# SOFTWARE

## Data Factory (DF)

### Feature Engineering

**Contributing task:** T8.3.11 Brain morphological features UCL  
**Description:** A privacy preserving approach for the generalised principal component analysis of large image datasets.  
**Dependencies:**

|  |  |  |
| --- | --- | --- |
| DATA | Hospital | Lille Hospital, Tel Aviv Hospital, Milano Hospital, Freiburg Hospital, CHUV Hospital |
|  | Reference | TBI (Traumatic Brain Injury), PPMI (Parkinson's Progression Markers Initiative), ADNI |
| SOFTWARE | Data Factory (DF) | Data Anonymisation, Workflow Engine, Data Pipeline processes, Data Quality Processes, Data Storage |

**Releases:**

|  |  |
| --- | --- |
| Planned functionalities at M18 | Further testing and code optimisation of image factorisation method (handle missing data, make more probabilistic, make computationally stable, etc). |
| Planned functionality at M12 | Initial implementation of image factorisation method without distributed computing (MS126). |
| Planned functionalities at M24 | A functioning distributed implementation of the image factorisation method, with features obtained by the approach available for data mining. |

**User and Use cases:**

|  |  |
| --- | --- |
| User | Use cases |
| Neuroscientist Clinical Researcher Developer - Software Developer - Methods SP8 Platform Developer | Compute a PCA-like analysis over thousands of 3D patient scans distributed across several hospitals, in situations where some parts of the data may be missing. Principal components serve as features for data mining. |

# SOFTWARE

## Data Factory (DF)

### Data Integration

**Contributing task:** T8.1.4 Data Integration AUEB  
**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 LDSM.  
**Dependencies:**

|  |  |  |
| --- | --- | --- |
| SOFTWARE | Hospital Databases Bundle (HDB) | Schema Mapping |

**Releases:**

|  |  |
| --- | --- |
| Planned functionalities at M18 | Query answering without rewriting with respect to database sources schema |
| Planned functionality at M12 | Rewriting of query with respect to ontology module and preliminary integration to the platform |
| Planned functionalities at M24 | Final version - Optimized query answering - and full integration to the platform |

**User and Use cases:**

|  |  |
| --- | --- |
| User | Use cases |
|  | Consider the SNOMED ontology and the concepts Dementia, Alzheimer Dementia and Parkinson Dementia. The SNOMED taxonomy defines that the concept of Alzheimer Dementia and of Parkinson Dementia are subsumed by the concept Dementia meaning that all Alzheimer/Parkinson Dementias are also Dementias. Moreover consider that the concept Alzheimer Dementia is mapped to the variable (of the MIP schema) Diagnosis with value specific to 'AD' and Parkinson Dementia is mapped to the variable Diagnosis with value specific to 'PR'. So, when a user poses a query using the SNOMED ontology that retrieves all patients with dementia, the OBDA system would rewrite this query to also return patients with Alzheimer disease and patients with Parkinson disease and finally would translate those queries to the MIP schema so that every patient with 'AD' and 'PR' value would return. |

**Contributing task:** T8.1.4 Data Integration AUEB  
**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 anonymization 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.  
**Dependencies:**

|  |  |  |
| --- | --- | --- |
| SOFTWARE | Data Factory (DF) | Data Integration |
|  | Hospital Databases Bundle (HDB) | Schema Mapping |

**Releases:**

|  |  |
| --- | --- |
| Planned functionalities at M18 | Preliminary version of module |
| Planned functionality at M12 | Report on Incremental Data Integration |
| Planned functionalities at M24 | Final version and integration to the Hospital Bundle |

**User and Use cases:**

|  |  |
| --- | --- |
| User | Use cases |
|  | A hospital participates in the MIP by providing a first batch of data. Consider that there exists a rule stating that the target diagnostic table is populated only with patients (source patient table) who have a diagnosis (source diagnostic table). Moreover consider that in this first data batch patient 'A' did not have any examinations, hence the target diagnostics table does not contain any tuple for 'A'. Consider now that in a second batch, a month later, diagnostics for patient 'A' are entered but patient information for 'A' is not resent (as it had been sent in the first batch a month earlier). In order to populate the target diagnostic table with information for 'A' it would be essential to run again the data exchange process for all the input data (first and second batch). However, this poses a significant computational burden. Hence, it would be ideal if an incremental data exchange approach would be used that would not need to re-run the data translation process from scratch but would be able to generate the information for patient 'A' and her diagnostics as they became available. |

**Contributing task:** T8.1.4 Data Integration AUEB  
**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.  
**Dependencies:**

|  |  |  |
| --- | --- | --- |
| SOFTWARE | Hospital Databases Bundle (HDB) | Schema Mapping |

**Releases:**

|  |  |
| --- | --- |
| Planned functionalities at M18 | Query answering without rewriting with respect to database sources schema |
| Planned functionality at M12 | Rewriting of query with respect to ontology module and preliminary integration to the platform |
| Planned functionalities at M24 | Final version - Optimized query answering - and full integration to the platform |

**User and Use cases:**

|  |  |
| --- | --- |
| User | Use cases |
|  | Consider the SNOMED ontology and the concepts Dementia, Alzheimer Dementia and Parkinson Dementia. The SNOMED taxonomy defines that the concept of Alzheimer Dementia and of Parkinson Dementia are subsumed by the concept Dementia meaning that all Alzheimer/Parkinson Dementias are also Dementias. Moreover consider that the concept Alzheimer Dementia is mapped to the variable (of the MIP schema) Diagnosis with value specific to 'AD' and Parkinson Dementia is mapped to the variable Diagnosis with value specific to 'PR'. So, when a user poses a query using the SNOMED ontology that retrieves all patients with dementia, the OBDA system would rewrite this query to also return patients with Alzheimer disease and patients with Parkinson disease and finally would translate those queries to the MIP schema so that every patient with 'AD' and 'PR' value would return. |

**Contributing task:** T8.1.4 Data Integration AUEB  
**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 anonymization 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.  
**Dependencies:**

|  |  |  |
| --- | --- | --- |
| SOFTWARE | Data Factory (DF) | Data Integration |
|  | Hospital Databases Bundle (HDB) | Schema Mapping |

**Releases:**

|  |  |
| --- | --- |
| Planned functionalities at M18 | Preliminary version of module |
| Planned functionality at M12 | Report on Incremental Data Integration |
| Planned functionalities at M24 | Final version and integration to the Hospital Bundle |

**User and Use cases:**

|  |  |
| --- | --- |
| User | Use cases |
|  | A hospital participates in the MIP by providing a first batch of data. Consider that there exists a rule stating that the target diagnostic table is populated only with patients (source patient table) who have a diagnosis (source diagnostic table). Moreover consider that in this first data batch patient 'A' did not have any examinations, hence the target diagnostics table does not contain any tuple for 'A'. Consider now that in a second batch, a month later, diagnostics for patient 'A' are entered but patient information for 'A' is not resent (as it had been sent in the first batch a month earlier). In order to populate the target diagnostic table with information for 'A' it would be essential to run again the data exchange process for all the input data (first and second batch). However, this poses a significant computational burden. Hence, it would be ideal if an incremental data exchange approach would be used that would not need to re-run the data translation process from scratch but would be able to generate the information for patient 'A' and her diagnostics as they became available. |

**Contributing task:** T8.1.4 Data Integration AUEB  
**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 standardization 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. The component can be used to accelerate the creation of the Knowledge Graph of SP5.  
**Dependencies:**

|  |  |  |
| --- | --- | --- |
| SOFTWARE | Hospital Databases Bundle (HDB) | Schema Mapping |

**Releases:**

|  |  |
| --- | --- |
| Planned functionalities at M18 | Ability to merge and extend existing schemata and create mappings for the overall scenario |
| Planned functionality at M12 | Creation of Global Mappings from privileged users and users' ability to endorse them |
| Planned functionalities at M24 | Visibility of user defined mappings to other users - 'friend' (or 'user I trust') feature and integration to the Web Portal |

