Call identifier: H2020-ICT-2016 - **Grant agreement no**: **732907**

Topic: ICT-18-2016 - Big data PPP: privacy-preserving big data technologies

Deliverable 4.5

Data Management Plan

Due date of delivery: April 30th, 2017

Actual submission date: June 9th, 2017

Start of the project: 1st November 2016
Ending Date: 31st October 2019

Partner responsible for this deliverable: ATHENA RC

Version: 3.0

Document Classification

Title	Data Management Plan
Deliverable	4.1
Reporting Period	-
Authors	Omiros Metaxas
Work Package	4
Security	Public
Nature	ORDP Open Research Data Pilot
Keyword(s)	Data profiling, pre-processing, anonymisation, sharing

Document History

Name	Remark	Version	Date
Omiros Metaxas	First Version	1.0	29/5/2019
Patrick Ruch	Second Version	2.0	07/06/2017
Anna Rizzo	Third Version	3.0	07/06/2017

List of Contributors

Name	Affiliation
Omiros Metaxas	Athena RC
Harry Dimitropoulos	Athena RC

List of reviewers

Name	Affiliation
Patrick Ruch	HES-SO
Anna Rizzo	Lynkeus
Antonella Trezzani	Lynkeus
Edwin Morley-Fletcher	Lynkeus

1 INTRODUCTION

This deliverable provides the MHMD data management plan version 1, outlining what kind of clinical and personal data will be collected or generated and how it will be handled, processed and shared. It describes the standards that will be incorporated and the related methodology for data collection, pre-processing, data profiling, anonymisation and data sharing that will be followed. This deliverable is based on the template and the guidelines provided by the European Commission and it will be updated during the project (live document).

2 DATA SUMMARY

2.1 Purpose of the data collection/generation and the relation to the objectives of the project

The MyHealthMyData (MHMD) project aims at fundamentally changing the way sensitive data are shared between individuals (patients) and/or healthcare stakeholders (i.e., medical institutions and other organizations). Proposed platform will bring together, integrate, harmonize and semantically consolidate clinical data from medical information systems, individual user data from personal data accounts and machine-generated data from Internet of Things (IoT) connected devices. At the same time, data management and data processing flow in MHMD will address three crucial, and sometimes interrelated or competitive, goals:

1. maximize data usage and data sharing unlocking the value of large volumes of biomedical data by allowing rapid merging of disparate, heterogeneous data sources and their lawful access by third party to support a proper privacy preserving Big Data analytical framework;
2. assess and ensure the quality of the heterogeneous, multi-modal biomedical and personal data;
3. prevent privacy breaches and comply with the European General Data Protection Regulation (GDPR) implementing both Privacy by Design and Privacy by Default principles.

2.2 Overall data management and data sharing architecture

MHMD challenge consists in simultaneously addressing all of the above-mentioned goals, having recourse to no one-size-fits-all solution, neither for all related data modalities, nor for all provided services. Different services have different processing requirements, while different data modalities may be considered more or less private or need different handling. To deal with all these issues, MHMD’s holistic and innovative data management and data sharing architecture combines:

1. a revolutionary decentralized data management and data sharing platform that enforces consent and peer-to-peer data transactions between healthcare stakeholders in a probative, secure and open manner offering very strong privacy safeguards and security guarantees,
2. a semi-automated data profiling and cleaning engine that ensures and assesses data quality while at the same time guaranteeing the most appropriate de-identification or encryption mechanism, according to each type of data or modality,

3. a well-designed privacy preserving and security layer that combines a multi-level anonymisation engine to support data privacy preserving data publishing to external parties and a privacy preserving complex data flow execution engine (i.e., differential privacy, Secure Multi-Party Computation (SMPC), homomorphic encryption) to support privacy preserving data mining and analytics within MHMD platform.

