at a hospital or center in order to enable contributing data to and using the Distributed Medical Informatics Platform (DMIP) through appropriate interfaces. The MI Hospital Bundle includes software for schema and data integration, in-situ querying, federated querying, and dataflow processing. Each hospital will be a node of the DMIP, and will communicate through appropriate modules in the hospital bundle with all other hospital nodes. The bundle consists of the following major components:

- **The in-situ query engine, RAW.** The engine serves as a database backend at each participating hospital and is responsible for executing queries on raw hospital data. The EPFL team is responsible for integrating all the components into a single software bundle and for overall testing of the bundle at participating hospitals (who will join during the ramp-up phase). RAW is a database system specially tailored for the needs of the HBP. Its main purpose is to offer efficient querying services directly on files inside the hospital. The type of research involved, biological signatures of diseases, requires more advanced data structures and more expressive operations than those provided by traditional databases.The query language enables users to apply powerful transformations over the output of a query. RAW uses a query language, similar to SQL (Structured Query Language) that provides support for a multitude of data models: collection types, hierarchies and multi-dimensional arrays. This flexibility enables queries to transparently access a great variety of datasets (e.g., relational tables, CSV and JSON files, array image data, etc.). Furthermore to be able to query from heterogeneous data formats, RAW utilizes code generation techniques. When a query is posed code generation plug-ins are invoked to produce code specific to both the file format and the query posed. Code generation acts as an enabler for queries targeting multiple data formats. In addition, it improves performance, as each query leads to execution of very specific code, avoiding generic methods that would take place (e.g., parsing the data files using a generic parser). RAW currently supports the following functionality:
    - Get list of the available schemas/ registered files patients, exams values, brain features
    - Submit queries and retrieve results
    - Submit paginated queries (streaming)
    - Register new files to be queried
- Schema Mapping and Data Exchange (MIPMap): MIPMap and WebMIPMap (see next bullet-point) provide interfaces for their users. MIPMap is tool that provides a user interface to data providers in order to allow them to translate their data to the MIP schema, while WebMIPMap provides a user interface to the MIP Web Portal users, allowing them to create mappings following the use cases described in the WebMIPMap description section. MIPMap is used in the Hospital Bundle as a declarative ETL tool (Extract, Transform, Load) that translates data provided by participating hospitals to the MIP schema, thus populating the Local Data Store Mirror of each hospital. Additionally it is utilized in the translation of the research data, used in the MIP, to also integrate them to the MIP schema, making them interoperable to other MIP data.

- The WebMIPMap is the web interface that can create mapping tasks which can later be downloaded and run on the desktop version, i.e. MIPMap. The use of WebMIPMap however is not restricted to the bundle. It is provided as a service in the MIP Web Portal to allow users to create mappings between the MIP schema and hospital research data schemata and/or ontologies. These mappings are used to translate terms of the MIP schema to hospital and research data schema terms in case their Local Data Store Mirror does not fully comply with the MIP schema. MIPMapRew is a service for rewriting queries posed at the Web Portal so that the nomenclature used by its predicates conforms to the schema of the hospital/research centres. Finally, WebMIPMap can be used to map MIP schema to existing ontologies.
- MIPMapReW is a service for rewriting queries posed at the Web Portal so that the nomenclature used by its predicates conforms to the schema of the hospital/research centres. As stated in the WebMIPMap description, it is highly likely that some hospitals/research centres might not fully comply with the MIP schema. In such cases WebMIPMap will be used to create the mappings between the two schemata and MIPMapRew will be responsible to rewrite the original query (created at the Web Portal) with the terms and nomenclature followed by the participating hospital, so that the appropriate results can be collected. Note, however that such a scenario will not be that common in the beginning of the Project, since according to the specifications of the MIP all the hospitals and research centres that are going to contribute data to the Platform will use the common MIP schema in order to make their data available.
- **EXAREME is a distributed processing engine, with a declarative language based on SQL with user-defined functions (UDFs) extended with parallelism and data pipeline primitives. In the context of MIP bundle, Exareme acts as the federation layer, responsible for the communication of each hospital and web portal. It does not allow communication amongst hospitals. Worker components are deployed on each hospital node and act as connectors with the RAW query engines. The Master component merges the partial hospital results and can be deployed in one or more hospital nodes.** Currently, Exareme is integrated with RAW and executes a variety of distributed algorithms in a privacy preserving manner.

- **The anonymization and query filter module, i.e.: the Anonymizer.** The goal of the anonymizer module deployed at a hospital is to (a) strip all data exported from the hospital's systems from personal identifiers and (b) control that incoming queries conform to privacy standards, i.e., the columns they read are limited and (c) strip all query results from personal identifiers as well. More precisely, identifiers are initially removed when exporting data from hospital information systems, i.e., even before the data is accessed by the medical informatics platform for the first time. In this process all personal identifiers are stripped from the data. Second, to ensure no personal information leaves the hospitals, incoming queries are checked to ensure they only request fields made available to the federated platform. Finally, to further reduce the information available about individual patients, the platform filters all results (ensuring they do not contain any personal patient information) before returning them to the users of the platform. Based on the above, the subcontractor (Gnubila) has developed the anonymization module including the following features:
    - Anonymizer: blacklist/whitelist manager, DICOM anonymizer, full text anonymizer.
    - Query Filter: Query Parser, Black Listed Fields exclusion, Aggregation Field check.
    - Response Cleaner: Privacy Information Webservice, Cleaner Process.
- **Federated Workflow:** Coming from the web portal, queries and analyses will first run through an Exareme instance (at any hospital), which distributes them to other Exareme instances (at other hospitals). At every hospital, the queries will run through the query filter to ensure that only allowed queries are passed on. If allowed, the queries are forwarded to RAW that is used to access the hospital data in situ. Once RAW returns the aggregated results, they are sent back to Exareme, which collects all results from different hospitals. As only aggregated results leave hospitals, the patient data privacy is preserved.

However, the data that are accessed by RAW (hospital LDSM) are not the original data provided by the hospital. The original hospital data have to undergo a process where they are initially anonymized and in the following integrated to the MIP schema to create the hospital LDSM. The anonymization component connects to the hospital data sources, and anonymizes the data. This happens for security reason as this server can be in a more restricted network, data can only be pushed ensuring that only anonymized data is accessed by the system. Then MIPMap integrates these data to the MIP schema, using a data exchange process, creating the Local Data Store Mirror of the hospital and allowing RAW to query them. The diagram below shows this federated workflow.

Human Brain Project