Horizon 2020 Project LETHE
“A personalized prediction and intervention model for early detection and reduction of risk factors causing dementia, based on AI and distributed Machine Learning.”

Research and Innovation Action
H2020-SC1-DTH-2020-1
GA 101017405

Duration: 48 months from 01/01/2021
Coordinator: Sten Hanke, FH JOANNEUM GESELLSCHAFT MBH

Deliverable ID.:	**D6.1**	
Deliverable title:	**Data Management plan and ORDP I**	
Planned delivery date:	30/06/2021 (M6)	
Actual delivery date:	30/06/2021 (M6)	
Editor:	Lelia Ataliani, Christos Koziaris, Takis Kotis (INFO)	
Contributing partners:	EGI, FORTH, KA, FHJ, KI, THL, UNIPG, MUW	
Internal reviewer:	Tiia Ngandu	
Checked and released by:	Lelia Ataliani – INFO	
Dissemination Level:	**X**	PU = Public;
		CO = Confidential
		CI = Classified

Document information and history

Deliverable description (from DoA)
This deliverable provides a general outline of the project policy and strategy for data management and describes the data management life cycle for all data sets that will be collected, processed or generated by the LETHE project. It also sums up and presents in a concise manner concepts about Fair data, project datasets, data security and compliance to GDPR and adherence to privacy principles regarding data collection and processing. The Deliverable includes also the selection of Research data to be provided with open access according to OpenAire guidelines.

Please refer to the Project Quality Handbook for guidance on the review process and the release numbering scheme to be used in the project.

Release number	Release date	Author [Person and Organisation]	Milestone*	Release description /changes made
V. 0.1	22/02/2021	Lelia Ataliani, Christos Koziaris Takis Kotis (INFO)	TOC	
V.0.2	01/03/2021	Christos Koziaris (INFO)		Updated data security related issues for LETHE system and protection of personal data.
V.0.3	09/03/2021	Lelia Ataliani, Takis Kotis (INFO)		Updated Data Summary and Data Management Strategy
V.0.4	05/04/2021	Christos Koziaris (INFO)		Updated Data Security section
V.0.5	28/04/2021	Medical Partners		Updated data fields and data description for retrospective data

V.0.6	18/05/2021	Takis Kotis (INFO)		Updated Fair data section
V.0.7	26/05/2021	Georgia Karanasiou (FORTH)		Overall comments to the deliverable and its contents
V.0.8	17/6/2021	Lelia Ataliani (INFO)		Updated sections regarding datasets, fair data and metadata
V.0.9	21/6/2021	Ngandu Tiia (THL)	Approved	Overall comments and suggestions from internal reviewer
To be submitted	30/6/2021	Lelia Ataliani, Christos Koziaris Takis Kotis Miltiadis Anastasiadis (INFO)	Revised	Final Version

** The project uses a multi-stage internal review and release process, with defined milestones. Milestone names include abbreviations/terms as follows:*

- *TOC = ”Table of Contents” (describes planned contents of different sections);*
- *Intermediate: Document is approximately 50% complete – review checkpoint;*
- *ER = “External Release” (i.e. to commission and reviewers);*
- *Proposed: document authors submit for internal review;*
- *Revised: document authors produce new version in response to internal reviewer comments*
- *Approved: Internal project reviewers accept the document.*

Table of Contents

1 Executive Summary

This document presents the project's Data Management Plan (DMP). Within LETHE, data management will be given high importance. The Data Management Plan (DMP) will give details about the data collected, processed or generated by the project. The DMP provides an analysis of the various datasets that: a) either exist, b) will be designed and implemented within the course of the LETHE project and c) datasets that will be produced, processed and analyzed during the pilot running phase. **This version of the document reflects the current state of the datasets paving the way for further updates during the lifecycle of the project based on the feedback received from pilot implementations, testing and evaluation.**

This report is a public deliverable targeted at the members of the Project and the wide audience interested in a project like LETHE. As such, the project partners will be able to familiarise themselves with LETHE Data Management Plan, and effectively contribute to its implementation. The aim of this report is to determine the objectives and procedures of LETHE Data Management Plan.

Within this document the methods & conventions, as well as the recommendations for categorising about the use, manipulation and inclusion of data sets in the LETHE project are presented. Moreover, it refers to regulatory aspects and operational information related to contact, personnel profiles details and about the ownership of the data within the project.
Finally, it serves as a guide for the partners of LETHE about the data lifecycle with respect to the creation, identification, caption and description, storage, preservation (including security and privacy), accessibility, discovery and analysis, re-use and transformation of data in the context of the different deployment sites.

Finally, the sections related to intellectual property rights (IPR) are included with the objective to identify the significant aspects of the project. Subsequently, the contents of the DMP are in full accordance with the signed Grant Agreement of the LETHE project with respect to EU Horizon 2020 recommendations, while the information provided will define common grounds in relation to the management of the data. In this respect, it is expected to motivate the participation and collaboration within the LETHE consortium partners and amongst external partners or participants in the project.

2 INTRODUCTION

This document has been prepared to outline how the research data gathered or generated during and after the project development will be managed. It describes the standards and the methodology for the data collection and generation, when and how they will be shared. This document follows the guidelines provided from the European Commission.

The aim of the Data Management Plan (DMP) is to consider the different aspects of data management from the beginning of the project to ensure that outcomes are well-managed in the present and prepared for preservation in the future. Within this document the methods & conventions, as well as the recommendations for categorising about the use, manipulation and inclusion of data sets in the LETHE project are presented.

Even if the document is due in M6 and project activities are at the beginning, a tentative description of the expected resources / data collected / data generated will be carried out, trying to present what data will be kept confidential and what data will be instead made available during project development.

In particular, this document specifies how LETHE research data will be handled in the framework of the project as well as after its completion. As a matter of fact, the report will indicate:

- what data will be collected, processed and/or created and from whom;
- which data will be shared and which one will be maintained confidential;
- how and where the data will be stored during the project;
- which backup strategy will be applied for safely maintaining the data;
- how the data will be preserved after the end of the project.

LETHE will make use of Open Access initiatives and in particular on the Open Research Data Pilot (https://www.openaire.eu/what-is-the-open-research-data-pilot) aiming at ensuring open access to discoverable data and scientific publications generated throughout the project lifecycle.

The Data Management Plan has to be considered as a living document, and any future update or change in the LETHE data management policy and/or dataset created will be included in the periodic reports or will be specified in the deliverables related to the specific tasks. The consortium foresees 2 more versions for the DMP – one on M24 and one on M42 – taking also into account the IPR strategy that will be defined for the LETHE exploitable results. Actually, the Consortium includes partners of various disciplines (universities, research organizations, IT companies, other non-research organizations, medical institutions and hospitals), aiming to design and develop innovative products and services through the project research and in this context, a proper management of the data generated is key in order to accelerate the uptake and diffusion of project results and outcomes to the market.

Moreover, this deliverable reports a preliminary strategy for the ethics and proper management of data generated in the framework of LETHE project activities, data (questionnaires, wearables data, etc) that is generated and collected from the participation of humans in the pilot studies foreseen in the project description of work.

2.1 Purpose and scope

DMP describes the data management life cycle for the data to be collected, processed and/or generated by the LETHE Horizon 2020 project. DMPs are a key element of good data manipulation and exchange. More specifically, the DMP shall be generated based on the EU Commission guidelines regarding the management of data requirements. According to these guidelines, the data that is going to be shared for scientific and commercial purposes should be easily discoverable, accessible, and intelligible.

Thus, the purpose of this deliverable (D6.1 - Data Management Plan) is to provide an analysis of the main elements of the data management strategy that will be used by the consortium regarding all the datasets that will be generated and/or collected by the project consortium.

2.2 EU Commission guidelines for data management

The basic guidelines [2] for data management plans are set up by the EU. In the table below LETHE summarizes the DMP components considered and analyzed as well as several issues addressed per component.

DMP Component	Issues to be addressed
Data Management Strategy	➢ State the purpose of the data collection/generation ➢ Specify the types and formats of data generated/collected ➢ Specify if existing data is being re-used (if any) ➢ Specify the origin of the data ➢ State the expected size of the data (if known) ➢ Exportation, Exploitation, availability of data ➢ Protection of Personal Data ➢ LETHE and GDPR compliance ➢ Gender Issues ➢ Ethical and Legal Issues
FAIR Data. Making data findable, including provisions for metadata.	➢ Making data findable ➢ Making Data Openly Accessible ➢ Research Data ➢ Making data interoperable ➢ Increase data re-use
Project Datasets	➢ Information about the consortium ➢ Project files ➢ Research activities ➢ Development data, implementations and codes

	➢ Pilot and testing activities
Allocation of resources	➢ Estimate the costs for making your data FAIR. Describe how you intend to cover these costs ➢ Clearly identify responsibilities for data management in your project ➢ Describe costs and potential value of long-term preservation
Data security	➢ Organizational measures ➢ Technical measures ➢ Data recovery, secure storage and transfer of sensitive data ➢ Safe storage for long term preservation
Other	Refer to other national/funder/sectorial/departmental procedures for data management that you are using (if any)
Data Governance	➢ Data governance model ➢ Data sharing pools

2.3 Internal communication tools

Communication inside the consortium is through mails – personal and mailing lists - and by using Microsoft Sharepoint community portal. For WP6 – Infotrend is the leading partner, thoroughly assisted and helped by other partners according to each partner expertise.

2.4 List of definitions & abbreviations

Term	
XML	eXtensive Markup Language
AaS	As a Service
DOI	Data Object Identifier
URI	Uniform Resource Identifier
DPIA	Data Protection Impact Assessment
GDPR	General Data Protection Regulation
ICT	Information & Communication Technology
PIA	Privacy Impact Assessment
PITR	Point-In-Time Recovery
POC	Person of Concern
RBAC	Role-Based Access Control

SOP	System Operating Procedure
VM	Virtual Machine
SOPs	Standard Operating Procedures

2.5 Terms and definitions

1.

“Anonymization” is the process of removing or modifying all personal identifiers and codes in such a way that individual data subjects cannot be identified and there is no reasonable likelihood that identification could take place based on the data, alone or in combination with other data.

“Data Protection Impact Assessment” is a tool and a process to assess the protection impacts on data subjects in processing their data, and for identifying remedial actions to avoid or minimize such impacts.

“Data Subject” means any individual falling within the scope of the Data Protection requirements whose personal data is subject to processing by Lethe.

“Data Controller” The natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data; where the purposes and means of such processing are determined by Union or Member State law, the controller or the specific criteria for its nomination may be provided for by Union or Member State law (article 4 n. 7 GDPR).

“Data Processor” The natural or legal person, public authority, agency or other body which processes personal data on behalf of the controller (article 4 n. 8 GDPR).

“Consent of the Data Subject” Any freely given, specific, informed and unambiguous indication of the data subject's wishes by which he or she, by a statement or by a clear affirmative action, signifies agreement to the processing of personal data relating to him or her (article 4 n. 11 GDPR).

“Lethe’s premises” is the term is used to point out those locations or premises of interested parties and contributors that are used for usage by the Lethe project and may host information systems, backups, databases, hard copy archives, etc. As Lethe is defined as a joint project by many interested parties and contributors there is no actual central premises.

“Personal data” is any data related to an individual who can be identified from that data; from that data and other information; or by means reasonably likely to be used related to that data. Personal data include biographical data (biodata) such as name, sex, marital status, date and place of birth, country of origin, country of asylum, individual registration number, occupation, religion and ethnicity, biometric data such as a photograph, fingerprint, facial or iris image, as well as any expression of opinion about the individual, such as assessments of the status, health data and/or specific needs. (The European Parliament and of the Council of the EU, 2016)

“Privacy by design and by default” means that the organization has to integrate or 'bake in' data protection into its processing activities and business practices, from the design stage right through the lifecycle. In other words, “Privacy by Design” states that any action an organization undertakes that involves processing personal data must be done with data protection and privacy in mind at every step. If more information than necessary to provide the service is disclosed, then "Privacy by Default" has been breached.

“Processing of personal data” means any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organization, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available to any party, alignment or combination, restriction, erasure or destruction. (The European Parliament and of the Council of the EU, 2016).

“Pseudonymization” is a process when modifying data so that it remains associated to a particular individual data subject without that individual being identified. This is done by assigning reference codes or pseudonyms to individual data subjects in lieu of their personally identifying data. These codes are kept separately and are subject to technical and organizational measures to ensure that data is not attributable to an identified or identifiable data subject.

3 Data Management Strategy

The general strategy for data management, according to the **Guidelines on Data Management in Horizon 2020 [1, 2]** will be based on the identification and classification of data generated and collected, standards and metadata to be used, exploitation and availability of data as well as how the data will be shared and archived, the preservation of the information as well as the ethical, legal compliance and the responsibilities in the implementation of the DMP. The LETHE DMP will cover all the data life cycle following the **H2020 guidelines regarding Open Research Data [3].** To formulate an effective DMP, which helps to keep track of the varieties of data generated by the project, it is useful to categorize such data according to their form, source and type.