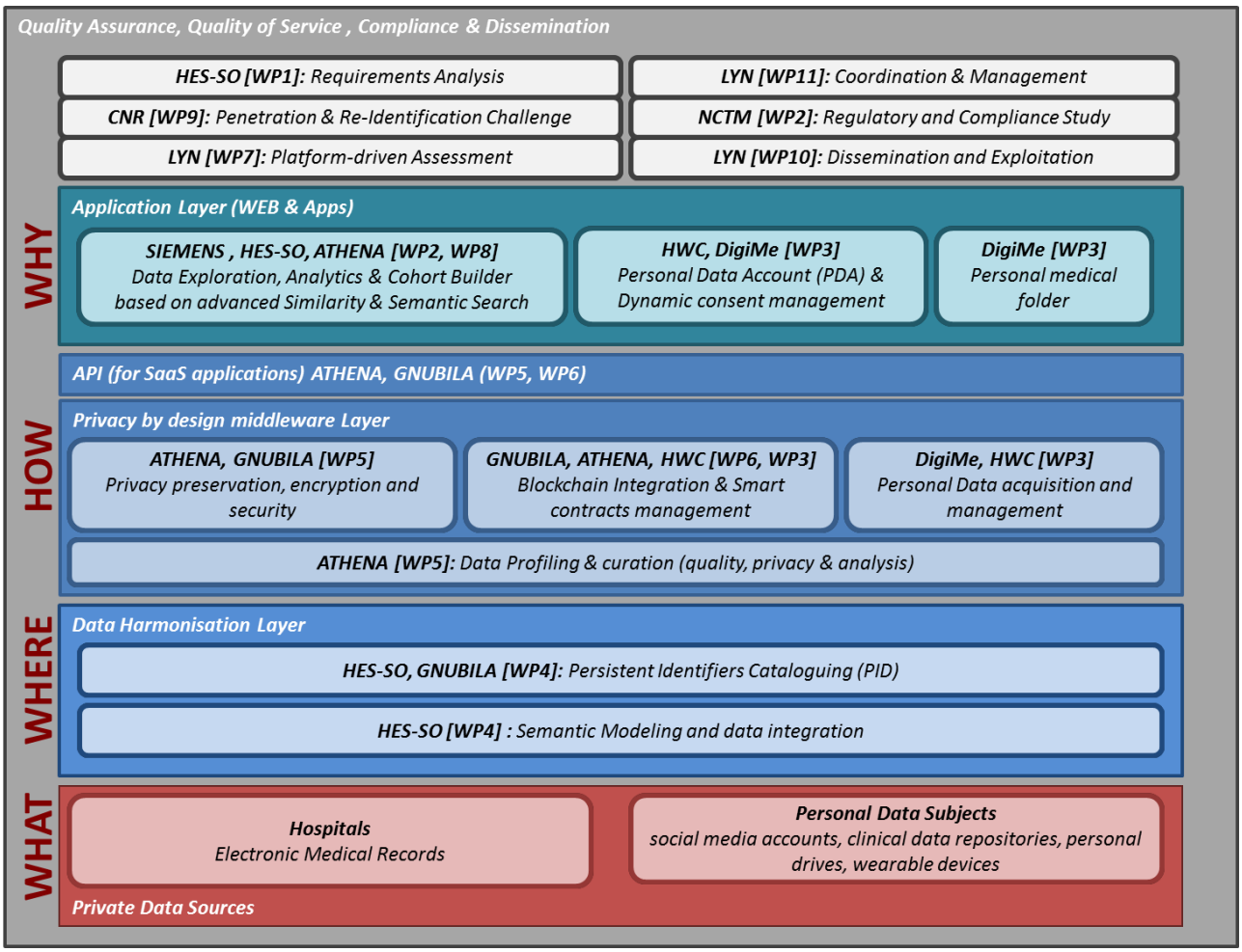

Figure 1: MHMD Architecture

Hence, relying on a federated data management infostructure where no central authority holds the entirety of data and a blockchain platform as a distributed, public and transparent ledger that orchestrates and monitors data sharing, MHMD will decentralize data storage and it will enable not only the project stakeholders but also data subjects to witness data sharing activities at any time. This key architectural specificity will allow MHMD to bring trust within a network of possibly highly heterogeneous and unsecure appliances. Transactions will be automated thanks to the provision of custom-tailored smart contracts.