**User and Use cases:**

|  |  |
| --- | --- |
| User | Use cases |
|  | Use case 1: Consider an administrator that creates a mapping of some variables of a research project schema to the MIP schema, and shares this schema with the MIP users, thus making it global. Consider now that a MIP user (with no administrative privileges) would like to create a mapping for additional variables of the same schemata/ontologies. It would be beneficial for the user to be able to load the global administrator's schema and consult its correspondences (either endorse them or ignore them) to create her own mapping. Use case 2: Users can share their mappings of novel research data schemas with other users ('friends') and/or endorse their mappings, to further data sharing. |

# SOFTWARE

## Data Factory (DF)

### Data Anonymisation

**Contributing task:** T8.1.3 Installation and adaptation of SP8 bundle at local hospitals with upgrade support EPFL  
**Description:** Local hospital anonymization validation tests.  
**Dependencies:**

|  |  |  |
| --- | --- | --- |
| DATA | Hospital | Lille Hospital, Tel Aviv Hospital, Milano Hospital, Freiburg Hospital, CHUV Hospital |
| SOFTWARE | Data Factory (DF) | Data Storage |

**Releases:**

|  |  |
| --- | --- |
| Planned functionalities at M18 | Anonymization process validated in three hospitals. |
| Planned functionality at M12 | Anonymization process validated in one hospital. |
| Planned functionalities at M24 | Anonymization process validated in all five hospitals. |

**User and Use cases:**

|  |  |
| --- | --- |
| User | Use cases |
| Developer - Software | Show validation report to the hospitals in order to build a trusting collaboration. |

# SOFTWARE

## Data Factory (DF)

### Workflow Engine

**Contributing task:** T8.1.3 Installation and adaptation of SP8 bundle at local hospitals with upgrade support EPFL  
**Description:** A component that creates a secure network between hospitals for the administration of the servers.  
**Dependencies:**

|  |  |  |
| --- | --- | --- |
| SOFTWARE | Data Factory (DF) | Data Anonymisation, Data Storage |
|  | Hospital Databases Bundle (HDB) | Local Database, Federated Query, Schema Mapping, Hospital Bundle Package |
| SERVICES | Security & Monitoring | Data governance |

**Releases:**

|  |  |
| --- | --- |
| Planned functionalities at M18 | Initial proof-of-concept developed. |
| Planned functionality at M12 | Technology defined. |
| Planned functionalities at M24 | The component is integrated with the bundle in the local hospital. |

**User and Use cases:**

|  |  |
| --- | --- |
| User | Use cases |
| Developer - Software | The system administrator can monitor securely the hospital bundle servers. |

**Contributing task:** T8.1.3 Installation and adaptation of SP8 bundle at local hospitals with upgrade support EPFL  
**Description:** This component will enable the local hospital services to be remotely managed (ie start, stop)  
**Dependencies:**

|  |  |  |
| --- | --- | --- |
| SOFTWARE | Hospital Databases Bundle (HDB) | Local Database, Federated Query, Schema Mapping, Hospital Bundle Package |
| SERVICES | Security & Monitoring | Data governance |

**Releases:**

|  |  |
| --- | --- |
| Planned functionalities at M18 | Initial proof-of-concept developed. |
| Planned functionality at M12 | dependent on delivery of: SOFTWARE > Data Factory (DF) > Workflow Engine > Encrypted overlay network Technology defined. |
| Planned functionalities at M24 | The component is integrated with the bundle in the local hospital. |

**User and Use cases:**

|  |  |
| --- | --- |
| User | Use cases |
| Developer - Software | The system administrator can manage securely the hospital bundle servers. |

# SOFTWARE

## Data Factory (DF)

### HPC

**Contributing task:** T8.4.5 Large-scale data analytics on massively parallel architecture ICL  
**Description:** This component essentially allows a user to upload (medical) data to the supercomputing infrastructure. It will either be implemented and made available through command line scripts (possibly Python) or a web page.  
**Dependencies:**

|  |  |  |
| --- | --- | --- |
| SOFTWARE | Algorithm Library | Statistical Analytics, Machine Learning Library |

**Releases:**

|  |  |
| --- | --- |
| Planned functionalities at M18 | none |
| Planned functionality at M12 | This component should be fully available at M12 (pending collaboration with SP7) |
| Planned functionalities at M24 | none |

**User and Use cases:**

|  |  |
| --- | --- |
| User | Use cases |
| Neuroscientist Clinical Researcher | Clustering of high-dimensional medical data. |

**Contributing task:** T8.4.5 Large-scale data analytics on massively parallel architecture ICL  
**Description:** THis component will clean, reformat and distribute the data in the supercomputing infrastructure. It will be based on scripts and will connect to the uploading component.  
**Dependencies:**

|  |  |  |
| --- | --- | --- |
| SOFTWARE | Algorithm Library | Statistical Analytics, Machine Learning Library |

**Releases:**

|  |  |
| --- | --- |
| Planned functionalities at M18 | none |
| Planned functionality at M12 | This component should be fully available at M12 (pending collaboration with SP7) |
| Planned functionalities at M24 | none |

**User and Use cases:**

|  |  |
| --- | --- |
| User | Use cases |
| Neuroscientist Clinical Researcher | Clustering of high-dimensional medical data. |

**Contributing task:** T8.4.5 Large-scale data analytics on massively parallel architecture ICL  
**Description:** This component will develop and deploy a library (containing multiple clustering/classification/machine learning algorithms) for data analaysis on the distributed/supercomputing infrastructure. It will be based on MPI and most likely implemented in C++. It will contain several algotihms including approximate ones. The component connects to the data uploader and the cleaner.  
**Dependencies:**

|  |  |  |
| --- | --- | --- |
| SOFTWARE | Algorithm Library | Statistical Analytics, Machine Learning Library |

**Releases:**

|  |  |
| --- | --- |
| Planned functionalities at M18 | none |
| Planned functionality at M12 | none |
| Planned functionalities at M24 | This component should be fully available at M24 (pending collaboration with SP7) |

**User and Use cases:**

|  |  |
| --- | --- |
| User | Use cases |
| Neuroscientist Clinical Researcher | Clustering of high-dimensional medical data. |

**Contributing task:** T8.4.5 Large-scale data analytics on massively parallel architecture ICL  
**Description:** This component will enable users to download analysis results. It will either be available as scripts on the command line or as a web page. It depends on the analytics library and thus connects to it.  
**Dependencies:**

|  |  |  |
| --- | --- | --- |
| SOFTWARE | Algorithm Library | Statistical Analytics, Machine Learning Library |

**Releases:**

|  |  |
| --- | --- |
| Planned functionalities at M18 | none |
| Planned functionality at M12 | none |
| Planned functionalities at M24 | This component should be fully available at M24 (pending collaboration with SP7) |

**User and Use cases:**

|  |  |
| --- | --- |
| User | Use cases |
| Neuroscientist Clinical Researcher | Clustering of high-dimensional medical data. |

# SOFTWARE

## Hospital Databases Bundle (HDB)

### Local Database

**Contributing task:** T8.1.4 Data Integration AUEB  
**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  
**Dependencies:**

|  |  |  |
| --- | --- | --- |
| SOFTWARE | Hospital Databases Bundle (HDB) | Local Database |
| SERVICES | Security & Monitoring | User Management |

**Releases:**

|  |  |
| --- | --- |
| Planned functionalities at M18 | Basic access control functionality |
| Planned functionality at M12 | Report on access restrictions in the MIP |
| Planned functionalities at M24 | Final release of ARM and integration to the platform |

**User and Use cases:**

|  |  |
| --- | --- |
| User | Use cases |
|  | Consider that a MIP user would like to run an EE query and identify the number of patients diagnosed with dementia. According to the access rules imposed by the MIP administrators only Advanced users (AU) are able to run queries that invoke aggregate data, while Simple users (SU) are only able to use the already pre-computed and stored aggregate results. Consider also an AU, who is a pathologist, and thus should not get access to brain imaging data to compute aggregates. All these restrictions can be implemented in the MIP using specific access control languages that also affect the way queries posed by users are reformulated. |