The data that will be collected, processed and generated during the project is diverse, but will fall into the following broad categories:

- Existing historical data sources – retrospective medical data – from patients' medical records and data from existing research projects.
- Data created by or added by users as part of their participatory activities.
- Data captured through user interaction with technologies like IoT devices, robots, questionnaires.
- Qualitative data about user interactions with the platform and user involvement with the project more broadly, obtained by a mixture of observation, surveys / questionnaires, and semi-structured interviews with participants.
- Processed data resulting from the application of big data analytics to combinations of the data listed above.
- Project data derived from research and development and presented in respective deliverables
- Dissemination activities data and exploitation data generated from project running and execution.

3.1 Expected Data Management Plan

Data Management is a complex process and all partners need to adhere to certain rules and regulations. In particular, each responsible partner will be requested to provide the following information regarding datasets in LETHE:

- **NAME:** name data/metadata/dataset;
- **DESCRIPTION:** brief description of data/metadata/dataset;

- **CREATED:** each partner should indicate if data/metadata/dataset was (or will be) created during the project (Yes/No);
- **GATHERED:** each partner should indicate if data/metadata/dataset was (or will be) collected from other sources (Yes/No);
- **TYPE:** each partner should indicate the type of data/metadata/dataset selecting some of the following options (more than option is possible): Document, Video, Images, Source code/Software, Algorithm, Raw Data, Dissemination material, etc;
- **FORMAT:** each partner should indicate the file extension of data/metadata/dataset (.pdf, .xls, .mat, specific customized format etc) and if a description of the data is available for its use
- **SIZE:** each partner should indicate the file extension the file dimension of data/metadata/dataset (order of magnitude: KB, MB or GB);
- **OWNER:** the lead beneficiary of the specific data/metadata/dataset (or "external" if the owner is not part of LETHE consortium) has to be indicated;
- **DISSEMINATION LEVEL:** each partner should indicate the dissemination level of the specific data/metadata/dataset collected or created during the project, by selecting one of the followings: Confidential, Public, Consortium, etc;
- **REPOSITORY DURING THE PROJECT (FOR PRIVATE/PUBLIC ACCESS):** each partner should indicate the location of data/metadata/dataset collected or created during the project, by selecting among LETHE file repository (MicrosoftTeams), open access repositories, partner repository (private cloud/ drop box/ internal area), etc;
- **BACK-UP FREQUENCY:** it refers to the frequency of updating data/metadata/dataset collected or created during the project (daily, monthly, yearly etc.)
- **REPOSITORY AFTER THE PROJECT:** the location of data/metadata/dataset collected or created during the project after its conclusion, by selecting among LETHE file repository (Nextcloud), open access repositories, partner repository (private cloud/ drop box/ internal area), etc;
- **PRESERVATION AFTER THE END OF THE PROJECT (IN YEARS):** if data/metadata/dataset collected or created during the project will be maintained, each partner must define for how many years they will be available.

A table template for all above variables will be created, stored and maintained on Teams project repository and all partners will update this table starting from M12 of project on 6-month level at least so the consortium keeps track of the progress and status of data collection. Infotrend will be the partner responsible for managing this table.

3.2 What is the purpose of the data collection/generation and its relation to the objectives of the project?

LETHE will provide a data driven risk factor prediction model for elderly individuals at risk of cognitive decline building upon big data analysis of cross sectional and longitudinal, observational and intervention datasets from 4 clinical centers in Europe. These retrospective datasets from the clinical partners will be used to initially feed the LETHE platform after being harmonized and creating a new master dataset of all retrospective datasets, that will in turn feed the risk factor prediction model.

Apart from already existing datasets that will be used, the project involves carrying out data collection (in the context of the trials and pilots) to support the subsequent evaluation of the platform in real-world contexts. Personal data will be collected and also pseydonymised data is still personal data. Personal identifying information will be kept only in the research center that collects the data, and then psudonymised data will be shared with those LETHE partners that need it for their tasks. This applies to both retrospective data and to new trial data. **All data collected – during the pilot trials - will be pseudo-anonymized and abstracted in a way that will not affect the final project outcome**. Finally, data collected will also be used for research purposes, through an analysis that will be carried out by research institutions of the project.

3.3 Will you re-use any existing data and how?

YES existing data will be re-used. A key part of LETHE project is to provide an initial model suite for progression of dementia indicators as well as related risk factors **based on retrospective clinical data – from medical centers / hospitals that are LETHE partners - and data from established studies like project FINGER.** This process is needed and is a key aspect in designing the Artificial and Machine Learning algorithms so that features and essential information from the existing data sources can be extracted.

In order to achieve this goal a homogenized master data record will be created – from the existing data sources that will be re-used - that will form the basis for creating the AI/ML algorithms and training programs. In order to effectively achieve this, certain steps will be taken and legally binded so that the end result cannot be confronted by any partner during project execution. These steps – fully explaining how re-use of existing data will take place - are:

1. Identify retrospective data sources to be used.
2. Partners – owners of retrospective data to extract what data is to be provided and implement pseudo-anonymization techniques.
3. Establish a secure and safe method of transmitting above data.
4. Create a database infrastructure to store collected retrospective multivariate and multimodal data.
5. Apply ETL (extract – transform – load) based techniques to raw collected and stored datasets.
6. Apply homogenization and harmonization techniques.
7. Define the structure and format of the joined master dataset.
8. Unify data records taken from various sources to an extent, that they can be used as a

joint basis for building predictive models.

9. Map existing datasets from retrospective data sources to the master dataset.
10. Find features which can be used for an initial prediction model suite in WP3 and WP4.

3.4 What is the origin of the data?

Data in LETHE is a key issue and a cornerstone of the work to be done and presented. Data in LETHE comprises of the following key areas:

- ➢ Data regarding project documents, mails, memos, minutes of meeting, presentations, research papers, etc analyzed in chapters 5.1 and 5.2
- ➢ Retrospective Data existing already in hospitals and medical centers and which is provided in LETHE for further analysis. Presented in Appendix 12.2
- ➢ Common master dataset – after homogenization of original retrospective data, presented in chapter 5.3.1
- ➢ Prospective data that will be collected from trial participants via measurements, questionnaires, interviews, as well as wearables, mobile application and cardiovascular sensing system presented in chapter 5.3.2
- ➢ Data from TEMI robot,
- ➢ cCOG web test leading to the CAIDE score matrix – presented in Appendix 12.3
- ➢ Data processed and is output of LETHE system.

Summarizing the origin of raw data sources are:

- Organizations, Systems, Devices and Applications that will provide the raw data required by LETHE project
- During the first phase of the project, the Clinical Partners will provide the Retrospective Data using different datasets, as csv/xls/json files.
- During the last phase, raw data will be:
  - Clinical Partners will provide additional Retrospective data coming from follow up protocols;
  - Sensors of the wearable devices will produce IOT data;
  - TEMI robot will provide data
  - Mobile Apps and other Web Apps (including cCoach and cTrain) will generate data as a result of interactions with the patients (tests, questions, etc).

3.5 Data to be collected/ generated

The overall data to be collected, processed, extracted, generated and stored in LETHE databases can be grouped in the following categories:

- ***DB1 Information about the consortium***: Data about the consortium, such as personal information, emails, deliverables, meeting minutes etc are handled and stored in a private and secure storage infrastructure (FHJ data center operating Microsoft Sharepoint Community Portal). Access is restricted to the members of the consortium.
- ***DB2 Project files****:* In this regard, all data gathered from meetings, workshops, and any type of internal communication will be protected according to each required level of confidentially (i.e. stored using the local cloud repository as in DB1). Fruitful and general outcomes of the project will be disseminated without restriction provided that no sensitive data is disclosed. In the case of confidential discussions or outcomes (e.g. EU Restricted deliverables), only specific partners will have access and files will be stored and encrypted in each corresponding organization premises (in the case of sensitive data).
- ***DB3 Research activities****:* LETHE research activities and their corresponding outcomes (deliverables, publications) are another source of data to be catered for. State-of-the-art methods and public resources will be used to carry such activities. In the case of research that involves data gathering from other platforms, such data will be stored and protected locally by the corresponding organization. In the case that specific materials need to be shared, each partner will provide their own restrictions and policies (according to national and EU laws), which should be agreed by the interested members. In each case, proper anonymization and/or encryption mechanisms will be applied to guarantee that any personal or sensitive data is disclosed. EU restricted data will be managed accordingly. These procedures will comply with LETHE ethics reports and deliverables, especially with regards to sensitive data storage and mass surveillance policies reported in deliverables.
- ***DB4 Development data, implementations and codes****:* Development and codes, as well as implementations derived from the project, will be performed in a private repository that will be defined on M12 (eg GitLab). Periodic versioning backups will be made for each module of the LETHE project, which will be stored in Gitlab.
- ***DB5 Pilot and testing activities***: In the case of pilots, the information of the users involved and the processed data will be stored and managed locally by each organization. Therefore, in the case of an end user using their own repositories their local policies and data management restrictions will apply. In any case, data will be deleted when the corresponding validation finishes

except if otherwise decided by partners and in cooperation with the PO. In addition, the outcomes of such pilots will be reported anonymized if some sensitive data needs to be shared. Other types of users and participants may be considered during the project lifetime, according to new system requirements, whose data will be stored and managed accordingly.

A global overview of the different data levels present in the project databases is depicted in the figure below. In this regard, the corresponding clarifications and descriptions are presented in the next sections.

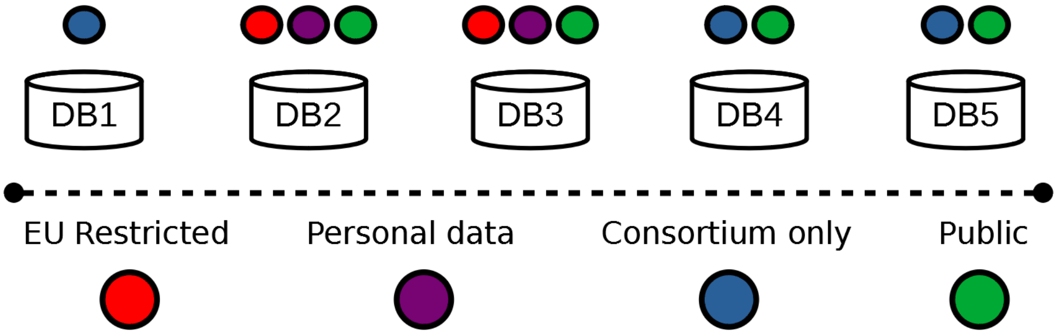

Figure – Correlation between LETHE databases contents and data restriction levels.

3.6 Guidance on data formats

- What format will your data be in?
  - Wherever possible LETHE will use data in open formats. We will strive to use non-proprietary formats using standard representations (Unicode) and meet the following requirements:
    1. Non-proprietary
    2. Open, documented standard
    3. Common usage by research community
    4. Standard representation (ASCII, Unicode)
    5. Uncompressed
  - Collected data from 4 medical centers will be delivered in either .csv or .xls format.
  - The format of the common master dataset to be further processed will be either numerical or sometimes alphanumerical.
  - Long-form surveys and questionnaires and observational data will generally be collected and will be mainly numerical or perhaps sometimes alphanumerical. Text and images are much more rare for both retrospective and trial data.

- Why have you chosen to use particular formats?
- We have a preference for open and commonly-used data formats for interoperability with existing analysis tools and long-term usability.
- Do the chosen formats and software enable sharing and long-term validity of data?
    - Yes: open data formats allow all consortium partners to easily access the data. Open formats can also anticipate greater sharing and access among a wider audience since they do not rely on a specific (generally costly) software to read them. Open data formats are the standards accepted by, and in widespread usage among several consortium partners and the intended user communities. Finally, using open lossless formats ensures the long-term usability of data as these standards are maintained by a robust community of users.

3.7 What is the expected size of the data?

The expected size of retrospective data – as described in chapter 3.4 - to be collected, generated, processed and stored per medical center is on the order of 10x Mbytes. The same holds for the common master dataset (homogenized dataset) that will be the prime dataset to be processed and analysed.

It must be though stressed that EGI data center to be used for handling respectively the various datasets has the necessary infrastructure to host and store large volumes of data.

3.8 Exportation, Exploitation, availability of data and re-use

In general, LETHE, following the requirements of disseminating the results and developments of the project, is willing to make open access as much material as possible, and open access to data will be the default option for the consortium especially with relation to dissemination tasks in WP9. Therefore, deliverables, publications and reports will be made public whenever possible. Nevertheless, as the project will engage individuals and respectively, process personal data, open access to the deliverables and project data might evoke potential issues (all ethical and legal aspects will be thoroughly examined by the Project Ethical Manager, Joint Data Controllers and LETHE DPO Security Officer). Therefore, the consortium will carefully examine to what extent data can be publicly shared.

Especially in the deliverables D7.6 Assessment and evaluation of pilots I and D7.7 Assessment and evaluation of pilots II, the data subjects and the GDPR framework will be presented. In these deliverables an analytical description of how the LETHE project implements the GDPR regulations is/will be included, along with the data subjects involved in testing and pilot operations, the types and sources of data that will be collected and processed, the data controllers, the data processors, the processing activities of the personal data that are taking place in the Lethe platform, and how the use of such technologies will protect the rights of both tested subjects and end-users (data recipients) in general.

According to Article 4 of the EU GDPR, the Data Controller, the Data Processor and the Data Recipient are defined as follows:

- Controller – "means the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data".
- Processor – "means a natural or legal person, public authority, agency or other body which processes personal data on behalf of the controller".
- Recipient – "means a natural or legal person, public authority, agency or another body, to which the personal data are disclosed, whether a third party or not".