2.3 Data description

MHMD will generate and integrate three main types of data.

1. **Pseudonymised (de-identified) clinical (routine) data** extracted from medical information systems (e.g., phenotype / demographic data, genomic data, medical images and signals, lab tests). Such data will be stored in a federated data storage platform where each hospital will have its own node.

2. **Individual personal data including machine-generated data from Internet of Things (IoT)** connected devices (wearables, smartphones): taking stock of MHMD's partner Digi.me (https://get.digi.me), MHMD will aggregate personal data from disparate sources (i.e., social media accounts, clinical data repositories, personal drives) and data derived from commonly used wearables, or personal monitoring devices, as they are stored on smartphones. Such data will be stored in a centralised, user-owned account.
3. **Derived data related to the usage and the processing of the data**: such data could be related to the different types of data profiles, pre-processing and mining data flows, analytics, biomedical and statistical simulation models, user profiles for app personalisation and privacy preservation, blockchain and security transactions.

2.4 Data Sourcing

The data sources to be explored are, in priority and chronological order:

1. **Hospital pseudonymised datasets**: already consented and available pseudonymised data from clinical partners having taken part in the MD-Paedigree (md-paedigree.eu) and Cardioproof (cardioproof.eu) E.U. funded projects (UCL, DHZB, OPBG);
2. **Individual user data**: individual digi.me users who will download the application and start sharing their data;
3. **Hospitals bringing additional data**: bringing in other individual users among their patients or involving other third parties (clinicians, hospitals, patients' associations).

2.5 Data extraction and data storage

2.5.1 Clinical data extracted from Healthcare Information Systems

The MHMD project will build upon and extend the already existing distributed data management and storage platform that interconnects several clinical centres in EU FP7 MD-Paedigree and FP7 CARDIOPROOF projects and the related biomedical data extraction, pre-processing and data integration flow. Based on this flow, routine clinical data are extracted from local Healthcare Information Systems within hospitals and are properly pseudo-anonymized (de-identified), normalized, curated, transformed and stored on a local node within the hospital. This architecture allows sourcing and preparing sensitive data at the hospital level and applying proper anonymisation onsite under the strict supervision of local IT and data controllers, who can quality check, quarantine, or even stop the sharing at any time. The verified data are then uploaded to a local (within hospital) node, which federates contents with the other connected centres. Beyond the data sourcing process, this architecture also makes it possible to deeply penetrate the local Healthcare Information System, by connecting it to the ETL routing system or proprietary RIS, PIS or PACS databases. As such, 3 of the participating hospitals in MHMD have integrated the solution to their routine systems.

This integrative architecture is a competitive and unique advantage for the project as it enforces privacy-by-design starting immediately from the data source and leaves full control to the data controllers over time. It also makes it possible to establish a 2-phase development strategy for the market place, starting from synthetic test data and then moving to exploitation with routine data. Besides, having real clinical centres involved, they will conform with their respective national laws as to the conservation of their respective medical sensitive data over time.

2.5.2 Individual Personal Data

The basic data management resides on the Personal Data Account (PDA) application of the DIGI.me that will retrieve in the background personal data to an encrypted local library, which the users can then add to a personal cloud of their choice (e.g. Dropbox, Google Drive, Microsoft OneDrive, or a home based personal cloud such as Western Digital MyCloud) to sync across all their devices. Hence, through the adoption of the digi.me app, MH-MD will gather personal data from sparse data sources, from actual biomedical data to data shared through social networks, from biometric data coming from wearable and mobile devices to privacy preferences gathered with specific questionnaires, etc.

A key benefit is that locally stored data do not interact or come into contact with any other interface servers or third-party storage houses. The User Interface (UI) will be engaging, while providing the users with an incentive to appreciate and benefit from their data. Hence, the MHMD architecture is such that no third party, nor MHMD itself, can directly access any user data held in the personal MHMD encrypted library. Data subjects can permission access to portions of that data to apps websites/businesses using a Permission Access Certificate (PAC) that is designed to ensure explicit and informed consent together with a clear requirement for “Right to Forget” and a protocol to activate that Right at a later date.