**Contributing task:** T8.1.1 Infrastructure to support just-in-time analytics on raw medical data EPFL  
**Description:** Basic primitives for computation over multidimensional queries in the local hospital database. This component needs data but not necessarily data from ALL hospitals.  
**Dependencies:**

|  |  |  |
| --- | --- | --- |
| DATA | Hospital | Lille Hospital, Tel Aviv Hospital, Milano Hospital, Freiburg Hospital, CHUV Hospital |
| SOFTWARE | Data Factory (DF) | Data Anonymisation, Workflow Engine, Data Pipeline processes, Data Quality Processes, Data Storage, Data Integration |
|  | Hospital Databases Bundle (HDB) | Schema Mapping |
| SERVICES | Upgrade-Deploy-Release | MIP Integrated Releases Hospital Databases Bundle |
|  | Security & Monitoring | Data governance |

**Releases:**

|  |  |
| --- | --- |
| Planned functionalities at M18 | The plug-in is completed and deployed as part of the Hospital Bundle. |
| Planned functionality at M12 | The infrastructure to run this plug-in in the query engine will be ready. |
| Planned functionalities at M24 | The plug-in is completed and deployed as part of the Hospital Bundle and validated by HBP users. |

**User and Use cases:**

|  |  |
| --- | --- |
| User | Use cases |
| SP8 Platform Developer | Allow ad-hoc dimensionality reduction and feature extraction, e.g. on imaging and genetic data. It can be later added to the data factory as the regular feature extraction pipelines. |

**Contributing task:** T8.1.2 Installation of RAW on local computing infrastructure for hospital analytics requirements EPFL  
**Description:** Extend the local query engine to enable the use of distributed computing frameworks (like Spark).  
**Dependencies:**

|  |  |  |
| --- | --- | --- |
| DATA | Hospital | Lille Hospital, Tel Aviv Hospital, Milano Hospital, Freiburg Hospital, CHUV Hospital |
| SOFTWARE | Data Factory (DF) | Data Anonymisation, Workflow Engine, Data Pipeline processes, Data Storage, Data Integration, HPC |
|  | Hospital Databases Bundle (HDB) | Schema Mapping, Hospital Bundle Package |
| SERVICES | Upgrade-Deploy-Release | Micro-services |
|  | Security & Monitoring | Data governance |

**Releases:**

|  |  |
| --- | --- |
| Planned functionalities at M18 | The federation software is integrated with the new and improved query engine API's . |
| Planned functionality at M12 | The query engine extensions are implemented. |
| Planned functionalities at M24 | Query engine optimized using real-world datasets and queries. |

**User and Use cases:**

|  |  |
| --- | --- |
| User | Use cases |
| Developer - Software | Allow for scalability of the query engine for hospitals with big amounts of data. |

**Contributing task:** T8.1.1 Infrastructure to support just-in-time analytics on raw medical data EPFL  
**Description:** Plug-in to enable local query engine to perfom queries directly on Nifti files.  
**Dependencies:**

|  |  |  |
| --- | --- | --- |
| SOFTWARE | Data Factory (DF) | Data Anonymisation, Data Pipeline processes, Data Storage, Data Integration |

**Releases:**

|  |  |
| --- | --- |
| Planned functionalities at M18 | Initial proof-of-concept query engine that can perform more complex queries like joins and group by over Nifti files. |
| Planned functionality at M12 | Initial proof-of-concept query engine that can perform selection queries over Nifti files. |
| Planned functionalities at M24 | Query engine that can perform complex queries and mathematical operations over Nifti files. |

**User and Use cases:**

|  |  |
| --- | --- |
| User | Use cases |
| Developer - Software | Data discovery directly on Nifti files Enable analysis over imaging data that are not provided directly by the existing pipeline. |

**Contributing task:** T8.1.1 Infrastructure to support just-in-time analytics on raw medical data EPFL  
**Description:** Plug-in to enable local query engine to perfom queries directly on genetic data files.  
**Dependencies:**

|  |  |  |
| --- | --- | --- |
| SOFTWARE | Hospital Databases Bundle (HDB) | Local Database, Schema Mapping |
|  | Algorithm Factory (AF) | Workflow Engine: Woken, Model Scoring |

**Releases:**

|  |  |
| --- | --- |
| Planned functionalities at M18 | Initial proof-of-concept query engine that can perform more complex queries like joins and group by over genetic data files. |
| Planned functionality at M12 | Initial proof-of-concept query engine that can perform selection queries over genetic data files. |
| Planned functionalities at M24 | Query engine that can perform complex queries and mathematical operations over genetic data files. |

**User and Use cases:**

|  |  |
| --- | --- |
| User | Use cases |
| Developer - Software | Data discovery directly on genetic data files Enable analysis over genetic data that are not provided directly by the existing pipeline. |

**Contributing task:** T8.1.1 Infrastructure to support just-in-time analytics on raw medical data EPFL  
**Description:** Library of functions for common operations on imaging data/Nifti files in the local query engine.  
**Dependencies:**

|  |  |  |
| --- | --- | --- |
| SOFTWARE | Data Factory (DF) | Data Anonymisation, Data Pipeline processes, Data Storage, Data Integration |

**Releases:**

|  |  |
| --- | --- |
| Planned functionalities at M18 | Initial proof-of-concept with a few functions defined in collaboration with Data Analysis groups. |
| Planned functionality at M12 | dependent on delivery of: SOFTWARE > Hospital Databases Bundle (HDB) > Local Database > Nifti data source for local query engine Definition of functionality requirements. This component needs the Nifti data source to be developed first. Therefore the development effort will be focused there for the period of M01-M12. |
| Planned functionalities at M24 | Final Nifti library with a satisfactory number of functions. |

**User and Use cases:**

|  |  |
| --- | --- |
| User | Use cases |
| Developer - Software | Data discovery directly on Nifti files Enable analysis over imaging data that are not provided directly by the existing pipeline. |

**Contributing task:** T8.1.1 Infrastructure to support just-in-time analytics on raw medical data EPFL  
**Description:** Library of functions for common operations on genetic data files in the local query engine.  
**Dependencies:**

|  |  |  |
| --- | --- | --- |
| SOFTWARE | Data Factory (DF) | Data Anonymisation, Data Pipeline processes, Data Storage, Data Integration |

**Releases:**

|  |  |
| --- | --- |
| Planned functionalities at M18 | Initial proof-of-concept with a few functions defined in collaboration with Data Analysis groups. |
| Planned functionality at M12 | dependent on delivery of: SOFTWARE > Hospital Databases Bundle (HDB) > Local Database > Genetic data data source for local query engine Definition of functionality requirements. This component needs the genetic data source to be developed first. Therefore the development effort will be focused there for the period of M01-M12. |
| Planned functionalities at M24 | Final genetic data library with a satisfactory number of functions. |

**User and Use cases:**

|  |  |
| --- | --- |
| User | Use cases |
| Developer - Software | Data discovery directly on genetic data files Enable analysis over genetic data that are not provided directly by the existing pipeline. |

**Contributing task:** T8.1.2 Installation of RAW on local computing infrastructure for hospital analytics requirements EPFL  
**Description:** Establish encrypted connections between two hospitals.  
**Dependencies:**

|  |  |  |
| --- | --- | --- |
| DATA | Hospital | Lille Hospital, Tel Aviv Hospital, Milano Hospital, Freiburg Hospital, CHUV Hospital |
| SOFTWARE | Data Factory (DF) | Data Anonymisation |
| SERVICES | Security & Monitoring | Data governance |

**Releases:**

|  |  |
| --- | --- |
| Planned functionalities at M18 | Initial proof-of-concept using this component is developed. |
| Planned functionality at M12 | The technical procedure describing the secure connection and which technologies to use is documented. |
| Planned functionalities at M24 | The component is integrated in the hospital bundle and/or data factory. |

**User and Use cases:**

|  |  |
| --- | --- |
| User | Use cases |
| Developer - Software | Secure connection to enable hospital hubs. |