Regarding data exportation, **LETHE will not transfer data to third countries**. Personal data will be processed in the territory of the European Union. However, the protection of their personal data will conform the EU data protection rules. Partners ensure that the implemented technical and organizational measures facilitate the adequate protection of their rights and freedoms.

3.9 Protection of Personal data

Protection of personal data in LETHE refers to guidelines and provisions that apply to patients' datasets provided by hospitals participating in the Consortium. Public datasets have already been screened for public access and authorization for data use will be achieved at the start of the LETHE project. Regarding public datasets, only datasets relevant for the study and for which the Consortium has achieved use grants from data owners will be included.

A detailed description for the protection of personal data (POPD) is presented ad-hoc in deliverable D10.1 which is specifically asked from the EU and is covering the concept of personal data protection.

3.9.1 *Confidentiality*

All parties involved in this study will maintain the strict confidentiality to assure that neither the person nor the family privacy of the subjects providing data for LETHE is violated; appropriate measures shall be taken to avoid the access of non-authorized persons to the data. GDPR regulations apply.

Investigators must guarantee that all personnel involved in this study will respect the confidentiality of any information concerning the study subjects.

3.9.2 *Privacy*

The provisions of the above mentioned GDPR will be adopted or – if more restrictive – national regulations in matters of personal data protection and privacy. For guidance the project will refer to the EU regulations and best practices as defined by the Data protection rules (https://ec.europa.eu/info/law/law-topic/data-protection_en). All data within LETHE project shall be handled using tools and processes with 'privacy by design' as the mindset. The project will share personal data, with applied internal encryption, with subject's ID recoded in a sequential ID code.

Data from study participants will be de-identified and encoded/pseudonymized using privacy-compliant procedures. In this case, the list matching patient personal data and study ID will be stored off-line in the center which collects the data, which will manage it in compliance to relevant local legislation, under the Data Controller of each Institution and according to GDPR.

The shared data will not contain any sensitive that can contribute to personal identification with the exception of gender and age: any information on name, address, birth date, phone numbers, etc. will not be shared. Phone/email contacts will be used upon consent of participants/tutors as foreseen by the study and only to facilitate communications between study participants/tutors. Participating hospitals hold responsibility to maintain pseudo-anonymity on shared data and on biological samples and to safely store and preserve from undue access any information which may disclose the patient's identity, in accordance to national and EU regulations (GDPR or any more restrictive regulation).

3.9.3 *Protection of data collected through mobile apps*

Data collected through mobile apps will be managed according to GDPR and relevant guidelines [6].

3.9.4 *Integrity*

Data quality and data integrity checks in accordance to GDPR art. 5 and 25, will be performed. Data integration will include checks to ensure that integrated data are not hampered as compared to original datasets. All actions taken by the cloud provider relating to LETHE data and services will be (unchangeably) logged and transparently provided to the Consortium.

3.9.5 *Transparency*

The DoA clearly indicates the process of collecting, transferring and processing data, whether this data is retrospective (existing in the medical partners' systems) or data resulting from using IoT, wearables, robots, questionnaires etc. Data sharing will be clearly defined in the LETHE architecture. Data Transfer Agreements (DTA) will clearly state data localization, data controllers and data processors, data owners/custodians, as well as responsible scientists and persons having access to data.

3.9.6 *Right to access, rectification and deletion of data*

The project will use data already checked and approved by Consortium partners and authorized by patients' informed consent. Non eligible data will be removed upfront before the start of the study.

Notwithstanding the provisions stated above, the data owners/custodians will have the possibility to rectify/delete data in case of patient's withdrawal, ascertained errors in the original dataset. LETHE services will be developed to allow the fulfilment of data subjects' rights, e.g. searching for data to comply with right of access or provide ability to block/delete data records. These procedures shall be managed under the control of Data Controller and the DPO of data providers and data management institutions.

3.9.7 *Data protection*

Data protection rules established by GDPR will apply and will be implemented. Access to data will be secured through a security layer which will include both individual ID and password and, when applicable, more sophisticated authorization and access methods (e.g. biometrics). Anonymous and pseudonymized data will be used for research purposes only. Access to patient's pseudonymized data will only be granted according to each hospital regulations and restrictions. Access will be time-limited and will provide different, individualized clearance levels based on the individual tasks to be performed. Differences in “clearance levels” are related to reading-only or writing privileges, full- or restricted-access to sub-groups of data, etc. Personal identifiable data will not be available nor accessible.
As regards processing personal data, protecting privacy in the electronic communications sector and retaining data generated or processed in connection with the provision of publicly available electronic communications services or of public communications networks (e.g. cloud, big data, open data, cookies etc.), we will ensure that the developed tools comply with the relevant legislation (in particular EU Directive 2002/58/EC and 2006/24/EC).

3.10 LETHE and GDPR Compliance

General Data Protection Regulation (GDPR) is considered and relevant ethical, legal and privacy concerns will be addressed respectively. LETHE management structure will comprise a number of Data Controllers and a person responsible for monitoring Data Protection principles and has significant expertise in GDPR. The DPO will be responsible for overseeing data protection strategy and implementation to ensure compliance with GDPR requirements. The DPO will also be in close collaboration with the Data Protection Officers and Data Controllers of the Beneficiary organizations.

Arrangements will be prepared by the researchers to carefully protect the confidentiality of participants and their data. All personal information collected will be considered privileged information and be dealt with in such a manner as not to compromise the personal dignity of the participant or to infringe upon his/her right to privacy. Before consent is obtained, the researchers will inform prospective participants of any potential risks that might mean that the confidentiality or anonymity of personal information may not be guaranteed and the purpose for which personal information provided will be used.

Regarding the rights to privacy and to the protection of personal data, LETHE will adhere to the provisions of:

- The EU Charter on Fundamental Rights (art. 7 and 8);
- The European Convention for the Protection of Human Rights and Fundamental Freedoms (art. 8);
- The CoE ConvenLi M et al, "Toward privacy-assured and searchable cloud data storage services," Network, IEEE, vol. 27, no. 4, pp. 56–62, 2013tion No. 108 for the Protection of Individuals with regard to Automatic Processing of Personal Data (1981);
- Regulation 2016/67 EC (General Data Protection Regulation)

LETHE is aware that it is likely that some standards, in particular but not exclusively the European data protection framework, may evolve within the lifetime of the project. The consortium is committed to monitoring and taking into account such evolutions as well as issues emerging from different enforcement regulations.

Key principles relating to processing of personal data that will be adhered to include:

- Informed consent, including a description of the purpose of the research, who is organising and funding the research, and explanation about what will happen to the results of the research
- Be fairly and lawfully processed, with a legal basis for processing identified
- Processed for limited and clearly pre-defined purposes
- Be adequate, relevant, and not excessive
- Be accurate and kept up to date
- Not be retained for longer than necessary
- Processed in line with a subject's rights
- Kept secure, thought technical and organisational measures
- Not be transferred to other countries without adequate protection
- Right to withdraw
- Right to be forgotten
- Data minimisation

Whenever participants are requested to submit personal data, they will be informed that this data is stored and processed by the partner(s) in charge of the activity and for the purposes of the project only. They will be also informed about their rights to access, modify and erase their personal data and how to enforce this.

Individual participants' data will not be mined or used for any purpose other than those explicitly and clearly needed for the running of the activity within LETHE. However, if data would be usefully held for future analysis, specific informed consent will be sought from participants for this purpose. Otherwise data will only be held for as long as is required by relevant regulations. Specific measures must be taken to protect collected data (personal or sensitive) and to ensure secure destruction/deletion when the scope of LETHE is fulfilled and, in any case, when requested by any participant or volunteer. Conservation and destruction of data (both physical and digital) must be done according to best practices so that destroyed data cannot be retrieved.

3.11 Ethical and Legal Issues

LETHE Consortium is fully aware of the ethical implications of the proposed research and respects the ethical rules and standards of HORIZON 2020, and those reflected in the Charter of Fundamental Rights of the European Union.

Summarizing, core ethical issues within LETHE will be addressed by fully complying with EU and national legislation. LETHE ethical issues will take into account the following main principles:

- ✓ Ensure transparency on all data collection and management practices performed by the project and notify all people and stakeholders involved;
- ✓ Confirm the (explicit and written) Informed Consent of patients involved in the project pilot evaluation phase – whenever appropriate based on the actual activities to be carried out -, while option to withdraw will be available at any time;
- ✓ Safeguard data protection, security and privacy issues through an integrated security and ethics management policy throughout technologies as well as data management practices in the project’s field of research (amnesia medical analysis and profiles of patient).

A detailed analysis is presented in D10.1

3.12 IPR Issues

A guiding rule for IPR management is that that organizations connected to LETHE network should have an advantage over those who do not. That means that generated knowledge of commercial interest must be safeguarded and protected for exploitation beyond the project. On the other hand, the partners of this project have come together in order to collaborate and benefit from their respective resources and competencies. Thus, added value through knowledge sharing and promoting exploitation are clear objectives and driving forces.

All LETHE consortium partners will ensure from the beginning of the project that their own "pre-existing know-how" (background), which will be used in the project, is identified and recognized by the other participants up front. "Pre-existing know-how" will relate to information developed before the starting of the project, whether it is patented or not, secret or not, as well as to results obtained outside the project after it has started, i.e. in parallel to it. A specific piece of knowledge resulting from the project (foreground) will belong to the contractor who generated it. If such piece of knowledge is jointly generated, it will be jointly owned, unless the concerned contractors agree on a different solution. In general, joint owners will agree among themselves on the allocation and the terms of exercising the ownership of the knowledge.

Transfers of ownership will be allowed but following their communication to the other contractors. Where knowledge to be developed in the project is capable of commercial application and having due regard to the legitimate interests of the partners concerned, it will be protected. It is expected though, that there will be situations where newspaper, scientific magazine, journal publication or other means of disseminating knowledge in the public domain will constitute appropriate alternatives, taking account of the nature of the results and the participants' interests.

This approach to knowledge and IPR management will be detailed and regulated in WP9 - exploitation. The major aspects are:

- Confidentiality: Each partner will treat information from other partners as confidential and not disclose it to third parties unless it is obvious that the information is already publicly available.
- Patents: Partners who will develop patentable knowledge will be encouraged to apply for patent or similar form of protection and shall supply details of each such patent application under preparation to the other partners.
- Access Rights: Partners grant to each of the other partners royalty-free access right to knowledge generated in the project to the extent needed to successfully perform the project. Access rights to a partner preexisting knowledge for use outside the project is, when needed and only to the extent necessary to make use of the project result, given on preferential conditions to the other partners.
- Ownership of Knowledge: Knowledge is owned by the partners who carried out the work generating the knowledge, or on whose behalf such work was carried out. If a partner wishes to assign knowledge to a third party, he should inform the other partners requesting their consent, which should not unreasonably be withheld.
- IP Ownership: Foreground IP shall be owned by the project partner carrying out the work leading to such Foreground IP. If any Foreground IP is created jointly by at least two project partners and it is not possible to distinguish between the contributions of each of the project partners, such work will be jointly owned by the contributing project partners. The same shall apply if, in the course of carrying out work on the project, an invention is made having two or more contributing parties contributing to it, and it is not possible to separate the individual contributions. Any such joint inventions and all related patent applications and patents shall be jointly owned by the contributing parties. Any details concerning the exposure to jointly owned Foreground IP, joint inventions and joint patent applications will be addressed in LETHE.

3.13 Metadata Standards and Data Documentation

In the context of Lethe initiative, and in order to achieve greater transparency in clinical research and, a diverse and expanding number of data is expected to be collected in Lethe DB repository, like Clinical Data, documents or other health data created by clinical studies (collectively known as data objects). To make the best use of such resources, it is also necessary for stakeholders to agree and deploy a simple, consistent metadata scheme.

The relevant data objects and their likely storage are described, and the requirements for metadata to support data sharing in clinical research are identified. **A scheme is proposed that is based heavily on the DataCite standard (The Metadata Working Group, 2015), with extensions to cover the needs of clinical researchers, specifically to provide:**

(a) study identification data, including links to clinical trial registries;

(b) data object characteristics and identifiers; and

(c) data covering location, ownership and access to the data object.

The components of the metadata scheme are described in the following table (Canham & Ohmann, 2016)

Table. Elements in the proposed metadata scheme for clinical research data objects.

Mandatory	Recommended	Optional
A.1 Source study title (3)	A.2 Study identifier records (3)	
	A.3 Study topics (3)	
B.1 DOI (1)	B.5 Version	B.2 Object other identifiers (3)
B.3 Object title		B.4 Object additional titles (3)
C.1 Creators		C.2 Contributors (3)
D.1 Creation year		D.2 Dates (3)
E.1 Resource type general	E.2 Resource type	E.4 Subjects (of data object) (3)
	E.3 Description (3)	

Mandatory	Recommended	Optional
	E.5 Language	
	E.6 Related identifiers (3)	
F.1 Publisher	F.2 Other hosting institutions (3)	F.7 Rights
F.3 Access type		
F.4 Access details (2)		
F.5 Access contact (2)		
F.6 Resources (3)		

Notes

DOI Digital object identifier

(1) Mandatory for publicly accessible data objects, recommended for all others

(2) Mandatory if access is non-public

(3) May be repeating

About LETHE metadata for retrospective and prospective data the consortium proposes to use a scheme based on the DataCite standard and partners are still looking for tools and sw that could help us in applying this standard in the LETHE project. Hence:

- The raw data will be "enriched" adding Metadata for dataset description, dataset lineage, etc.
- The Metadata could be used to build an internal Data Catalog.
- The Metadata could be sent and published to DataCite to improve the open sharing of these datasets, when and where required (FAIR Data Principles requirement of LETHE project).