2.6 Data usage and utility

The ultimate goal of MHMD is to extract valuable and accurate information from clinical routine data targeting specific similarity analysis and knowledge discovery uses cases related to precision medicine and biomedical research. Individual personal data will be used in conjunction with those coming from clinical data repositories, and contribute to the overall data pool, supporting cross-domain knowledge discovery analyses. For instance, geolocation and physical activity data, as well as purchases and social media activities, can provide valuable indicators to classify medical risk profiles. Finally, the proposed platform will allow patients to share their data with medical institutions and other organizations while still enjoying very strong privacy safeguards.

3 MAKING DATA FINDABLE, ACCESSIBLE, INTEROPERABLE AND RESUSABLE [FAIR DATA]

3.1 Data Modelling, Harmonisation and Integration

For the purpose of research and business, distributed biomedical and personal data need to be normalized. The already existing MD-Paedigree/Cardioproof Infostructure has been designed and implemented with this specific purpose in mind and is currently deployed to serve both projects’ needs. This infrastructure will be extended to ingest and semantically integrate additional, non-medical data sources.

At the heart of the system, a patient centric data model will be developed capturing and integrating all biomedical data following a dynamic Subjective-Objective-Assessment-Plan (SOAP) model of an Electronic Medical Record supporting vertical integration and temporal evolution. Whenever possible, well-established biomedical onto-terminological resources such as ATC, SNOMED CT, ICD-10, MESH, etc. will be incorporated either directly or as semantic annotations. In addition, efficient data storage and handling of non-traditional data types such as geolocation data, images and streams will be supported, e.g., data from wearable devices.

For personal data, MHMD will further extend the already existing digi.me Personal Data Account semantic modelling scheme taking into consideration possible overlaps on biomedical data modelling. In addition to the patient specific data, application specific data will be modelled and integrated.

3.2 Data Cataloguing and Persistent Identifiers

MHMD will develop a catalogue service indexing available data in the centres with Persistent Identifiers. The data model will be used to populate and browse the MHMD global data catalogue, and it will be mapped to the Persistent IDentifiers (PIDs), to create non repudiable, persistent, unique and standard identifiers to selected data points. The resulting data catalogue will be browsable by advanced semantic-enabled engines and interfaces, allowing to segment, group, and thus create, specific cohorts of data. PIDs will be used in transactions in lieu of the actual data and will thus ensure that no sensitive data is compromised nor exposed at any time in the transaction processes.

3.3 Accessibility and Data sharing

MHMD's goal is to create the first open biomedical information network centred on the connection between organisations and the individual, aiming at encouraging hospitals to start making pseudo-anonymized or anonymised data available for open research, while prompting citizens to become the ultimate owners and controllers of their health data.

Regarding personal data, the GDPR legislation identifies two alternatives regarding the application of the EU regulation:

- Anonymised (irreversibly de-identified or "sanitized") data, for which re-identification is made impossible with current "state of the art" technology. For these types of data, the GDPR does not apply, so long as the data subject cannot be re-identified, even by matching his/her data with other information held by third parties. Data security, however, is not defined by the legal authority.
- Pseudonymised (partially de-identified) data: they constitute the basic privacy-preserving level allowing for some data sharing, and represent data where direct identifiers (e.g. Names, SSN) or quasi-identifiers (e.g. unique combinations of date and zip codes) are removed and data is mismatched with a substitution algorithm, impeding correlation of readily associated data to the individual's identity. For such data, GDPR applies and appropriate compliance must be ensured.

In the context of MHMD both options will be considered and addressed through well-defined data sharing flows as follows:

Accessing Anonymized Data: MHMD will consider possible re-use, sharing and correct citation/crediting of specific subsets of Anonymised datasets in an Open Science environment ensuring compliance with the European efforts and policies related to OpenAccess and OpenData. In more detail, MH-MD will consider the adoption of the appropriate policies in the entire data flow and under specific consent will provide access to experimental anonymised datasets through research data repositories and horizontal infrastructures (e.g., OpenAIRE, ZENODO). Such datasets could be either related to a small number of variables targeting specific clinical research use cases or contain aggregated / statistical information (e.g., for an epidemiological research). Well established anonymisation techniques will be incorporated ensuring specific privacy guarantees (e.g., k-Anonymity) while optimizing data utility.