**Contributing task:** T8.1.2 Installation of RAW on local computing infrastructure for hospital analytics requirements EPFL  
**Description:** This component will enable hospitals by finding a way to share processing between hospitals.  
**Dependencies:**

|  |  |  |
| --- | --- | --- |
| SOFTWARE | Hospital Databases Bundle (HDB) | Federated Query, Schema Mapping, Hospital Bundle Package |
| SERVICES | Security & Monitoring | Data governance |

**Releases:**

|  |  |
| --- | --- |
| Planned functionalities at M18 | Initial proof-of-concept developed. |
| Planned functionality at M12 | dependent on delivery of: SOFTWARE > Hospital Databases Bundle (HDB) > Local Database > Secure connection between two hospitals. Technology defined. |
| Planned functionalities at M24 | The component is integrated with the bundle in the local hospital. |

**User and Use cases:**

|  |  |
| --- | --- |
| User | Use cases |
| Developer - Software | Hospital with less resources can off load some of their processes to another hospital/hub. |

**Contributing task:** T8.1.4 Data Integration AUEB  
**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.  
**Dependencies:**

|  |  |  |
| --- | --- | --- |
| SOFTWARE | Hospital Databases Bundle (HDB) | Local Database |
|  | Web Exploration and Analytics | Data Access |
| SERVICES | Security & Monitoring | User Management |

**Releases:**

|  |  |
| --- | --- |
| Planned functionalities at M18 | Basic access control functionality |
| Planned functionality at M12 | Report on access restrictions in the MIP |
| Planned functionalities at M24 | Final release of ARM and integration to the platform |

**User and Use cases:**

|  |  |
| --- | --- |
| User | Use cases |
|  | Consider that a MIP user would like to run an EE query and identify the number of patients diagnosed with dementia. According to the access rules imposed by the MIP administrators only Advanced users (AU) are able to run queries that invoke aggregate data, while Simple users (SU) are only able to use the already pre-computed and stored aggregate results. Consider also an AU, who is a pathologist, and thus should not get access to brain imaging data to compute aggregates. All these restrictions can be implemented in the MIP using specific access control languages that also affect the way queries posed by users are reformulated. |

# SOFTWARE

## Hospital Databases Bundle (HDB)

### Schema Mapping

**Contributing task:** T8.1.4 Data Integration AUEB  
**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 LDSM.  
**Dependencies:**

|  |  |  |
| --- | --- | --- |
| SOFTWARE | Hospital Databases Bundle (HDB) | Schema Mapping |

**Releases:**

|  |  |
| --- | --- |
| Planned functionalities at M18 | Query answering without rewriting with respect to database sources schema |
| Planned functionality at M12 | Rewriting of query with respect to ontology module and preliminary integration to the platform |
| Planned functionalities at M24 | Final version - Optimized query answering - and full integration to the platform |

**User and Use cases:**

|  |  |
| --- | --- |
| User | Use cases |
|  | Consider the SNOMED ontology and the concepts Dementia, Alzheimer Dementia and Parkinson Dementia. The SNOMED taxonomy defines that the concept of Alzheimer Dementia and of Parkinson Dementia are subsumed by the concept Dementia meaning that all Alzheimer/Parkinson Dementias are also Dementias. Moreover consider that the concept Alzheimer Dementia is mapped to the variable (of the MIP schema) Diagnosis with value specific to 'AD' and Parkinson Dementia is mapped to the variable Diagnosis with value specific to 'PR'. So, when a user poses a query using the SNOMED ontology that retrieves all patients with dementia, the OBDA system would rewrite this query to also return patients with Alzheimer disease and patients with Parkinson disease and finally would translate those queries to the MIP schema so that every patient with 'AD' and 'PR' value would return. |

**Contributing task:** T8.1.4 Data Integration AUEB  
**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 anonymization 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.  
**Dependencies:**

|  |  |  |
| --- | --- | --- |
| SOFTWARE | Data Factory (DF) | Data Integration |
|  | Hospital Databases Bundle (HDB) | Schema Mapping |

**Releases:**

|  |  |
| --- | --- |
| Planned functionalities at M18 | Preliminary version of module |
| Planned functionality at M12 | Report on Incremental Data Integration |
| Planned functionalities at M24 | Final version and integration to the Hospital Bundle |

**User and Use cases:**

|  |  |
| --- | --- |
| User | Use cases |
|  | A hospital participates in the MIP by providing a first batch of data. Consider that there exists a rule stating that the target diagnostic table is populated only with patients (source patient table) who have a diagnosis (source diagnostic table). Moreover consider that in this first data batch patient 'A' did not have any examinations, hence the target diagnostics table does not contain any tuple for 'A'. Consider now that in a second batch, a month later, diagnostics for patient 'A' are entered but patient information for 'A' is not resent (as it had been sent in the first batch a month earlier). In order to populate the target diagnostic table with information for 'A' it would be essential to run again the data exchange process for all the input data (first and second batch). However, this poses a significant computational burden. Hence, it would be ideal if an incremental data exchange approach would be used that would not need to re-run the data translation process from scratch but would be able to generate the information for patient 'A' and her diagnostics as they became available. |

**Contributing task:** T8.1.4 Data Integration AUEB  
**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.  
**Dependencies:**

|  |  |  |
| --- | --- | --- |
| SOFTWARE | Hospital Databases Bundle (HDB) | Schema Mapping |

**Releases:**

|  |  |
| --- | --- |
| Planned functionalities at M18 | Query answering without rewriting with respect to database sources schema |
| Planned functionality at M12 | Rewriting of query with respect to ontology module and preliminary integration to the platform |
| Planned functionalities at M24 | Final version - Optimized query answering - and full integration to the platform |

**User and Use cases:**

|  |  |
| --- | --- |
| User | Use cases |
|  | Consider the SNOMED ontology and the concepts Dementia, Alzheimer Dementia and Parkinson Dementia. The SNOMED taxonomy defines that the concept of Alzheimer Dementia and of Parkinson Dementia are subsumed by the concept Dementia meaning that all Alzheimer/Parkinson Dementias are also Dementias. Moreover consider that the concept Alzheimer Dementia is mapped to the variable (of the MIP schema) Diagnosis with value specific to 'AD' and Parkinson Dementia is mapped to the variable Diagnosis with value specific to 'PR'. So, when a user poses a query using the SNOMED ontology that retrieves all patients with dementia, the OBDA system would rewrite this query to also return patients with Alzheimer disease and patients with Parkinson disease and finally would translate those queries to the MIP schema so that every patient with 'AD' and 'PR' value would return. |

**Contributing task:** T8.1.4 Data Integration AUEB  
**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 anonymization 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.  
**Dependencies:**

|  |  |  |
| --- | --- | --- |
| SOFTWARE | Data Factory (DF) | Data Integration |
|  | Hospital Databases Bundle (HDB) | Schema Mapping |

**Releases:**

|  |  |
| --- | --- |
| Planned functionalities at M18 | Preliminary version of module |
| Planned functionality at M12 | Report on Incremental Data Integration |
| Planned functionalities at M24 | Final version and integration to the Hospital Bundle |

**User and Use cases:**

|  |  |
| --- | --- |
| User | Use cases |
|  | A hospital participates in the MIP by providing a first batch of data. Consider that there exists a rule stating that the target diagnostic table is populated only with patients (source patient table) who have a diagnosis (source diagnostic table). Moreover consider that in this first data batch patient 'A' did not have any examinations, hence the target diagnostics table does not contain any tuple for 'A'. Consider now that in a second batch, a month later, diagnostics for patient 'A' are entered but patient information for 'A' is not resent (as it had been sent in the first batch a month earlier). In order to populate the target diagnostic table with information for 'A' it would be essential to run again the data exchange process for all the input data (first and second batch). However, this poses a significant computational burden. Hence, it would be ideal if an incremental data exchange approach would be used that would not need to re-run the data translation process from scratch but would be able to generate the information for patient 'A' and her diagnostics as they became available. |

**Contributing task:** T8.1.4 Data Integration AUEB  
**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 standardization 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. The component can be used to accelerate the creation of the Knowledge Graph of SP5.  
**Dependencies:**

|  |  |  |
| --- | --- | --- |
| SOFTWARE | Hospital Databases Bundle (HDB) | Schema Mapping |