More detailed descriptions of the metadata will become available as part of the work in various project WPs apart from wp6 like definition of case scenarios, and reference architecture (WP2), platform implementation (WP3, WP4, WP5, WP8) and testing (WP7).

3.14 Consideration of gender aspects

The consortium addresses sex, gender and equality issues during the proposal itself in the constitution of the consortium itself and the Scientific Advisory Board. The consortium will be respectful of gender equality and will ensure there is no discrimination on the basis of a person’s sex during the project’s research and activities. Gender balance will be respected in research teams and taken into account in the decision-making process. Therefore, LETHE will comply with the Charter of Fundamental Rights of the European Union and Directive 2002/73/EC on equal treatment of men and women in employment, vocational training and promotion, and working conditions and will seek to promote the role of women in the scientific and technological field. All partners in LETHE are committed to providing equal opportunities for all, irrespective of gender, colour, race, religion or belief, ethnic or national origins, marital/civil partnership status, sexuality, disability, or age. Please note that the majority of researchers in partners and especially in clinical partners are female.

4 FAIR Data

Part of the role of a DMP is to define a framework concerning the handling of research data generated or acquired as the project progresses, but also after the end of it. Subjects for investigation are: the nature of the data in question, which data will be collected and to whom they will be useful, the use of metadata to render data easily retrievable, standardisation, whether and which data will be open-access, how they will be stored and preserved, etc.

Aiming to actively be part of the Open Research Data Pilot, the Lethe DMP complies with the H2020 guidelines for making data Findable, Accessible, Interoperable, Re-usable (FAIR). To achieve that, the FAIR template provided by the European Commission[1] is followed. In the Guidelines on FAIR Data Management in Horizon 20203, the European Commission states: "Where will the data and associated metadata, documentation and code be deposited? Preference should be given to certified repositories which support open access where possible." For LETHE project these repositories are:

- EGI Federated Cloud at a site located in the EU/EEA area.
- Source code will be provided in a Github repository at the FHJ environment.
- Project documentation in in sharepoint at the FHJ
- All deliverables which are stated as public in the DoA will be available at the webpage

By default, Horizon 2020 projects participate in the Open Research Data Pilot and they must deposit the following data in a research data repository:

1. All data needed to validate the results presented in scientific publications, including the metadata that describe the research data deposited. This is called the "underlying data". In LETHE partners will decide in due time and as project advances which data will be deposited.
2. Any other data (for instance curated data not directly attributable to a publication, or raw data), including the associated metadata, as specified and within the deadlines laid down in the DMP. For LETHE this is viable after month 12 in project implementation.
3. Projects should also provide information via the chosen repository about the tools that are needed to validate the results, e.g. specialised software or software code, algorithms and analysis protocols. Where possible, they should provide these instruments themselves, or alternatively, provide direct access to them.

[1] http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

4.1 Making Data Findable, including Provisions for Metadata

To make the published **data findable, we will first publish the corresponding repositories on GitHub**. Therefore, each repository will be public and contentiously available. Depending on the underlying data, the information will be stored in a standard format, e.g. PCAP files, along with some relevant documentation about the contents. Finally, the README file of each repository will detail the dataset and include the relative keywords to make it more discoverable from search engines.

To make the published datasets discoverable and raise awareness in the scientific community, whenever possible, these datasets will be linked to publications made by the LETHE partners, and proper promotions through social media will be made. Moreover, the datasets will be shared on platforms like ResearchGate1 which will engage more researchers. Notably, ResearchGate issues a Digital Object Identifier (DOI) for such datasets, allowing further persistence.

4.1.1 Are the data produced and/or used in the project discoverable with metadata, identifiable and locatable by means of a standard identification mechanism (e.g. persistent and unique identifiers such as Digital Object Identifiers)?

LETHE data gathered and processed will be stored on secure physical database servers. All information will be stored on tables with indexes and ids. All information retrieved from the relational databases will be done using standard relational database extraction mechanisms like ids, from tables in the database.

4.1.2 What naming conventions do you follow?

During the project we will use filenames that reflect the structure of the trials and pilots. These internal conventions will be specific to the project team. The naming conventions for the published data sets will be finalised in due course as the project advances.

4.1.3 Will search keywords be provided that optimize possibilities for re-use?

Potential users will find out about LETHE data and results through dissemination and exploitation strategies and actions the consortium will employ. The consortium will also create an open space as part of the website where project results and analysis of data will be stored and can be seen by other interested parties apart from the consortium members.

4.1.4 Do you provide clear version numbers?

LETHE software platform will by design incorporate a versioning mechanism. Published datasets will use sequential version numbers if updated, and a new DOI will be assigned.

4.2 Making Data Openly Accessible

LETHE opts for a “green” open access model for scientific and technical publications. Outcomes will be available for access in the LETHE project website, without prejudice of IPR and copyright considerations regarding publications in peer-reviewed journals and conferences. When applicable, the scientific and technical publications will also be made available through public repositories widely known and accessed like Openaire2. The choices will also be made according to the target audience to reach with a publication. Publication and dissemination of any foreground will be granted with the approval of the Consortium, making sure, when applicable, that any period of secrecy needed is respected. Adequate references to the EU shall be given in any dissemination. **LETHE will comply with EU laws and guidelines and provide all information possible in an open access manner. LETHE will provide open access where possible but need – at the same time – to protect the IPRs of the partners as well as the confidentiality of some sensitive and critical information.**

4.2.1 Which data produced and/or used in the project will be made openly available as the default? If certain datasets cannot be shared (or need to be shared under restrictions), explain why, clearly separating legal and contractual reasons from voluntary restrictions.

Researchers, information managers and other stakeholders can rely on a framework of various international certification standards for digital repositories to assess and improve the quality of their work processes and management systems. “Trustworthy Digital Repository” (TDR) is a term often used in this respect.

Beneficiaries must also provide open access, through the repository, to the bibliographic metadata that identify the deposited publication. The purpose of the bibliographic metadata requirement is to make it easier to find publications and ensure that EU funding is acknowledged. Information on EU funding must therefore be included as part of bibliographic metadata so that Horizon 2020 can be properly monitored, statistics produced, and the programme's impact assessed.

To monitor any embargo periods, the publication date and embargo period must be provided. The persistent identifier (for example a Digital Object Identifier) identifies the publication. It enables a link to be provided to an authoritative version of the publication.

Open Access is one of the main principles of Horizon 2020; by Open Access we mean the provision of free of charge online access to scientific information for any user. The beneficiaries’ obligation to granting open access is differentiated between scientific publications and research data.

- Scientific publication: Publication of academic and research work, most often in the form of an article, research paper and otherwise, in scientific journals or in other forms (e.g. textbook, conference proceedings, etc.).
- Research data: This refers to the recorded factual material commonly accepted in the scientific community as necessary to validate research findings. Examples of research data generated from a project like Lethe could include: Questionnaires, Algorithms, Methodologies, Source Code etc.

All participating projects' beneficiaries are required to ensure open access for their peer-reviewed scientific publications relating to their results, as defined in Article 29.2 of the H2020 - General MGA[2]. LETHE will fully comply with this guideline and will publish all actions taken in the course of the project to prove the open data access status that has taken place.

There are two routes to open access for scientific publications[3]:

1. Gold open access / open access publishing - the practice of immediately publishing in open access mode (in open access journals or in 'hybrid' journals), shifting the payment of publication costs from readers' subscriptions to author fees. These costs are usually borne by the researcher's university or research institute or the agency funding the research.
2. Green open-access / self-archiving – the practice of depositing of a published article or a final peer-reviewed manuscript in an open-access online repository (by the author or a representative). A 6-12-month embargo period before the data is granted open-access may be considered appropriate by some scientific publishers.

Therefore, the open access to publications process is as follows:

1. Publications are deposited in online repositories.
2. Open access route is selected.
3. Open access is granted to publications.

[2] H2020 Multi-Beneficiary General Model Grant Agreement v5.0, available at: http://ec.europa.eu/research/participants/data/ref/h2020/mga/gga/h2020-mga-gga-multi_en.pdf

[3] http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-Access-data-management/open-Access_en.htm

Note that the steps mentioned above are not strictly successive, but may occur simultaneously, depending on the selected open-access route and a possible embargo period set by the consortium. Provision for the GDPR[4], the newly enacted EU regulation about data, is also included.

Organisations acquiring and/or processing data of natural persons are required to adopt more robust data management and security systems. At the same time GDPR empowers citizens, by enhancing monitoring and control over their own data. As stated[5]:

1. This Regulation lays down rules relating to the protection of natural persons with regard to the processing of personal data and rules relating to the free movement of personal data.
2. This Regulation protects fundamental rights and freedoms of natural persons and in particular their right to the protection of personal data.
3. The free movement of personal data within the Union shall be neither restricted nor prohibited for reasons connected with the protection of natural persons regarding the processing of personal data.

As previously noted, significant changes on data, which may arise in the course of the project and the development of the platform, are to be reported in the form of new versions of the present deliverable due in M24 and M48.

4.2.2 How will the data be made accessible (e.g. by deposition in a repository)?

Regarding research data for projects participating in the Open Research Data (ORD) pilot, it is obligatory to ensure open-access to all data needed for result validation. Whether other parts of data will be made open-access, is left to the discretion of the beneficiaries, as they must ensure that the main objective of the project will not be jeopardised by the publicity. Ethical and privacy concerns raised by publication of particular data, as well as protection of Intellectual Property Rights (IPR) are also a great deterrent to granting open access. Justification for excluding particular parts of data from being open access must be included in the DMP. The open-access research data must be deposited in online repositories, available for access, mining, exploiting, processing and disseminating, free of charge for any user, accompanied by the appropriate information —

[4] General Data Protection Regulation, available at: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=EN

[5] Article 1, GDPR

via the repository — regarding the specific tools and instruments that beneficiaries have at their disposal, considered to be necessary for validating the results. Where possible, these tools or instruments should be provided.

4.2.3 Open access in the research data pilot

Horizon2020 has launched an Open Research Data Pilot (ORDP) aiming at improving and maximising access to and re-use of research data generated by projects (e.g. from experiments, simulations and surveys). These data are typically small sets, scattered across repositories and hard drives throughout Europe. The success of the EC's Open Data Pilot is therefore dependent on support and infrastructures that acknowledge disciplinary approaches on institutional, national, and European levels. The pilot is an excellent opportunity to stimulate and nurture the data-sharing ecosystem and has the potential to connect researchers interested in sharing and re-using data with the relevant services within their institutions (library, IT services), data centres and data scientists. The pilot should serve to promote the value of data sharing to both researchers and funders, as well as to forge connections between the various players in the ecosystem.

LETHE project recognizes the value of regulating research data management issues. Accordingly, in line with the rules laid down in the Model Grant Agreement, the beneficiaries will deposit the underlying research data needed to validate the results presented in the deposited scientific publications in a clear and transparent manner.

Open Research Data Pilot project aims at supporting researchers in the management of research data throughout their whole lifecycle, providing answers to key issues such as “what”, “where”, “when”, “how” and “who”.

WHAT
The Open Data Pilot covers all research data and associated metadata resulting from EC-funded projects, if they serve as evidence for publicly available project reports and deliverables and/or peer reviewed publications. To support discovery and monitoring of research outputs, metadata and publications will be made available in Open Access. Individual level research data will most likely not be openly available. Data repositories might consider supporting the storage of related project deliverables and reports, in addition to research data.
WHERE

All research data has to be registered and deposited into at least one open data repository. This repository should: provide public access to the research data, where necessary after user registration; enable data citation through persistent identifiers; link research data to related publications (eg. journals, data journals, reports, working papers); support acknowledgement of research funding within metadata elements; offer the possibility to link to software archives; provide its metadata in a technically and legally open format for European and global re-use by data catalogues and third-party service providers based on wide-spread metadata standards and interoperability guidelines. Data should be deposited in trusted data repositories. These repositories should provide reliable long-term access to managed digital resources and be endorsed by the respective disciplinary community and/or the journal(s) in which related results will be published (e.g., Data Seal of Approval, ISO Trusted Digital Repository Checklist).
WHEN
Research data related to research publications should be made available to the reviewers in the review process. In parallel to the release of the publication, the underlying research data should be made accessible through an Open Data repository. If the project has produced further research datasets (i.e. not necessarily related to publications) these should be registered and deposited as soon as possible, and made openly accessible as soon as possible, at least at the point in time when used as evidence in the context of publications.
HOW
The use of appropriate licenses for Open Data is highly recommended (e.g. Creative Commons CC0, Open Data Commons Open Database License). More specifically the consortium will follow a policy to publish open source (as written in the DoA) and also certain models might be published as Juypter notebooks (like written in the DoA).
WHO
Responsibility for the deposit of research data resulting from the project lies with the project coordinator and partners producing research related results.