Accessing Pseudonymised (partially de-identified) data: all clinical data stored in the system will be pseudonymised and will be only accessible within MHMD data management and data processing platform through specific privacy preserving APIs. MHMD relies on a decentralized, blockchain-based infrastructure that monitors and orchestrates data sharing transactions and a multi-level privacy preserving and security layer that provides secure access with specific privacy guarantees on the data. This way it ensures that data will only be accessed and used from specific stakeholders and applications (data processors) and for well-defined and specific purposes in alignment with the data subject's 'dynamic' consent. Dynamic Consent allows to extend traditional consents, combining them into a novel user workflow in which patients may or may not allow access to their data based on a range of key parameters:

- What will data be used for
- What will be done with the data
- What data will be retained
- What data will be shared with 3rd parties and for what purpose
- How will the right to be forgotten be implemented

Hence, MHMD will give the opportunity and assurance to the data subjects (e.g., patients, hospitals, individuals) that they are able to control their data in a flexible and agile manner, being enabled to monitor and re-evaluate the clauses included in the initial agreement / consent.

3.4 Data Profiling and Data Quality

MHMD will incorporate the already existing DCV Data profiling and Data Cleaning engine provided by ATHENA RC to assess and ensure the quality of the data. DCV is able to analyse the content, structure, and relationships within data to uncover patterns, inconsistencies, anomalies, and redundancies and automate the curation processes using a variety of advanced data cleaning methods. MHMD will work on expanding already existing data profiling capabilities, defining a formal methodology to support classification of medical data and correspondent security and privacy provisions suggested in each category. The MHMD methodology will be framed by regulatory analysis and yield indication for policies in those areas where current regulations are not addressing fine grained operational constraints.

3.5 Allocation of resources and responsibilities

- **Federated data management:** Gnúbila (a data privacy solution designer and independent software vendor) will develop, deploy and maintain the federated data management MHMD Infostructure for the clinical centres. Extending its FedEHR federated platform and its FedEHR Anonymizer product, that have already been deployed at the participating hospitals of MD-Paedigree and Cardioproof projects, Gnùbila will provide solutions to extract, de-identify, demilitarise and share medical sensitive data cross-enterprise and transnational.

- **Clinical Data modelling and data integration:** HES-SO (University of Applied Sciences Western Switzerland) (leader of WP4) will be responsible for the clinical data sourcing and preparation, the construction of a clinical data catalogue and the normalization of the clinical data with reference terminologies.

- **Personal data management:** Digi.me will provide and extend the already existing digi.me software and platform that will gather personal data from sparse data sources, from actual biomedical data to data shared through social networks and from biometric data coming from wearable and mobile devices to privacy preferences gathered with specific questionnaires. Digi.me will also provide expertise and knowledge as required concerning personal data, data normalisation, and health value exchange.

- **Data Profiling & Data Quality Assurance:** ATHENA RC will provide the necessary tools, techniques and methodologies for data profiling (including data sensitivity and privacy profiling) and data curation, extending the already existing DCV data profiling and data cleaning web based tool (deployed in MD-Paedigree project).

- **Privacy Preserving solutions and data security:** ATHENA RC (leader of the related WP5) will provide the anonymisation tool (AMNESIA) and the related techniques for privacy preserving data publication as well as a privacy preserving complex data flow execution engine (EXAREME) targeting privacy preserving data mining within MHMD. In addition, ATHENA will provide the required API for privacy preserving data access.

- **Blockchain Infrastructure and Smart Contracts:** Gnùbila (leader of the related WP6) will provide, integrate and deploy the blockchain platform which will handle consent and data transactions between the concerned centres. ATHENA RC will participate at the specification of the blockchain related policies, requirements and guidelines. Lynkeus will participate at the Smart Contracts specification.