**Releases:**

|  |  |
| --- | --- |
| Planned functionalities at M18 | Ability to merge and extend existing schemata and create mappings for the overall scenario |
| Planned functionality at M12 | Creation of Global Mappings from privileged users and users' ability to endorse them |
| Planned functionalities at M24 | Visibility of user defined mappings to other users - 'friend' (or 'user I trust') feature and integration to the Web Portal |

**User and Use cases:**

|  |  |
| --- | --- |
| User | Use cases |
|  | Use case 1: Consider an administrator that creates a mapping of some variables of a research project schema to the MIP schema, and shares this schema with the MIP users, thus making it global. Consider now that a MIP user (with no administrative privileges) would like to create a mapping for additional variables of the same schemata/ontologies. It would be beneficial for the user to be able to load the global administrator's schema and consult its correspondences (either endorse them or ignore them) to create her own mapping. Use case 2: Users can share their mappings of novel research data schemas with other users ('friends') and/or endorse their mappings, to further data sharing. |

# SOFTWARE

## Hospital Databases Bundle (HDB)

### Federated Query

**Contributing task:** T8.1.5 Distributed complex workflow engine UoA  
**Description:** The master component transforms, schedules and dispatches the queries to workers  
**Dependencies:**

|  |  |  |
| --- | --- | --- |
| SOFTWARE | Hospital Databases Bundle (HDB) | Hospital Bundle Package |

**Releases:**

|  |  |
| --- | --- |
| Planned functionalities at M18 | Master component is integrated with 5 worker components deployed in 5 hospitals |
| Planned functionality at M12 | Master component released packaged and integrated with a worker component |
| Planned functionalities at M24 | Master component is validated by HBP users |

**User and Use cases:**

|  |  |
| --- | --- |
| User | Use cases |
| Developer - Software Developer - Methods SP8 Platform Developer | None |

**Contributing task:** T8.1.5 Distributed complex workflow engine UoA  
**Description:** The workers reside on the hospital nodes and act as a bridge with the RAW query engine which executes the queries in situ.  
**Dependencies:**

|  |  |  |
| --- | --- | --- |
| SOFTWARE | Hospital Databases Bundle (HDB) | Hospital Bundle Package |

**Releases:**

|  |  |
| --- | --- |
| Planned functionalities at M18 | Worker/ Bridge Component is deployed in 5 hospitals |
| Planned functionality at M12 | Worker/ Bridge Component released, packaged and integrated with Master component and RAW query engine |
| Planned functionalities at M24 | Worker/ Bridge Component is validated by HBP users |

**User and Use cases:**

|  |  |
| --- | --- |
| User | Use cases |
| Developer - Software Developer - Methods SP8 Platform Developer | None |

**Contributing task:** T8.1.6 User Defined Functions (UDFs) and query templates UoA  
**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.  
**Dependencies:**

|  |  |  |
| --- | --- | --- |
| SOFTWARE | Hospital Databases Bundle (HDB) | Hospital Bundle Package |
|  | Algorithm Library | Predictive Models, Feature reduction, Statistical Analytics, Biological Diagnostic Tools, Machine Learning Library |

**Releases:**

|  |  |
| --- | --- |
| Planned functionalities at M18 | Deployment of UDFs component |
| Planned functionality at M12 | Implementation of UDFs component |
| Planned functionalities at M24 | Validation of UDFs component |

**User and Use cases:**

|  |  |
| --- | --- |
| User | Use cases |
| Developer - Software Developer - Methods SP8 Platform Developer | None |

**Contributing task:** T8.1.6 User Defined Functions (UDFs) and query templates UoA  
**Description:** The template composer converts the template, which describes parameterized distributed workflows, into an ExaDFL query script. The template composer is responsible for the isolated execution of each algorithm template.  
**Dependencies:**

|  |  |  |
| --- | --- | --- |
| SOFTWARE | Hospital Databases Bundle (HDB) | Hospital Bundle Package |
|  | Algorithm Library | Predictive Models, Feature reduction, Statistical Analytics, Biological Diagnostic Tools, Machine Learning Library |

**Releases:**

|  |  |
| --- | --- |
| Planned functionalities at M18 | Deployment of template composer component |
| Planned functionality at M12 | Implementation of template composer component |
| Planned functionalities at M24 | Validation of template composer component |

**User and Use cases:**

|  |  |
| --- | --- |
| User | Use cases |
| Developer - Software Developer - Methods SP8 Platform Developer | The user specifies particular values to the parameters of a parameterized algorithm implemented in the query engine and then it executes the algorithm. |

**Contributing task:** T8.1.7 Query templates and workflow management UoA  
**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).  
**Dependencies:**

|  |  |  |
| --- | --- | --- |
| SOFTWARE | Hospital Databases Bundle (HDB) | Hospital Bundle Package |

**Releases:**

|  |  |
| --- | --- |
| Planned functionalities at M18 | Deployment of query template repository |
| Planned functionality at M12 | Implementation of query template repository |
| Planned functionalities at M24 | Validation of query template repository |

**User and Use cases:**

|  |  |
| --- | --- |
| User | Use cases |
| Developer - Software Developer - Methods SP8 Platform Developer | Researchers who are willing to contribute a new algorithm will be able to submit to the code review system a request to add a new algorithm/template or modify an existing one. Senior researchers will then review the proposed changes and accept or deny the modification request. |

**Contributing task:** T8.1.7 Query templates and workflow management UoA  
**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.  
**Dependencies:**

|  |  |  |
| --- | --- | --- |
| SOFTWARE | Hospital Databases Bundle (HDB) | Hospital Bundle Package |

**Releases:**

|  |  |
| --- | --- |
| Planned functionalities at M18 | Deployment of management component of query template repository |
| Planned functionality at M12 | Implementation of management component of query template repository |
| Planned functionalities at M24 | Validation of management component of query template repository |

**User and Use cases:**

|  |  |
| --- | --- |
| User | Use cases |
| Developer - Software Developer - Methods SP8 Platform Developer | The system administrators will be able to monitor the system usage and load and take the appropriate actions. Credentials with limited access will be used for the web portal. |

# SOFTWARE

## Web Exploration and Analytics

### Data Access

**Contributing task:** T8.1.4 Data Integration AUEB  
**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  
**Dependencies:**

|  |  |  |
| --- | --- | --- |
| SOFTWARE | Hospital Databases Bundle (HDB) | Local Database |
| SERVICES | Security & Monitoring | User Management |

**Releases:**

|  |  |
| --- | --- |
| Planned functionalities at M18 | Basic access control functionality |
| Planned functionality at M12 | Report on access restrictions in the MIP |
| Planned functionalities at M24 | Final release of ARM and integration to the platform |

**User and Use cases:**

|  |  |
| --- | --- |
| User | Use cases |
|  | Consider that a MIP user would like to run an EE query and identify the number of patients diagnosed with dementia. According to the access rules imposed by the MIP administrators only Advanced users (AU) are able to run queries that invoke aggregate data, while Simple users (SU) are only able to use the already pre-computed and stored aggregate results. Consider also an AU, who is a pathologist, and thus should not get access to brain imaging data to compute aggregates. All these restrictions can be implemented in the MIP using specific access control languages that also affect the way queries posed by users are reformulated. |

**Contributing task:** T8.1.4 Data Integration AUEB  
**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 LDSM.  
**Dependencies:**

|  |  |  |
| --- | --- | --- |
| SOFTWARE | Hospital Databases Bundle (HDB) | Schema Mapping |

**Releases:**

|  |  |
| --- | --- |
| Planned functionalities at M18 | Query answering without rewriting with respect to database sources schema |
| Planned functionality at M12 | Rewriting of query with respect to ontology module and preliminary integration to the platform |
| Planned functionalities at M24 | Final version - Optimized query answering - and full integration to the platform |