4.2.4 Research Data Repository

All data collected during the project will be in the first instance stored and preserved in an online data repository/cloud platform linked to the project website with access limited to the LETHE Consortium, managed by FHJ and Info and intended for internal uses. Particular attention will be paid to the confidential and/or sensitive data and the consortium will not disclose or share this information to third parties.

In the internal LETHE Consortium Repository a specific folder has been dedicated for the collection of data to be included in the future LETHE Open Research Data Platform.

Furthermore, it is important to remark that this Data Management Plan will be updated at each reporting period.

Concerning the open access of discoverable data, different online public repository possibilities will be investigated in subsequent stages of the project. Some examples of suitable repositories under evaluation are shown below:

- ZENODO (http://www.zenodo.org/) is the open access repository of OpenAIRE (the Open Access Infrastructure for Research in Europe, https://www.openaire.eu/). The goal of OpenAIRE portal is to make as much European funded research output as possible available to all. Institutional repositories are typically linked to it. Moreover, dedicated pages per project are visible on the OpenAIRE portal, making research output (whether it is publications, datasets or project information) accessible through the portal. This is possible due to the bibliographic metadata that must accompany each publication.
- LIBER (www.libereurope.eu) supports libraries in the development of institutional research data management policies and services. It also enables the exchange of experiences and good practices across Europe. Institutional infrastructures and support services are an emerging area and will be linked to national and international infrastructure and funder policies. Building capacities and skills, as well as creating a culture of incentives for collaboration on research data, management are the core targets of LIBER.

4.3 Research Data

Research data refers to data that is collected, observed, or created within a project for purposes of analysis and to produce original research results. Data are plain facts. When they are processed, organized, structured and interpreted to determine their true meaning, they become useful and they are called information.

In a research context, research data can be divided into different categories, depending on their purpose and on the process through which they are generated. It is possible to have:

- Observational data, which are captured in real-time, for example, sensor data, survey data, sample data;
- Simulation data, generated from test or numerical models;
- Derived data.

Research data may include all of the following formats:

- Text or word documents, spreadsheets
- Laboratory notebooks, field notebooks, diaries
- Questionnaire, transcripts, codebooks
- Audiotapes, videotapes
- Photographs, films,
- Test responses
- Slides, specimen, samples
- Collection of digital objects acquired and generated during the research process
- Data files
- Database contents
- Models, algorithms, scripts
- Contents of software application such as input, output, log files, simulations
- Methodologies and workflows
- Standard operating procedures and protocols

4.3.1 Key principle for open access to research data

According to the “Guidelines on FAIR Data Management in Horizon 2020”, research data must be findable, accessible, interoperable, re-usable[6].

The findable, accessible, interoperable and reusable (FAIR) guiding principles are reported in the following table[7].

FINDABLE
F1 (meta)data is assigned a globally unique and eternally persistent identifier

[6] http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

[7] http://www.nature.com/articles/sdata201618

F2 data is described with rich metadata **F3** (meta)data is registered or indexed in a searchable resource **F4** metadata specify the data identifier
ACCESSIBLE
A1 (meta)data is retrievable by their identifier using a standardized communications protocol **A1.1** the protocol is open, free, and universally implementable **A1.2** the protocol allows for an authentication and authorization procedure, where necessary. **A2** metadata are accessible, even when the data is no longer available
INTEROPERABLE
I1 (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation **I2** (meta)data use vocabularies that follow FAIR principles **I3** (meta)data include qualified references to other (meta)data.
RE-USABLE
R1 meta(data) have a plurality of accurate and relevant attributes. **R1.1** (meta)data is released with a clear and accessible data usage license **R1.2** (meta)data is associated with their provenance **R1.3** (meta)data meet domain-relevant community standards

4.3.2 Roadmap and procedures for data sharing

LETHE will generate a relevant amount of data mainly determining predictive and prevention models for cognitive impairment and dementia. Part of these data could be made available not only for the purpose of the project, but also for other tools and studies and presented in a specific section of the project website.

According to the aforementioned principles information on data management is disclosed by detailing the next elements:

- ○ **Data set reference and name:** Identifier for the data set to be produced
- ○ **Data set description:** its origin (in case it is collected), nature and scale and

to whom it could be useful, whether it underpins a scientific publication. Information on the existence (or not) of similar data and the possibilities for integration and reuse will be also included.

- **Standards and metadata:** reference to existing suitable standards of the discipline. If these do not exist, an outline on how and what metadata will be created has to be given.
- **Data sharing:** Description of how data will be shared, including access procedures, embargo periods (if any), outlines of technical mechanisms for dissemination and necessary software and other tools for enabling re-use, and definition of whether access will be widely open or restricted to specific groups. The repository where data will be stored will be identified, if already existing, indicating in particular the type of repository (institutional, standard repository for the discipline, etc.). In case the dataset cannot be shared, the reasons for this should be mentioned (e.g. ethical, IP, privacy related, security-related etc.).
- **Archiving and preservation** (including storage and backup): Procedures that will be put in place for long-term preservation of the data. Indication of how long the data should be preserved, what is its approximated end volume, what are the associated costs and how these are planned to be covered.

The above list is set as a guideline for any data generated within LETHE for the whole project duration. Obviously, the sharing of data will be strictly linked to the level of confidentiality of the data itself. In particular, the level of confidentiality of gathered data will be checked by the partners responsible in order to verify if data can be disclosed or not. For the purpose, a written confirmation to publish data in LETHE Open Access Repository will be asked via e-mail by the task leader to the data owner. It will be possible to make such data available only following the received confirmation provided by the data owner. No confidential data generated within the project will be made available in digital form.

4.4 Making Data Interoperable

4.4.1 Is the data produced in the project interoperable, that is allowing data exchange and re-use between researchers, institutions, organisations, countries, etc. (i.e. adhering to standards for formats, as much as possible

There should be a clear difference between prospective and retrospective data. Retrospective data should be processable (data structure). Prospective data should follow international standards (e.g. FHIR, Personal Connected Health Alliance (former Continua), CDA if necessary).
A clear decision can be made if data structure is available. Changes of standards may be possible later in the project.

4.4.2 What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?

It would be necessary that coding systems/standards provide unique identifiers for each system (e.g. OIDs, URLs,…)
Additionally, there is a clear commitment to use all mentioned Standards / vocabularies / methodologies by all project members as stated in 3.13.

4.4.3 Will you be using standard vocabularies for all data types present in your data set, to allow inter-disciplinary interoperability?

In LETHE project standard vocabularies will be used for all data types for all datasets. Only plain text data will be used which is already inter-operable. Data stored and/or generated through Project Databases will be stored in standard database format (SQL), and will feature export functionalities to word processing and other applications, making it inter-operable and accessible. Furthermore data in no-SQL format to be used is also interoperable since international standards will be used to manage this data types.

If necessary and useful: ICD-10, SNOMED-CT, LOINC, ATC/DDD (if provided), MEDDRA
A clear decision can be made if data structure is available. Changes of standards may be possible later in the project.

4.4.4 In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies?

In LETHE we have not planned to develop ontologies, but we will use existing ones in case they are needed. We prefer to use Reference Sets.

4.5 Increase Data Re-use (Through Clarifying Licences)

The data re-use policy of LETHE will strictly be defined by the consortium and through IPRs to be signed between partners. This process will take place later in time and as project progresses.

For software code, whenever deemed necessary, LETHE partners will publish the corresponding code on a GitHub repository using open licenses that allow market re-usage. In this regard, licenses like MIT, Apache, and BSD etc. will be favoured, however they will be subject to other code dependencies.

Similarly, whenever research data will be generated that will be used for a publication, the data will be pseudo-anonymized and then if possible, shared on a public repository like GitHub, to allow other researchers simulate the experiments and further extend the results.

4.5.1 How will the data be licensed to permit the widest re-use possible?

Data generated and used within the project will be made publicly available, in the cases that this is possible, following the FAIR data directive. When this is not possible, licensing options will be examined on a case by case basis, taking into account all applicable factors and IPR principles and agreements.

LETHE aims to support the use of open source tools and models. The R&D partners will openly share concepts, software programs, API interface descriptions, outputs from AI/ML models etc. The data will be made available on LETHE website once is ready as well as on EU OpenAIRE portal https://www.openaire.eu/.

4.5.2 When will the data be made available for re-use? If an embargo is sought to give time to publish or seek patents, specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

The data will be made available for re-use immediately after project completion.

4.5.3 How long is it intended that the data remains re-usable?

Data will remain re-usable at least until the end of the project. After project completion and based on IPR agreements between consortium partners, data can remain re-usable, an action that will be reported on the DMP version to be submitted on M42.

4.5.4 Are data quality assurance processes described?

All LETHE deliverables, audio-visual content and reports/other publications will be peer-reviewed by the project partners and in some cases by external reviewers as well. Through this approach we expect to ensure high data quality within the project, promoting project data re-use and sharing.

5 Project Datasets

The current version of the DMP addresses the following aspects on a dataset basis and presents the current status of reflection within the consortium concerning the set of data managed by the project.

- Dataset reference, name and reference: Identifier for the data set to be gathered / processed / generated.
- Dataset description: Description of the data that will be generated or collected, its origin (in case it is collected), nature and scale and to whom it could be useful and whether it underpins a scientific publication. Information on the existence (or not) of similar data and the possibilities for integration and reuse.
- Standards and metadata: it includes the reference to existing suitable standards of the discipline. If these do not exist, an outline of how and what metadata will be created.
- Data sharing: A detailed description of how data will be shared, including access procedures, outlines of technical mechanisms for dissemination and necessary software and other tools for enabling re-use, and definition of whether access will be wide open or restricted to specific groups. Identification of the repository where data will be stored, if already existing and identified, indicating, in particular, the type of repository (institutional, standard repository for the discipline, etc.). In case the dataset cannot be shared, the reasons for this should be mentioned (e.g., ethical, rules of personal data, intellectual property, and commercial, privacy-related, security-related).
- Archiving and preservation: Description of the procedures that will be put in place for the long-term preservation of the data. Indication of how long the data should be preserved, what is its approximated end volume, what the associated costs are and how these are planned to be covered.

All data collected in phase I and processed (retrospective data and homogenized common master dataset) and in phase II of the project from participants, IoT, wearables, TEMI robot etc. will be stored in the data infrastructure environment of EGI. Like already stated in the proposal of course we cannot provide all data as open source. This is justifiable basically based to IPRs we need to protect. Source code will be provided in a Github repository at the FHJ environment and Project documentation in sharepoint environment at the FHJ.

Currently, LETHE plans to manage 5 different datasets. Next sections describe those datasets following a structure in accordance with the Guide of Horizon 2020 for the Data Management Plan [1, 2].

5.1 Dataset 1: Information about the consortium

1	**Dataset reference**
	Information about the consortium
2	**Dataset Description**
	This dataset includes information about the consortium (e-mails, phones, project resources and etc)
3	**Standards and Metadata**
	Excels, SQLs and text files containing data about the consortium.
4	**Data Sharing**
	Data is only for the consortium
5	**Archiving and preservation (including storage and backup)**
	All respective data storage and backup takes place on FHJ secure servers infrastructure and making use of Microsoft Sharepoint platform.

5.2 Dataset 2: Project files

1	**Dataset reference**
	Project files
2	**Dataset Description**
	This dataset includes all meetings reports, research activities from their creation until their dissemination, the information and data collected in workshops and other communication activities.

3	**Standards and Metadata**
	Partners will follow the metadata and standards notation stated in section 3.14 in order to name all project files.
4	**Data Sharing**
	Confidential data will not be shared and if it has to, it will be protected accordingly (e.g. encrypted). The rest of the data can be publicly disseminated.
5	**Archiving and preservation (including storage and backup)**
	Sensitive data will be stored locally by each organization under their responsibility and strict protection policies. Non-sensitive data can be stored in all partners' IT systems and disseminated through LETHE communication channels

5.3 Dataset 3: Research activities

1	**Dataset reference**
	Research activities
2	**Dataset Description**
	Research activities related to data collection, usage and processing entail the following categories: ➢ Retrospective Data existing already in hospitals and medical centers and which is provided in LETHE for further analysis. Presented in Appendix 12.2 ➢ Common master dataset – after homogenization of original retrospective data, presented in chapter 5.3.1 ➢ Prospective data that will be collected from participants of the pilot trial, wearables, mobile application and cardiovascular sensing system presented in chapter 5.3.2 ➢ Data from TEMI robot presented in chapter 5.3.3 ➢ cCOG web test leading to the CAIDE score matrix – presented in Appendix 12.3
3	**Standards and Metadata**
	Partners will follow the metadata and standards notation stated in section 3.14 in order to name all data files, database tables, web questionnaires and files as outputs from machine learning processing of datasets.

4	Data Sharing
	For LETHE Data Sharing the 'Professional Cloud Environment' will be provided by the EGI Federated Cloud that will host an encrypted central storage. This storage will then offer secured access for data transfers and sharing between partners and between partners and third parties following agreements signed. Further the encrypted central storage within EGI Fedcloud has the advantage that could provide an environment to process the data in the cloud should partners wish to. In a more generic view non-sensitive data will be shared between the consortium and research outcomes will be made publicly available when necessary. If during the research activities some organization collects sensitive/personal data such data will only be shared if it is strictly mandatory and only after prior pseudo-anonymization of data and encryption.
5	**Archiving and preservation (including storage and backup)**
	Data will be stored and preserved in the secure IT environment of partner EGI and all partners – based on roles and accreditations – will have access through secure channels and using at least 2F authentication.