4 DATA PROTECTION, PRIVACY PRESERVATION AND DATA SECURITY

MHMD is dealing with highly sensitive biomedical and personal data hence data security and privacy preservation will be addressed in every step of the data processing flow, from harvesting and curation to sharing and analysis. Following and implementing privacy-by-design and privacy-by-default guidelines, MHMD will develop an innovative architecture for data storage, access, and sharing, having recourse to federated data management and blockchain / smart contracts technology, and combining it with multi-level anonymisation and encryption techniques, whose efficiency and usability will be quantitatively measured during the project's duration. In addition, a complete methodology for re-identification and penetration threats modelling and test will be developed and the resulting system will be openly challenged, to spot possible breaches.

4.1 Privacy preserving data sharing and decentralized monitoring and orchestration

As described in section 3.3, MHMD will combine and support two specific data access / sharing flows:

- Privacy preserving data publishing where specific anonymized subsets of data will be exposed to external parties

- Privacy preserving complex data flow execution within MHMD platform, where specific applications will be able to process and analyse the pseudo-anonymized data through a well-defined secure API that implements multi-level privacy preservation techniques (including Secure Multi-Party Computation (SMPC), differential privacy and homomorphic encryption) targeting data mining and analytics.

A key novelty of MHMD will be the incorporation of these mechanisms in its overall privacy policy in conjunction with cryptographic and data fishing prevention techniques.

The entire platform will rely on a blockchain infrastructure to orchestrate and monitor data sharing transactions (where transactions will be made of anonymous consent(s) and their related PID(s)). Relying on the blockchain as a distributed, public and transparent ledger will enable not only the project stakeholders but also data subjects to witness data sharing activities at any time, while decentralizing decision making on the actual transactions. Transactions will be automated thanks to the provision of custom-tailored smart contracts. This way, MHMD will promote decentralised privacy preserving data sharing and analytics, increasing transparency and strengthening individuals' right to control and be aware of the processing of their data.

4.2 Sensitivity and security data profiling

MHMD will provide a formal methodology to support privacy related profiling of medical and personal data and adjust correspondent security and privacy provisions. Such methodology will be framed by regulatory analysis and yield indication for policies in those areas where current regulations are not addressing fine grained operational constraints. Hence, MHMD will classify data types and assign them to different security and privacy preserving modules, based on their relevance, sensitivity, risk for the individual, and practical value, and will also craft recommended best practices for the protection of each data type. MHMD's privacy profiling methodology and related privacy preserving execution flow will impact both the way that privacy related options are communicated to data subjects (providing a clear, easily understandable privacy preservation scale per type and method) and the way that privacy preservation techniques are applied (ensuring that engineers can easily understand how to build privacy-friendly applications implementing the concepts of Privacy by design and Privacy by default principles in practice).

4.3 Software development

All software modules will encapsulate state-of-the art security, authentication and authorization mechanisms. The robustness of such modules is ensured by years of developments in the field (the basic building-blocks stem from previously funded EU projects or from already functioning commercial solutions) and will be tested through dedicated penetration / hacking tests and challenges. In addition, data protection methods will be made available through a set of secure APIs and Smart Contracts.

4.4 Fingerprinting and watermarking

MHMD's internal monitoring functions will be paired with scanning and tracking functionalities, capable of identifying data that were leaked or fraudulently acquired, by making use of fingerprinting and watermarking as a reactive method, i.e. as means to discover and attribute data leakages. Watermarks embed a unique identification feature to the dataset, allowing to determine data identity and provenance. Fingerprinting is

similar to watermarking, but is further personalised to a specific user of a dataset, thus allowing to identify the specific source a dataset has been obtained from.

4.5 Penetration/hacking challenges

MHMD will organize penetration/hacking challenges, open to the participation of external competitors. Self-hacking tests are also foreseen. For these penetration challenges only synthetic datasets will be used. Both penetration tests and patient re-identification scenarios will be executed to thoroughly stress test the infrastructure, software and platform functions.

5 ETHICAL ASPECTS

To be covered in the context of the ethics review, ethics section of DoA and ethics deliverables.