**User and Use cases:**

|  |  |
| --- | --- |
| User | Use cases |
|  | Consider the SNOMED ontology and the concepts Dementia, Alzheimer Dementia and Parkinson Dementia. The SNOMED taxonomy defines that the concept of Alzheimer Dementia and of Parkinson Dementia are subsumed by the concept Dementia meaning that all Alzheimer/Parkinson Dementias are also Dementias. Moreover consider that the concept Alzheimer Dementia is mapped to the variable (of the MIP schema) Diagnosis with value specific to 'AD' and Parkinson Dementia is mapped to the variable Diagnosis with value specific to 'PR'. So, when a user poses a query using the SNOMED ontology that retrieves all patients with dementia, the OBDA system would rewrite this query to also return patients with Alzheimer disease and patients with Parkinson disease and finally would translate those queries to the MIP schema so that every patient with 'AD' and 'PR' value would return. |

**Contributing task:** T8.1.4 Data Integration AUEB  
**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.  
**Dependencies:**

|  |  |  |
| --- | --- | --- |
| SOFTWARE | Hospital Databases Bundle (HDB) | Local Database |
|  | Web Exploration and Analytics | Data Access |
| SERVICES | Security & Monitoring | User Management |

**Releases:**

|  |  |
| --- | --- |
| Planned functionalities at M18 | Basic access control functionality |
| Planned functionality at M12 | Report on access restrictions in the MIP |
| Planned functionalities at M24 | Final release of ARM and integration to the platform |

**User and Use cases:**

|  |  |
| --- | --- |
| User | Use cases |
|  | Consider that a MIP user would like to run an EE query and identify the number of patients diagnosed with dementia. According to the access rules imposed by the MIP administrators only Advanced users (AU) are able to run queries that invoke aggregate data, while Simple users (SU) are only able to use the already pre-computed and stored aggregate results. Consider also an AU, who is a pathologist, and thus should not get access to brain imaging data to compute aggregates. All these restrictions can be implemented in the MIP using specific access control languages that also affect the way queries posed by users are reformulated. |

# SOFTWARE

## Algorithm Library

### Machine Learning Library

**Contributing task:** T8.4.5 Large-scale data analytics on massively parallel architecture ICL  
**Description:** Analytics/clustering algorithms for the efficient and scalable large scale analysis of medical data.  
**Dependencies:**

|  |  |  |
| --- | --- | --- |
| SOFTWARE | Data Factory (DF) | HPC |
|  | Algorithm Library | Machine Learning Library |

**Releases:**

|  |  |
| --- | --- |
| Planned functionalities at M18 | Additional distributed, MPI base algorithms as well as more mature implementations. |
| Planned functionality at M12 | Initial prototypes of a set of algorithms run distributed on HPC via MPI. |
| Planned functionalities at M24 | Deployment of production grade algorithms on supercomputing infrastructure, making it available fro general use. |

**User and Use cases:**

|  |  |
| --- | --- |
| User | Use cases |
| Neuroscientist Clinical Researcher Developer - Software Developer - Methods | Large scale clustering/analysis of large amounts of (cleaned) medical data. |

**Contributing task:** T8.3.5 Methods for distributed rule-based disease signature discovery JSI  
**Description:** Distributed versions of the tree- and rule-based methods for predictive clustering for solving different tasks of predicting structured outputs (e.g. multi-target regression).  
**Dependencies:**

|  |  |  |
| --- | --- | --- |
| DATA | Hospital | Lille Hospital, Tel Aviv Hospital, Milano Hospital, Freiburg Hospital, CHUV Hospital |
| SOFTWARE | Algorithm Library | Predictive Models |
|  | Algorithm Factory (AF) | Model Scoring, X-Validation module, Model Testing (PFA parsing), PFA translation, Model Training / Parameter Estimation, Package of Algorithms as Docker images |

**Releases:**

|  |  |
| --- | --- |
| Planned functionalities at M18 | First tests and validations on research data. |
| Planned functionality at M12 | First working prototype developed. |
| Planned functionalities at M24 | Analysis of distributed data. |

**User and Use cases:**

|  |  |
| --- | --- |
| User | Use cases |
| Neuroscientist Clinical Researcher | Discovery of disease signatures on distributed data. |

**Contributing task:** T8.3.8 Methods for disease progression modeling JSI  
**Description:** Machine learning methods for describing and modelling the temporal dynamics of disease and its clinical and biological markers.  
**Dependencies:**

|  |  |  |
| --- | --- | --- |
| DATA | Hospital | Lille Hospital, Tel Aviv Hospital, Milano Hospital, Freiburg Hospital, CHUV Hospital |
|  | Reference | PPMI (Parkinson's Progression Markers Initiative), ADNI |
| SOFTWARE | Algorithm Library | Predictive Models |
|  | Algorithm Factory (AF) | Model Scoring, X-Validation module, Model Testing (PFA parsing), PFA translation, Model Training / Parameter Estimation, Package of Algorithms as Docker images |

**Releases:**

|  |  |
| --- | --- |
| Planned functionalities at M18 | First tests and validations on research data. |
| Planned functionality at M12 | First working prototype developed. |
| Planned functionalities at M24 | Analysis of longitudinal data. |

**User and Use cases:**

|  |  |
| --- | --- |
| User | Use cases |
| Neuroscientist Clinical Researcher | Discovery of disease signatures in longitudinal data. |

**Contributing task:** T8.3.6 Methods for redescription mining JSI  
**Description:** Methods for redescription mining - 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.  
**Dependencies:**

|  |  |  |
| --- | --- | --- |
| DATA | Hospital | Lille Hospital, Tel Aviv Hospital, Milano Hospital, Freiburg Hospital, CHUV Hospital |
|  | Reference | PPMI (Parkinson's Progression Markers Initiative), ADNI |
| SOFTWARE | Algorithm Factory (AF) | Model Scoring, X-Validation module, Model Testing (PFA parsing), PFA translation, Model Training / Parameter Estimation, Package of Algorithms as Docker images |

**Releases:**

|  |  |
| --- | --- |
| Planned functionalities at M18 | First tests and validations on research data. |
| Planned functionality at M12 | First working prototype developed. |
| Planned functionalities at M24 | Analysis of multi-view data. |

**User and Use cases:**

|  |  |
| --- | --- |
| User | Use cases |
| Neuroscientist Clinical Researcher | Discovery of disease signatures. |

**Contributing task:** T8.3.7 Methods for heterogeneous networks JSI  
**Description:** Mehods for mining text-enriched heterogeneous information networks.  
**Dependencies:**

|  |  |  |
| --- | --- | --- |
|  | Reference | PPMI (Parkinson's Progression Markers Initiative), ADNI |
| SOFTWARE | Algorithm Library | Predictive Models |
|  | Algorithm Factory (AF) | Model Testing (PFA parsing), PFA translation, Package of Algorithms as Docker images |

**Releases:**

|  |  |
| --- | --- |
| Planned functionalities at M18 | First tests and validations on research data. |
| Planned functionality at M12 | First working prototype developed. |
| Planned functionalities at M24 | Analysis of text-enriched heterogeneous information networks. |

**User and Use cases:**

|  |  |
| --- | --- |
| User | Use cases |
| Neuroscientist Clinical Researcher | Discovery of disease signatures on text-enriched data. |

# SOFTWARE

## Algorithm Library

### Brain Anatomy

**Contributing task:** T8.3.10 Methods for linkage of local SNP data (individual SNPs) to imaging data through SNP LUMC  
**Description:** Algorithm that generates a 3D expression heatmap of of an SNP name, gene name or co-expression module.  
**Dependencies:**

|  |  |  |
| --- | --- | --- |
| SOFTWARE | Web Exploration and Analytics | Image & Genetic Viewer |
|  | Algorithm Library | Feature reduction, Statistical Analytics, Brain Anatomy |

**Releases:**

|  |  |
| --- | --- |
| Planned functionalities at M18 | M 8.3.2: Initial proof-of-concept algorithm for generating 3D gene expression heatmaps from gene co-expression modules. |
| Planned functionality at M12 | M 8.3.2 Initial proof-of-concept algorithm for generating 3D gene expression heatmaps of single genes and SNPs |
| Planned functionalities at M24 | M 8.3.11 Gene Expression Maps of Disease Link to Brain Atlases |