5.3.1 Common Master Retrospective Dataset

Retrospective data will be collected from medical partners and will in turn be further processed to create prediction modelling work and as a result new datasets. Retrospective health data related to: cognition, cognition-related markers, demographics, functional status, health status, lifestyle, mood, quality of life and study visits. Categorization entails:

- Theme / dataset name – is the name of the dataset
- Variable / group of variables are a level below theme/dataset name

Within LETHE there are four (4) sources for retrospective clinical data and data from established studies provided by the MUW, UPG, KI and THL. A key part of LETHE project is to provide an initial model suite for progression of dementia indicators as well as related risk factors based on these retrospective datasets, which can be achieved by producing a homogenized common master data record ("Master Dataset") out of these four (4) individual datasets. Information about this master dataset is provide on the table below:

Theme / dataset name	Variable/ group of variables	Data field description Alphanumeric , number etc
Clinical markers	Height	number
Clinical markers	Weight	number
Clinical markers	Glucose	number
Clinical markers	Albumine	number
Clinical m	Vitamin D	number
Clinical m	Hemoglobin	number
Clinical m	BMI	number
Clinical m	Cholesterol	number
Clinical m	Creatinine	number
Clinical m	Calcium	number
Clinical m	TSH	number
Clinical m	Vitamin B12	number
Clinical m	Folic acid	number
Clinical m	Triglycerides	number
Clinical m	Blood pressure	number
Cognition	Dementia (type)	number
Cognition	MCI	number
Cognition	SCI	number
Cognition	Clinical dementia rating (CDR)	number
Cognition	Digit Symbol	number
Cognition	Digit Span	number
Cognition	FAS test (verbal fluency by letter)	number
Cognition	RAVLT	number
Cognition	Subjective memory	number
Cognition	MRI: total gray matter volume	number
Cognition	MRI: regional volumes (e.g. hippocampal)	number
Cognition	TMT - A	number
Cognition	TMT - B	number
Cognition	Verbal fluency by category	number
Cognition	CSF	number
Cognition	MMSE	number
Cognition	ApoE	number

Demographics	Date of birth/ age	number
Demographics	living alone	number
Demographics	marriage status	number
Demographics	Education level	number
Demographics	Occupation	number
Demographics	Age, years	number
Demographics	Education years	number
Demographics	Sex	number
Functional status	ADL and IADL	number
Functional status	ADL	number
Functional status	SPPB or similar	number
Functional status	Care need	number
Health status	Current medications	number
Health status	Hypothyreosis, hyperthyreosis or other thyroid disease	number
Health status	Obstructive sleep apnea syndrome	number
Health status	Malignancy/cancer	number
Health status	Hyperlipidemia	number
Health status	Heart bypass	number
Health status	TIA	number
Health status	Accidents (treated by a doctor)	number
Health status	Dementia	number
Health status	Parkinson's disease	number
Health status	Myocardial infarction	number
Health status	Diabetes type 1	number
Health status	Diabetes type 2	number
Health status	Depression	number
Health status	Hypertension	number
Health status	Head injury	number
Health status	Hospitalization (within last year)	number
Health status	Stroke	number
Lifestyle	Sleep and related	number
Lifestyle	Alcohol use	number
Lifestyle	Smoking Ever	number
Lifestyle	Smoking Current	number
Mood and QOL	Depression/GDS	number

5.3.2 Prospective Wearable Data sources

Here below there is a potential list of wearable data sources, but it will be modified during the project evolution (decisions regarding choice of wearables has not yet been done) based on research and discussions currently held between consortium partners. On M24 when the updated version of DMP will be presented these prospective wearable related datasets will be explicitly specified.

***Sensing* Technology (Active and Passive)**	**Short description and background**	**TRL[23]**	**Prospective Data collected according to followed Lifestyle[24]**	**Description**
Smartphone sensory system (IMU, GPS)	**Passive** collection of behaviour data, **Active** collection of input data in **daily interactive dialogues** regarding mental health (mood etc.), smoking cessation, social activities, subjective parameters runs LETHE App	7-8 (out of the market solution)	Daily tracking of route s, smoking cessation, mood, location, persons meet, social activities participated etc.	LETHE will exploit location technologies and specifically the smartphone sensory system (GPS and IMU) to address the problem of elder's wandering and help to keep them safe and secure. LETHE will incorporate in its mobile app a route tracking service which will help the target group in their outdoor daily activities. The app compares locations and routes against a preset geofence or virtual boundary, i.e. safe zones and will trigger timely interventions either though it or by engaging the caregivers when necessary.

Fitbit Charge 4 (wearable)	**Passive** collection of activity data, sleep related data and health data	7-8 (out of the market solution)	Daily Steps, Daily Activity, Daily Stairs, Sleep Quality, Sleep duration, Sleep onset, Sleep stages, heart rate, SpO2	During the project LETHE will adopt ready market wearable solutions to collect activity tracking data and metadata. LETHE will integrate this wearable using the RADAR-Base framework.
Kardiaobile FDA-cleared **single-lead** or/and six-lead ECG/EKG	**Passive** collection of cardiovascular related data from two lead ECG And **Active** collection of blood pressure collected through the mobile LETHE app	7-8 (out of the market solution)	Daily ECG patterns recordings, Daily heart rate variability (HRV), Daily blood pressure measurements	LETHE will exploit a portable, easy to use and non-obtrusive medical device certified single-lead or/and six-lead ECG device. During the project lifetime an ML risk stratification algorithm will be developed based on digital biomarkers along with the clinical status of the patient collected to his/her personal health record. The algorithm could detect electrocardiogram anomalies exploiting deep learning algorithms, convolutional networks or high performance supervised classifiers assessing the cardiovascular risk in different settings.

5.3.3 TEMI Robot

LETHE will exploit the technology of companion robots as one of its proposed interventional pillars. Due to functionalities such as automatic map generation the setup time of a robot in unknown environments is kept to a minimum. TEMI is a commercially available robot that has been used in different real settings by the project partner KAASA. TEMI has several capabilities and functionalities that can be used in LETHE like:

- video calls,

- smarthome services,
- remind patients – during pilot runs - of taking the Combinostics test.
- Transfer data collected to a server for further processing.
- automatic map generation,
- connection with physiological sensors that have been integrated by KAASA already and will be used in LETHE.

The plan is to use TEMI based on the prediction and intervention model. If we look at TEMIS from the tablet / smartphone perspective, we add another degree of freedom to the new model that LETHE project brings.

5.4 Dataset 4: Development data, implementations and codes

1	**Dataset reference**
	Development data, implementations and codes
2	**Dataset Description**
	This dataset includes all the implementations, codes and development outcomes regarding to the architecture of LETHE. A key part of the work is the respective prospective type dataset that will result after applying AI/ML algorithms on the datasets stored and processed in the EGI Federated Cloud - encrypted central storage. In general, the prediction model will deliver as the main result a measure for dementia progression and a corresponding measure for the uncertainty of the prediction. In addition, a log-file will be generated during training of the individual models. This log-file includes model specific information (e.g. network architecture for a neural network, hyperparameters) and training history (e.g. training, validation and test metrics, learning curves). Furthermore, the final (trained) models will be stored in a model file e.g. for a deep neural network this would result in a table of the learned weights for each layer and the network architecture. The models which will be selected highly depend on the research question to be answered.
3	**Standards and Metadata**
	Documents, s/w programs, architectures, s/w code as per framework presented in chapter 3.13
4	**Data Sharing**
	Data will be shared in GitLab between required consortium members.

5	Archiving and preservation (including storage and backup)
	The code of the project will be stored in GitLab and periodic backups, as well as associated documentation, will be stored on EGI's secure IT infrastructure.

5.5 Dataset 5: Pilot and testing activities

1	Dataset reference
	Pilot and testing activities
2	Dataset Description
	This dataset stores data related to the pilot and testing activities of the project. It includes pilot setup and execution deliverables, pilot results evaluation reports, risk factor datasets and other user and participants information.
3	Standards and Metadata
	Partners will follow the metadata and standards notation stated in section 3.14 in order to name all project files.
4	Data Sharing
	Pilot and testing outcomes will be shared between members if they do not disclose sensitive data. If organizations use real data, all experiments will be performed in their premises and the outcomes of such tests will be shared prior sanitization of data.
5	Archiving and preservation (including storage and backup)
	Data about real case scenarios and participant users during pilot trials will be stored in secure EGI servers infrastructure and will be managed by respective partners according to strict protection measures, such as an isolated storage/computer resource as well as encryption mechanisms and of course authentication and strict access rights as WP2 architecture design is specifying.

6 Data Security

Following discussions with WP2 partners and discussions about LETHE architecture it has been agreed that the Website, the Application and Data Base will be hosted in a Cloud Computing and Data Storage 'as a service' Provider. The following paragraphs describe organizational and technical measures to be applied by the Data controller(s), and additionally the requirements for the data hosting and data security issues, that the chosen provider should cover as part of the data management plan and the key topic of security.

6.1 Organizational measures

The organizational measures referred are not exhaustive, however essential. Data controller(s), assisted by their data protection focal points and other relevant staff, are encouraged to:

1. Ensure that relevant data security measures are covered in Standard Operating Procedures, e.g. procedures for physical and electronic file management;
2. Ensure that trainings in data protection are organized or attended, including for Implementing partners
3. Raise the awareness for the responsible use of Lethe's ICT assets and resources including email, internet, portable devices and ICT equipment;
4. Ensure the conduct of Data Protection Impact Assessments;
5. Implement methods of safe transfer for personal data of Data Subjects;
6. Routinely review and upgrade data security measures, e.g. through random monitoring and inspections and testing, assessing and evaluating the effectiveness of existing measures;
7. Share relevant SOPs with and keep the DPO informed of organizations measures.

Furthermore, in order to prevent unauthorized access, modification, replication or destruction of LETHE data in general, several measures are proposed and must be put in place by all partners - whether being data controllers, data processors, developers, integrators, pilots etc. These include:

- **Identification security**: Data is stored in online repositories which are password protected and/or grant access only upon correct identification. Different layers of security are implemented in order to protect data of higher sensitivity (users' personal data, etc.)
- **Location security**: Access to the premises of the partners, where LETHE physical files/confidential information is being stored, is restricted.
- **Workstation security**: People working on LETHE are strongly encouraged to remain protected against a possible data breach by password protecting all computers and through the use of an up to date antivirus software. Additionally, the sharing of confidential information via email is highly discouraged.

6.2 Technical measures

Under technical measures, should be the maintenance of physical security of premises, portable equipment, individual case files and records and ICT security through a number of control measures. This section elaborates on physical and electronic file management and distinguishes storage, access and user control that apply to both forms of file management. Data controller(s) may delegate the implementation of technical measures to their data protection focal points together with, for instance, registration and IT staff.

6.2.1 Physical file management

Storage control responsible personnel is advised to observe and the following:

1. Hard Copy files should be kept in a lockable storage room or location designated for this purpose within Lethe's premises, safe from water, fire and temperature damage;
2. Access to the storage room to be controlled, monitored or restricted, for example, through access cards, physical control barriers, local or remote monitoring systems, with only authorized personnel granted access to enter;
3. The storage location needs to be kept locked when unattended. Copies of the key(s)/access cards are normally kept only by the Filing/Registration staff and the Representative and/or senior protection staff;
4. Outside the storage room, files should be kept in a locked cabinet or drawer when personnel dealing with it is not at his/her desk or out of office, even for short breaks;
5. Access to Lethe's premises should be regulated, visitors logged in and out, and accompanied by Lethe personnel inside the premises and offices.

Access control to physical files (within and outside the designated storage location):

1. Lethe's project members should have access to physical files that have been assigned to them, in line with their duties and responsibilities;

User control. Tracking and recording the movement of physical files:

1. A file check-out/check-in procedure should be in place, with an up-to-date record of who has, and have in the past had, access to individual POC/data subject files;
2. The information that should be registered is the file number, date, and initials/name of the personnel requesting the file onto the file movement log upon release, and note its date of return and initials/name of the personnel who returned it;
3. Requests, releases, transfers and returns of files should normally be recorded on a File Action Sheet. File movement logs should be sought stored electronically wherever possible. Larger operations may also consider implementing an electronic

tracking system by attaching barcodes to their files, and issuing identification with barcodes to personnel;

4. Lethe personnel may not remove individual POC/data subject files from Lethe's premises. Exceptions may be authorized by the data controller or local DPO's based on a written request. There should be a limit to the number of files an individual project member may have in his/her possession at any given time.

-

6.2.2 Electronic file management

Personal data stored in electronic format are particularly vulnerable to accidental, unlawful or illegitimate destruction, loss, alteration, as well as unauthorized disclosure, due to the ease with which it can be copied, transferred, and even posted on the Internet. Access to such data should therefore be carefully restricted, managed and monitored. Data controllers, with close support from IT Officers, are responsible for ensuring that databases and supporting IT infrastructure are established and used according to standard, including the following measures:

6.2.2.1 Storage control

1. Operations are advised to only use Lethe's organisation's tools, document management applications, and network drives with controlled accessibility. The use of non-Lethe approved tools can undermine data security;
2. Server locations need to be physically secure, with adequate electrical, water and fire safety. IT Officers are responsible for adequate back-up procedures;
3. Offices with reliable access to the internet are advised to store electronic files in Cloud Drives; offices without such access should establish a restricted shared drive. Personal data of POCs/data subjects should not be stored on personal or network drives.

6.2.2.2 Access control to electronic files

1. Access to electronic files should be tiered, so that personnel only have access to what they need to for the purposes of performing their duties and responsibilities;
2. Operations are recommended to establish procedures for the submission and review of user access requests to ensure that users are only given access to the data they need. Access rights are normally defined by the Heads of Units, approved by the data controller(s), and updated by a database administrator;
3. A regular review of access rights is recommended, e.g. every 6 months, to ensure that personnel who no longer require access have their permissions revoked.

6.3 Data security by Cloud Computing and Data Storage Provider: Data recovery, secure storage and transfer of sensitive data

6.3.1 Data Hosting and Cloud Computing and Data Storage Services Used

A Cloud Computing and Data Storage will be used as a service provider. As such, EGI as the hosting provider will provide VMs which host the website, the application and the database. In more details, the Cloud Computing and Data Storage provider should provide the following products/services for Lethe:

- **VMs**. Usually Linux (or other)-based virtual machines (VMs) that run on top of virtualized (or not) hardware. Each VM should be created as a new server that can be used, either standalone or as part of a larger, cloud-based infrastructure.
- **Managed Databases**. Managed Databases should be fully managed, by high performance databases preferably in cluster service. Using managed databases is a powerful alternative to installing, configuring, maintaining, and securing databases manually. Clusters include daily backups with point-in-time recovery (PITR), standby nodes for high availability, and end-to-end SSL encryption. Managed databases are multi-region and scalable, and their automated failover means even single-node plans add resiliency to your infrastructure. When a new managed database cluster is created, the cluster should be added to a private network or VPC network for the datacenter region.
- **Cloud Firewalls**. Cloud Firewalls are a network-based, stateful firewall service provided for the VM (Environment) access. Cloud firewalls block all traffic that isn't expressly permitted by rules. Firewalls actually place a barrier between servers and other machines on the network to protect them from external attacks. Cloud Computing and Data Storage's Firewalls, are Cloud Firewalls, that is network-based and stop traffic at the network layer before it reaches the server.

6.3.2 Data backup and recovery

The provider should provide backup services also.

- **Backups**. Backups are automatically-created disk images of VMs. Enabling backups for VMs enables system-level backups at weekly intervals, which provides a way to revert to an older state or create new VMs.

In brief, The Provider should utilize a snapshot-based backup system that creates a point-in-time image based on the current state of a VM. This process happens automatically within a pre-determined scheduling window and is completed in the background while the VM is running. This should provide system-level backups of the server without powering down as follows:

– A snapshot of the live system is taken, creating a crash-consistent, point-in-time image.
– The snapshot is backed up off-disk.
– The snapshot is deleted once the backup is complete.

– A crash-consistent backup allows the system to capture all of the data on disk exactly as it was at a single point in time. This means that the data is backed up in a consistent state.

This is called a crash-consistent backup because it saves every piece of data that was committed to the disk at the moment that the snapshot occurs. The data saved is consistent with the data that would be available if the system crashed at that exact point and had to recover on boot. Backups should be taken once per week, minimum, and each backup should be retained for 4-6 weeks. Backups should be stored in the same datacenter as the corresponding VM.

- **Snapshot and Backup Security**. Snapshots and Backups should be stored on internal non-publicly visible network on NAS/SAN servers. External customers can directly manage the regions where their snapshots and backups exist which allows the customer to control where their data resides within the datacenter for security and compliance purposes.

6.3.3 Secure transfer of sensitive data

There is a high risk of data breaches when personal data is communicated or transferred, for instance from a data contributor to the repository. E-mails and SMS messages may be intercepted during transmission and/or retained by surveillance programmes, thus putting data subjects at risk of exposure. On this issue, in order to ensure and respect confidentiality, personal data must be transferred only through the use of protected means of communication.

In order to reduce the risk of personal data breaches during communication and transfer of personal data, LETHE personnel isrecommended to:

1. In principle, use only Lethe's developed and approved tools to transfer personal data;
2. Exercise caution regarding the use of third party file-sharing tools;
3. It is impossible to guarantee the confidentiality of any electronic message transmitted outside the Lethe system via the internet. No information of a confidential nature should be sent by e-mail via the internet. More secure alternatives include the use of a secure file transfer protocol (SFTP) service, and encrypted portable media devices;
4. Personal data should not be transferred using personal email accounts (e.g. Gmail, Yahoo or Hotmail), or through social media accounts (e.g. Facebook, Twitter)
5. If e-mail is used, ensure that additional measures are taken to protect the content, such as encrypting the email or its attachment. When sharing password protected files, the password should be sent via an alternative means of communication (such as phone call or text message);
6. SMS should be avoided as a means to communicate personal data.

7. Seek advice from the IT Officer, DPO, on which tools to use for different purposes and in different operational scenarios.

6.4 Data security by Cloud Computing and Data Storage Provider: Safe storage in certified repositories for long term preservation and curation

The provider should also care about overall security of its services. Physical Security measures and Certifications are factors that contribute to this objective.

6.4.1 Security Certifications

Cloud Computing and Data Storage Provider should be certified for Security (ISO/IEC 27001:2013). The certification for information security should be publicly available. The certificate should state that the Cloud Computing and Data Storage Provider, located in Netherlands is compliant with the requirements as stated in the standard: ISO/IEC 27001:2013. Additional and complementary compliance statements are required such as GDPR.

6.4.2 Data Centers Physical Location

Furthermore, the VM, DataBase and Datasets should be hosted in a Datacenter physically located within EU. The specific datacenter should also be certified with the above mentioned certifications and compliance statements.

6.4.3 Physical Security

- **Physical Security**. Datacenters usually are co-located in well-respected datacenter facility providers in the world. The provider should leverage the capabilities of such providers including physical security and environmental controls to secure infrastructure from physical threat or impact. Therefore, the site should be staffed 24/7/365 with on-site physical security to protect against unauthorized entry. Security controls provided by datacenter facilities should include but not limited to:
    - 24/7 Physical security guard services
    - Physical entry restrictions to the property and the facility
    - Physical entry restrictions to our co-located datacenter within the facility
    - Full CCTV coverage externally and internally for the facility
    - Biometric readers with two-factor authentication
    - Facilities are unmarked as to not draw attention from the outside
    - Battery and generator backup
    - Generator fuel carrier redundancy
    - Secure loading zones for delivery of equipment

6.4.4 Infrastructure Security, Logs and System Monitoring

Systems should be protected through key-based authentication and access should be limited by Role-Based Access Control (RBAC). RBAC policy should ensure that only the users who require access to a system are able to login. In any case, the provider should consider any system which houses customer data (Lethe Datasets) that to be of the highest sensitivity. As such, access to these systems should be extremely limited and closely monitored.

- **Infrastructure Security**. The provider's infrastructure should be secured through a defense-in-depth layered approach. Access to the management network infrastructure should be provided through multi-factor authentication points which restrict network-level access to infrastructure based on job function utilizing the principle of least privilege. All access to the ingress points should be closely monitored, and are subject to stringent change control mechanisms.

Additionally, hard drives and infrastructure should be securely erased before being decommissioned or reused to ensure that data remain secure.

- **Access Logging**. Systems controlling the management network at the Provider should log to a centralized logging environment to allow for performance and security monitoring. The logging should include system actions as well as the logins and commands issued by the provider's system administrators.
- **Security Monitoring**. The Provider's Security team should utilize monitoring and analytics capabilities to identify potentially malicious activity within infrastructure. User and system behaviors should be monitored for suspicious activity, and investigations should be performed following any incident reporting and response procedures.
- **VMs Security & Provider's Employee Access**. The security and data integrity of the project's (Lethe) VMs should also be under control. Provider's technical support staff should not have access to the backend hypervisors where virtual servers reside nor direct access to the NAS/SAN storage systems where snapshots and backup images reside. Only selected engineering teams should have direct access to the backend hypervisors based on their role.

7 Scientific publications

As reported in the DoA, a dissemination and communication plan has been set up in order to raise awareness on the project outcomes among specialized audience. In this framework, the consortium commits itself to perform publications in peer reviewed international journals in order to make the outcomes available to the scientific community. Fully in line with the rules laid down in LETHE Grant Agreement and respecting personal data, confidentiality and IPR rights wherever applicable, each beneficiary will ensure open access to all peer reviewed scientific publications relating to its results.

The project will make use of a mix of the three different possibilities for open access, namely:

1) **Open access publishing** (without author processing charges): partners may opt for publishing directly in open access journals, i.e. journals which provide open access immediately, by default without any charges.
2) **Gold open access publishing:** partners may also decide to publish in journals that sell subscriptions, offering the possibility of making individual articles open accessible (hybrid journals). In such case, authors will pay the fee to publish the material for open access, whereby highest level journals offer this option.
3) **Self-archiving/ "green" open access publishing:** alternatively, beneficiaries may deposit the final peer reviewed article or manuscript in an online disciplinary, institutional or public repository of their choice, ensuring open access to the publication within a maximum of six months.

Moreover, the relevant beneficiary will deposit at the same time the research data presented in the deposited scientific publication into a data repository. The Consortium will evaluate which of these data will be part of the data to be published on LETHE Open Research Data Platform mainly according to Ethics and confidentiality reasons.

7.1 Bibliographic metadata

Metadata for scientific peer reviewed publications must be provided in order to maximize the discoverability of publications and to ensure EU funding acknowledgment.

The inclusion of information relating to EU funding as part of the bibliographic metadata is necessary also for adequate monitoring, production of statistics and assessment of the impact of Horizon 2020.

All the following information must be included in the metadata associated to each LETHE publication following the metadata framework process presented in chapter 3.13.

Information about the grant number, name and acronym of the action:

- European Union (UE);
- Horizon 2020 (H2020);
- Innovation Action (IA);
- LETHE;
- Project No 101017405.

Information about the publication date and embargo period if applicable:

- Publication date
- (eventual) Length of embargo period

Information about the persistent identifier:

- Persistent identifier, if any, provided by the publisher (for example an ISSN number).

8 Data Governance

8.1 Data Governance Model

This paragraph examines and synthetizes current discourses and practices on the governance of data. It scrutinizes different approaches for accessing, controlling, sharing and using data in today's platform economy and depicts four emerging models of data governance (Micheli, et al., 2020). Finally proposes the governance model that fits the 'Lethe' requirements and goals set.

As Micheli observes, (Micheli, et al., 2020), the current platform economy is mainly characterized by the asymmetry of power of a few technology corporations and telecommunication companies that have established de-facto quasi-data monopolies. The negative societal implications of this system, including biases in algorithmic decision-making, nudging and manipulation, and privacy violations are increasingly highlighted by research (Beer, 2017); (Taylor, 2017). Additionally scandals such as Cambridge Analytica raise the awareness among public opinion and policy makers, at least in Europe, that the distortions of this model need to be addressed. The General Data Protection Regulation (GDPR) is an important step in this direction, even if with some limitations (Delacroix & Lawrence, 2019), and further new measures are being prepared in the European Union, including a Digital Services Act and a Data Act (European Commission (EC), 2018).

Based on Micheli et al., proposition (Micheli, et al., 2020), in order to classify the data governance models, the following questions should be addressed:

(1) What configurations of roles and relationships between stakeholders can we identify in the emerging models of data governance?
(2) To what extent are other actors beyond corporate data platforms able to participate?
(3) What kind of value is pursued and how is it redistributed across actors and society?
(4) What mechanisms and arrangements are set in place to generate value from the data?

Although the dominant model of data governance in current 'platform society' is the one established by a few corporate big tech platforms, other actors beyond 'big tech' are progressively becoming involved in controlling personal data and producing value from it through different data governance models. By focusing on these alternative models and looking instead at the practices for data access and control developed by societal actors, it is understood that these practices are a fertile context for developing socio-technical imaginaries for data that might influence how (big) data will be governed in the future.

This section describes the data governance models identified following the five dimensions described in the following table:

Dimension	Definition
Stakeholders	The individuals, institutions, organisations or groups who are affected by, or have an effect on, the way data is governed and the value created.
Governance goals	The objectives held by actors/parties who influence how data is governed.
Value from the data	The resources expected to be generated from the use of data and how these are distributed among actors and across society.
Governance mechanisms	The different instruments adopted to achieve specific governance goals, including the underlying principles.
Reciprocity	The power relation between stakeholders for data access and use.

These models should be understood as ideal types. They are analytical constructs that emphasize certain traits in order to synthetize phenomena that differ for the degree of affiliation to those traits. They are not intended as an exhaustive description of the state of the art, but as a contribution in synthetizing emerging data governance models. The analysis includes models that differ, to varying degrees, from the current dominant one. Therefore, we do not account for cases in which platforms engage in data sharing with other actors, but retain full control over data, deciding unilaterally which other stakeholders to bring inside, what data they can access and what they can do with it. The four models described are labelled:

- data sharing pools (DSPs),
- data cooperatives (DCs),
- public data trusts (PDTs)
- and personal data sovereignty (PDS).

9 These models are depicted in a summary table in paragraph 13, ALLOCATION OF RESOURCES

9.1 What are the costs for making data FAIR in your project?

Costs for making data FAIR relate to papers in international journals, exhibitions, conferences and any other promotional activities including manpower to publish results and make data available in OPENAIRE website.