**User and Use cases:**

|  |  |
| --- | --- |
| User | Use cases |
| Developer - Methods | Discovery of disease associations between gene expression data en imaging data |

# SOFTWARE

## Algorithm Library

### Statistical Analytics

**Contributing task:** T8.3.1 Tools to mine replicable selection and integration of hierarchical features, inter and across domains using FDR. TAU  
**Description:** Methodology for Medical big data analysis and disease sub-type identification  
**Dependencies:**

|  |  |  |
| --- | --- | --- |
| DATA | Hospital | Lille Hospital, Tel Aviv Hospital, Milano Hospital, Freiburg Hospital, CHUV Hospital |
|  | Reference | PPMI (Parkinson's Progression Markers Initiative), ADNI |

**Releases:**

|  |  |
| --- | --- |
| Planned functionalities at M18 | Methodology improvemnts - scaling up the ability to work with larger datasets. |
| Planned functionality at M12 | MS 8.3.2: Initial Proof-of-concept and results of the different 3-C Algorithms |
| Planned functionalities at M24 | Validation of 3-C strategy on different data |

**User and Use cases:**

|  |  |
| --- | --- |
| User | Use cases |
| Clinical Researcher Developer - Methods SP8 Platform Developer | None |

**Contributing task:** T8.3.1 Tools to mine replicable selection and integration of hierarchical features, inter and across domains using FDR. TAU  
**Description:** Incorporating longitudinal information (day-to-day and multi-patients visits).  
**Dependencies:**

|  |  |  |
| --- | --- | --- |
| DATA | Hospital | Lille Hospital, Tel Aviv Hospital, Milano Hospital, Freiburg Hospital, CHUV Hospital |
|  | Reference | PPMI (Parkinson's Progression Markers Initiative), ADNI |
| SOFTWARE | Algorithm Library | Statistical Analytics |

**Releases:**

|  |  |
| --- | --- |
| Planned functionalities at M18 | Clinical measurements data construction |
| Planned functionality at M12 | Review of existing methods |
| Planned functionalities at M24 | MS8.3.11: Analyses of 3-C demonstrating the use of Longitudinal Data |

**User and Use cases:**

|  |  |
| --- | --- |
| User | Use cases |
| Clinical Researcher Developer - Methods SP8 Platform Developer General Public | None |

**Contributing task:** T8.3.1 Tools to mine replicable selection and integration of hierarchical features, inter and across domains using FDR. TAU  
**Description:** methodology for enriching current models  
**Dependencies:**

|  |  |  |
| --- | --- | --- |
| DATA | Hospital | Lille Hospital, Tel Aviv Hospital, Milano Hospital, Freiburg Hospital, CHUV Hospital |
|  | Reference | PPMI (Parkinson's Progression Markers Initiative), ADNI |
| SOFTWARE | Algorithm Library | Feature reduction, Statistical Analytics, Machine Learning Library |

**Releases:**

|  |  |
| --- | --- |
| Planned functionalities at M18 | Increasing the variety of data sources that can be integrated into the previous analyses, in particular genomics |
| Planned functionality at M12 | Developing methodologies for enriched data MS8.3.3 |
| Planned functionalities at M24 | Analysis of multi-domain data |

**User and Use cases:**

|  |  |
| --- | --- |
| User | Use cases |
| Neuroscientist Clinical Researcher Developer - Methods | None |

**Contributing task:** T8.3.2 Developing methods for high-dimensional data with possible informative missing values TAU  
**Description:** Symetry targeted monotone transformations, and the advantage gained in variance stability,linearity and clustering.  
**Dependencies:**

|  |  |  |
| --- | --- | --- |
| SOFTWARE | Data Factory (DF) | Data Pipeline processes, Data Quality Processes |
|  | Algorithm Library | Statistical Analytics |

**Releases:**

|  |  |
| --- | --- |
| Planned functionalities at M18 | Functions improvements |
| Planned functionality at M12 | Improvements and use in different datasets |
| Planned functionalities at M24 | Prepare methods for integration into MIP, test and validate results. (preparation specification to be agreed with and provided by MIP implementation team) |

**User and Use cases:**

|  |  |
| --- | --- |
| User | Use cases |
| Neuroscientist Clinical Researcher Developer - Methods | None |

**Contributing task:** T8.3.2 Developing methods for high-dimensional data with possible informative missing values TAU  
**Description:** Develop statistical procedures and workflows to help and guide the discovery of possible patterns in missing values  
**Dependencies:**

|  |  |  |
| --- | --- | --- |
| DATA | Hospital | Lille Hospital, Tel Aviv Hospital, Milano Hospital, Freiburg Hospital, CHUV Hospital |
|  | Reference | TBI (Traumatic Brain Injury), PPMI (Parkinson's Progression Markers Initiative), ADNI |
| SOFTWARE | Web Exploration and Analytics | Image & Genetic Viewer, Research & Modeling application |
|  | Algorithm Library | Statistical Analytics |

**Releases:**

|  |  |
| --- | --- |
| Planned functionalities at M18 | Missing values: structure discovery |
| Planned functionality at M12 | Missing values - extraction |
| Planned functionalities at M24 | Missing values: imputation and visualization MS8.3.10 |

**User and Use cases:**

|  |  |
| --- | --- |
| User | Use cases |
| Neuroscientist Clinical Researcher Developer - Methods SP8 Platform Developer General Public | None |

**Contributing task:** T8.3.3 Introducing selective inference into dimensionality reduction and clustering methods TAU  
**Description:** Evaluating disease signature clusters (by combining new approaches with tools developed for the visualization and manipulation of hierarchical clustering)  
**Dependencies:**

|  |  |  |
| --- | --- | --- |
| DATA | Hospital | Lille Hospital, Tel Aviv Hospital, Milano Hospital, Freiburg Hospital, CHUV Hospital |
|  | Reference | PPMI (Parkinson's Progression Markers Initiative), ADNI |
| SOFTWARE | Algorithm Library | Statistical Analytics |

**Releases:**

|  |  |
| --- | --- |
| Planned functionalities at M18 | Exploration of new methods for incorporating Knowledge into the process |
| Planned functionality at M12 | Re-evaluation of 3-C methodology as a data and knowledge combined methodology |
| Planned functionalities at M24 | Report on incorporating Knowledge into the process |

**User and Use cases:**

|  |  |
| --- | --- |
| User | Use cases |
| Developer - Methods | None |

**Contributing task:** T8.3.4 Statistical methods for 'Disease Signature' confidence assessment TAU  
**Description:** Define and propose a model for disease signature  
**Dependencies:**

|  |  |  |
| --- | --- | --- |
| SOFTWARE | Algorithm Library | Statistical Analytics |

**Releases:**

|  |  |
| --- | --- |
| Planned functionalities at M18 | Summary of consultations with multi-disciplinary experts and opinion leaders |
| Planned functionality at M12 | Literature Review |
| Planned functionalities at M24 | Suggest definition of the 'Disease signature' and proposed estimation methods. |

**User and Use cases:**

|  |  |
| --- | --- |
| User | Use cases |
| Clinical Researcher General Public | None |

# SOFTWARE

## Algorithm Factory (AF)

### Workflow Engine: Woken

**Contributing task:** T8.1.2 Installation of RAW on local computing infrastructure for hospital analytics requirements EPFL  
**Description:** A local data hub is a computer cluster, that is deployed within the same country as the hospital and connected together via a secure network. The objective is to allow small hospitals to share IT infrastructure.  
**Dependencies:**

|  |  |  |
| --- | --- | --- |
| DATA | Hospital | Lille Hospital, Tel Aviv Hospital, Milano Hospital, Freiburg Hospital, CHUV Hospital |
| SOFTWARE | Data Factory (DF) | Data Anonymisation, Workflow Engine, Data Pipeline processes, Data Quality Processes, Data Storage, Data Integration, Data Categorization, Feature Engineering, HPC |
|  | Hospital Databases Bundle (HDB) | Local Database, Federated Query, Schema Mapping, Hospital Bundle Package |
|  | Algorithm Factory (AF) | Workflow Engine: Woken, Package of Algorithms as Docker images |
| SERVICES | Upgrade-Deploy-Release | MIP Integrated Releases Algorithm Factory |
|  | Security & Monitoring | Data governance |

**Releases:**

|  |  |
| --- | --- |
| Planned functionalities at M18 | Software extension to support provenance tracking through hospital hubs have been implemented. |
| Planned functionality at M12 | A secure procedure has been defined to connect a hospital to a hub. Related data management processes are also defined. |
| Planned functionalities at M24 | The software required by the procedure is developed, tested and packaged, ready to be used if necessary. |

**User and Use cases:**

|  |  |
| --- | --- |
| User | Use cases |
| Developer - Software | Small hospitals with restricted resources can still contribute to the human brain project. |

# SERVICES

## Security & Monitoring

### User Management

**Contributing task:** T8.1.4 Data Integration AUEB  
**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  
**Dependencies:**

|  |  |  |
| --- | --- | --- |
| SOFTWARE | Hospital Databases Bundle (HDB) | Local Database |
| SERVICES | Security & Monitoring | User Management |

**Releases:**

|  |  |
| --- | --- |
| Planned functionalities at M18 | Basic access control functionality |
| Planned functionality at M12 | Report on access restrictions in the MIP |
| Planned functionalities at M24 | Final release of ARM and integration to the platform |

**User and Use cases:**

|  |  |
| --- | --- |
| User | Use cases |
|  | Consider that a MIP user would like to run an EE query and identify the number of patients diagnosed with dementia. According to the access rules imposed by the MIP administrators only Advanced users (AU) are able to run queries that invoke aggregate data, while Simple users (SU) are only able to use the already pre-computed and stored aggregate results. Consider also an AU, who is a pathologist, and thus should not get access to brain imaging data to compute aggregates. All these restrictions can be implemented in the MIP using specific access control languages that also affect the way queries posed by users are reformulated. |

# SERVICES

## Upgrade-Deploy-Release

### MIP Integrated Releases Hospital Databases Bundle

**Contributing task:** T8.1.3 Installation and adaptation of SP8 bundle at local hospitals with upgrade support EPFL  
**Description:** This component will prepare and install the first working version of the FCDI (Hospital Bundle) at the participating hospitals in a bottom-up fashion. This effort includes going to hospitals, installing the software, understanding the network topology,configuring the data sources and getting the right permissions. It will also integrate new features as they become available in the other tasks of WP8.1 into the FCDI.  
**Dependencies:**

|  |  |  |
| --- | --- | --- |
| SOFTWARE | Data Factory (DF) | Workflow Engine |
|  | Hospital Databases Bundle (HDB) | Local Database, Federated Query, Schema Mapping, Hospital Bundle Package |
|  | Algorithm Factory (AF) | Workflow Engine: Woken |
| SERVICES | Upgrade-Deploy-Release | Micro-services |

**Releases:**

|  |  |
| --- | --- |
| Planned functionalities at M18 | Continuous process |
| Planned functionality at M12 | Continuous process |
| Planned functionalities at M24 | Continuous process |

**User and Use cases:**

|  |  |
| --- | --- |
| User | Use cases |
| Developer - Software | Base deployment infrastructure of the hospital bundle. |

# SERVICES

## Upgrade-Deploy-Release

### MIP Integrated Releases Web Exploration & Analytics

**Contributing task:** T8.1.5 Distributed complex workflow engine UoA  
**Description:** It interfaces the master component with the web portal.  
**Dependencies:**

|  |  |  |
| --- | --- | --- |
| SOFTWARE | Hospital Databases Bundle (HDB) | Federated Query, Hospital Bundle Package |

**Releases:**

|  |  |
| --- | --- |
| Planned functionalities at M18 | First version of web portal connector component deployed |
| Planned functionality at M12 | Define communication protocol with web portal |
| Planned functionalities at M24 | Web portal connector component is validated by HBP users |

**User and Use cases:**

|  |  |
| --- | --- |
| User | Use cases |
| Developer - Software Developer - Methods SP8 Platform Developer | The user is able to see a catalogue of algorithms, select one of them, to execute the selected algorithm by specifying particular values to its parameters and see the results of the exexution. |

# DATA

## MDR (Meta Data Register)

### Ontology&Standards

**Contributing task:** T8.1.4 Data Integration AUEB  
**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 LDSM.  
**Dependencies:**

|  |  |  |
| --- | --- | --- |
| SOFTWARE | Hospital Databases Bundle (HDB) | Schema Mapping |

**Releases:**

|  |  |
| --- | --- |
| Planned functionalities at M18 | Query answering without rewriting with respect to database sources schema |
| Planned functionality at M12 | Rewriting of query with respect to ontology module and preliminary integration to the platform |
| Planned functionalities at M24 | Final version - Optimized query answering - and full integration to the platform |

**User and Use cases:**

|  |  |
| --- | --- |
| User | Use cases |
|  | Consider the SNOMED ontology and the concepts Dementia, Alzheimer Dementia and Parkinson Dementia. The SNOMED taxonomy defines that the concept of Alzheimer Dementia and of Parkinson Dementia are subsumed by the concept Dementia meaning that all Alzheimer/Parkinson Dementias are also Dementias. Moreover consider that the concept Alzheimer Dementia is mapped to the variable (of the MIP schema) Diagnosis with value specific to 'AD' and Parkinson Dementia is mapped to the variable Diagnosis with value specific to 'PR'. So, when a user poses a query using the SNOMED ontology that retrieves all patients with dementia, the OBDA system would rewrite this query to also return patients with Alzheimer disease and patients with Parkinson disease and finally would translate those queries to the MIP schema so that every patient with 'AD' and 'PR' value would return. |

**Contributing task:** T8.3.9 Ontologies for describing data on neurological diseases, patients JSI  
**Description:** A mid-level ontology for describing various types of data on patients with neurological diseases.  
**Dependencies:**

**Releases:**

|  |  |
| --- | --- |
| Planned functionalities at M18 | First dataset annotated according to the ontology. |
| Planned functionality at M12 | Prototype of the ontology for describing data on patients with neurological diseases developed |
| Planned functionalities at M24 | Ontology and several datasets annotated according to the ontology. |

**User and Use cases:**

|  |  |
| --- | --- |
| User | Use cases |
| Developer - Software Developer - Methods SP8 Platform Developer | Annotation of different types of data in the MIP. |

# DATA

## MDR (Meta Data Register)

### Common Variables & Metadata

**Contributing task:** T8.1.4 Data Integration AUEB  
**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 anonymization 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.  
**Dependencies:**

|  |  |  |
| --- | --- | --- |
| SOFTWARE | Data Factory (DF) | Data Integration |
|  | Hospital Databases Bundle (HDB) | Schema Mapping |

**Releases:**

|  |  |
| --- | --- |
| Planned functionalities at M18 | Preliminary version of module |
| Planned functionality at M12 | Report on Incremental Data Integration |
| Planned functionalities at M24 | Final version and integration to the Hospital Bundle |

**User and Use cases:**

|  |  |
| --- | --- |
| User | Use cases |
|  | A hospital participates in the MIP by providing a first batch of data. Consider that there exists a rule stating that the target diagnostic table is populated only with patients (source patient table) who have a diagnosis (source diagnostic table). Moreover consider that in this first data batch patient 'A' did not have any examinations, hence the target diagnostics table does not contain any tuple for 'A'. Consider now that in a second batch, a month later, diagnostics for patient 'A' are entered but patient information for 'A' is not resent (as it had been sent in the first batch a month earlier). In order to populate the target diagnostic table with information for 'A' it would be essential to run again the data exchange process for all the input data (first and second batch). However, this poses a significant computational burden. Hence, it would be ideal if an incremental data exchange approach would be used that would not need to re-run the data translation process from scratch but would be able to generate the information for patient 'A' and her diagnostics as they became available. |