9.2 How will these be covered? Note that costs related to open access to research data are eligible as part of the Horizon 2020 grant (if compliant with the Grant Agreement conditions).

All costs associated with making data FAIR have been taken into account and will be covered by partners' individual budgets under the category other costs.

9.3 Who will be responsible for data management in your project?

Infotrend is the LETHE DMP responsible partner and the so called project DPO office. Having said that though all partners collecting, sharing and processing data are also responsible for their respective share of work and of course the project coordinator has a key role in project data management especially for data related to deliverables, minutes of meeting and presentations.

9.4 Are the resources for long term preservation discussed (costs and potential value, who decides and how what data will be kept and for how long)?

Although a key issue, long term preservation has not been dealt so far within the consortium as the project is at an early phase in development and progress. This issue will be discussed towards the end of the project and will be reported in the DMP version due M48.

10 Conclusions

This deliverable represents the LETHE Data Management Plan at month 6. The scope of this Data Management Plan is to describe the data management life cycle for the data to be collected, processed and/or created in the framework of the LETHE project.

In particular, this document specifies how LETHE research data will be handled in the framework of the project as well as after its completion. More in detail, the report indicated:

- what data will be collected, processed and/or created and from whom
- which data will be shared and which one will be maintained confidential
- how and where the data will be stored during the project
- which backup strategy will be applied for safely maintaining the data
- how the data will be preserved after the end of the project
- security issues related to data collected/processed and handled.
- Data governance issues

Moreover, the deliverable presents a preliminary strategy for the proper management of some data generated in the framework of LETHE project activities that incidentally can come from the participation of humans and related sensible data.

Last, but not the latest a summary of the project procedures to be followed in case of activities involving personal data have been summarized in the deliverable.

The present Data Management Plan has to be considered as a living document and any future update or change in LETHE data management policy will be included in the periodic reports or will be specified in the deliverables related to the specific tasks as well as in 2 new versions of DMP to be submitted M24 and M48.

11 References

[1] EU DMP Online Manual: https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm

[2] Guidelines on FAIR Data Management in Horizon 2020: https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

[3] OpenAIRE - EC Open Research Data Pilot: https://www.openaire.eu/what-is-the-open-research-data-pilot

[4] Gitlab: https://github.com/

[5] Github: https://about.gitlab.com/

[6] E.g. Ref. Faq CNIL dated August 17th, 2018 related to mobile apps use for healthcare purposes (Ref. https://www.cnil.fr/fr/applications-mobiles-en-sante-et-protection-des-donnees-personnelles-les-questions-se-poser)

[7] Beer, D., 2017. The social power of algorithms. *Information, Communication & Society,* 20(1), pp. 1-13.

[8] Delacroix, S. & Lawrence, N., 2019. Bottom-up data trusts: Disturbing the 'one size fits all' approach to data governance. *International Data Privacy Law,* 9(4), pp. 236-252.

[9] European Commission (EC), 2018. *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions "Artificial Intelligence for Europe" COM/2018/237 Final..* Luxembourg: Publications Office.

[10] International Organization for Standardization, 2013. *ISO/IEC 27001:2013 | Information technology — Security techniques — Information security management systems — Requirements.* [Online]
Available at: https://www.iso.org/standard/54534.html
[Accessed 01 04 2021].

[11] Micheli, M., Ponti, M., Craglia, M. & Suman, A. B., 2020. Emerging models of data governance. *Big Data and Society,* Issue Jul-Dec, pp. 1-15.

[12] Shkabatur, J., 2019. The global commons of data. *Stanford Technology Law Review,* Issue 22, pp. 1-46.

[13] Taylor, L., 2017. What is data justice? The case for connecting digital rights and freedoms globally.. *Big Data & Society,* 4(2), pp. 1-14.

[14] The European Parliament and of the Council of the EU, 2016. *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Da.* [Online]
Available at: https://eur-lex.europa.eu/eli/reg/2016/679/oj
[Accessed 04 2021].

[15] Guidelines on FAIR Data Management in Horizon 2020: https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

[16] ISO Organization, 2019. *ISO 8601 - Date and Time Format.* [Online]
Available at: https://www.iso.org/iso-8601-date-and-time-format.html
[Accessed 01 06 2021].

[17] The Metadata Working Group, 2015. *DataCite Metadata Schema for the Publication and Citation of Research Data. (Version 3.1, August 2015).* [Online]
Available at: http://schema.datacite.org/meta/kernel-3/doc/DataCite-MetadataKernel_v3.1.pdf.
[Accessed 01 06 2021].

[18] Canham, S. & Ohmann, C., 2016. A Metadata Schema for Data Objects in Clinical Research. *Trials,* Issue 17.

12 APPENDICES

Appendix: Data governance models, page 70. **The data governance model that is closer to 'LETHE' requirements is the DSP, which is analyzed further in the next paragraph.**

12.1 Data sharing pools

Different actors join a DSP to 'analyse each other's data, and help fill knowledge gaps while minimizing duplicative efforts' (Shkabatur, 2019). By creating these partnerships, they ease the economic need for exclusive rights and obtain limited co-ownership stakes in the resulting data pool. Data is treated and exchanged as a commodity with the aim of producing data-driven innovation, new services, and economic benefits for all the parties involved or in the Lethe case to facilitate scientific research. DSPs are described as horizontal joint initiatives among data holders to aggregate data from different sources to create more value through their combination (Shkabatur, 2019). Their overall rationality is attuned with dominant discursive regimes of Big Data and lies in the assumption that 'the greatest advantages of data sharing may be in the combination of data from multiple sources, compared or 'mashed up' in innovative ways'.

Governance mechanisms for DSPs include technical architectures, such as data sharing platforms and Application Programming Interfaces (APIs), which facilitate a centralised data exchange within business ecosystems. However, a key mechanism is the contract, a legal and policy framework, that defines the modalities for data sharing, how data can be handled, and for which purposes. These contracts could be 'repeatable frameworks of terms and mechanisms to facilitate the sharing of data' between entities, which are especially useful for organisations that do not have the knowhow and legal support to leverage data. Although these frameworks have been referred to as data trusts, there is not a full consensus whether they could be assimilated to actual legal trust structures or a 'marketing tool' facilitating the responsible sharing of data (Delacroix & Lawrence, 2019).

In DSPs, one of the classic rhetoric of Big Data is embraced: data creates more value if aggregated. In that spirit, two or more data holders (both private and public) join forces and establish data sharing agreements. They analyse each other's data filling knowledge gaps and fostering data-driven innovation. On the surface this model promotes reciprocity between, potentially many, data holders, as it is based on horizontal relationships. Yet, it also fosters power asymmetries. Data holders with more resources or that possess more valuable datasets have greater power to set the terms on how data is accessed and used. Furthermore, data subjects (and citizens in general) do not have a voice in this model; they are not included in the relation and are at best depicted as recipients of the innovations developed through it.

13 ALLOCATION OF RESOURCES

13.1 What are the costs for making data FAIR in your project?

Costs for making data FAIR relate to papers in international journals, exhibitions, conferences and any other promotional activities including manpower to publish results and make data available in OPENAIRE website.

13.2 How will these be covered? Note that costs related to open access to research data are eligible as part of the Horizon 2020 grant (if compliant with the Grant Agreement conditions).

All costs associated with making data FAIR have been taken into account and will be covered by partners' individual budgets under the category other costs.

13.3 Who will be responsible for data management in your project?

Infotrend is the LETHE DMP responsible partner and the so called project DPO office. Having said that though all partners collecting, sharing and processing data are also responsible for their respective share of work and of course the project coordinator has a key role in project data management especially for data related to deliverables, minutes of meeting and presentations.

13.4 Are the resources for long term preservation discussed (costs and potential value, who decides and how what data will be kept and for how long)?

Although a key issue, long term preservation has not been dealt so far within the consortium as the project is at an early phase in development and progress. This issue will be discussed towards the end of the project and will be reported in the DMP version due M48.

14 Conclusions

This deliverable represents the LETHE Data Management Plan at month 6. The scope of this Data Management Plan is to describe the data management life cycle for the data to be collected, processed and/or created in the framework of the LETHE project.

In particular, this document specifies how LETHE research data will be handled in the framework of the project as well as after its completion. More in detail, the report indicated:

- what data will be collected, processed and/or created and from whom
- which data will be shared and which one will be maintained confidential
- how and where the data will be stored during the project
- which backup strategy will be applied for safely maintaining the data
- how the data will be preserved after the end of the project
- security issues related to data collected/processed and handled.
- Data governance issues

Moreover, the deliverable presents a preliminary strategy for the proper management of some data generated in the framework of LETHE project activities that incidentally can come from the participation of humans and related sensible data.

Last, but not the latest a summary of the project procedures to be followed in case of activities involving personal data have been summarized in the deliverable.

The present Data Management Plan has to be considered as a living document and any future update or change in LETHE data management policy will be included in the periodic reports or will be specified in the deliverables related to the specific tasks as well as in 2 new versions of DMP to be submitted M24 and M48.

15 References

[1] EU DMP Online Manual: https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm

[2] Guidelines on FAIR Data Management in Horizon 2020: https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

[3] OpenAIRE - EC Open Research Data Pilot: https://www.openaire.eu/what-is-the-open-research-data-pilot

[4] Gitlab: https://github.com/

[5] Github: https://about.gitlab.com/

[6] E.g. Ref. Faq CNIL dated August 17th, 2018 related to mobile apps use for healthcare purposes (Ref. https://www.cnil.fr/fr/applications-mobiles-en-sante-et-protection-des-donnees-personnelles-les-questions-se-poser)

[7] Beer, D., 2017. The social power of algorithms. *Information, Communication & Society,* 20(1), pp. 1-13.

[8] Delacroix, S. & Lawrence, N., 2019. Bottom-up data trusts: Disturbing the 'one size fits all' approach to data governance. *International Data Privacy Law,* 9(4), pp. 236-252.

[9] European Commission (EC), 2018. *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions "Artificial Intelligence for Europe" COM/2018/237 Final..* Luxembourg: Publications Office.

[10] International Organization for Standardization, 2013. *ISO/IEC 27001:2013 | Information technology — Security techniques — Information security management systems — Requirements.* [Online]
Available at: https://www.iso.org/standard/54534.html
[Accessed 01 04 2021].

[11] Micheli, M., Ponti, M., Craglia, M. & Suman, A. B., 2020. Emerging models of data governance. *Big Data and Society,* Issue Jul-Dec, pp. 1-15.

[12] Shkabatur, J., 2019. The global commons of data. *Stanford Technology Law Review,* Issue 22, pp. 1-46.

[13] Taylor, L., 2017. What is data justice? The case for connecting digital rights and freedoms globally.. *Big Data & Society,* 4(2), pp. 1-14.

[14] The European Parliament and of the Council of the EU, 2016. *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Da.* [Online]
Available at: https://eur-lex.europa.eu/eli/reg/2016/679/oj
[Accessed 04 2021].

[15] Guidelines on FAIR Data Management in Horizon 2020: https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

[16] ISO Organization, 2019. *ISO 8601 - Date and Time Format.* [Online]
Available at: https://www.iso.org/iso-8601-date-and-time-format.html
[Accessed 01 06 2021].

[17] The Metadata Working Group, 2015. *DataCite Metadata Schema for the Publication and Citation of Research Data. (Version 3.1, August 2015).* [Online]
Available at: http://schema.datacite.org/meta/kernel-3/doc/DataCite-MetadataKernel_v3.1.pdf.
[Accessed 01 06 2021].

[18] Canham, S. & Ohmann, C., 2016. A Metadata Schema for Data Objects in Clinical Research. *Trials,* Issue 17.

16 APPENDICES

16.1 Appendix: Data governance models

Model	Key actors	Goals	Value	Mechanisms
Data sharing pools (DSPs)	▪ Business entities ▪ Public bodies	▪ Fill knowledge gaps through data sharing ▪ Innovate and develop new services	▪ Private profit ▪ Economic growth	▪ Principle of 'data as a commodity' ▪ Partnerships ▪ Contracts (e.g. repeatable frameworks)
Data cooperatives (DCs)	▪ Civic organisations ▪ Data subjects	▪ Rebalance power unbalances of the current data economy ▪ Address societal challenges ▪ Foster social justice and fairer conditions for value production	▪ Public interest ▪ Scientific research ▪ Empowered data ▪ Subjects	▪ Principles from the cooperative movement ▪ Data commons ▪ 'Bottom-up' data trusts ▪ GDPR Right to data portability
Public data trusts (PDTs)	▪ Public bodies ▪ Inform policy-making	▪ Address societal challenges ▪ Innovate ▪ Adopt a responsible approach to data	▪ Public interest ▪ More efficient public service delivery	▪ Principle of 'data as a public infrastructure' ▪ Trust building initiatives ▪ Trusted intermediaries ▪ Enabling legal framework
Personal data sovereignty (PDS)	▪ Business entities ▪ Data subjects	▪ Data subjects self-determination ▪ Rebalance power unbalances of the current data economy ▪ Develop new digital services centred on users need	▪ Empowered data subjects ▪ Economic growth ▪ Private profit ▪ Knowledge	▪ Principle of 'technological sovereignty' ▪ Communities and movements ▪ Intermediary digital services (personal data spaces) ▪ GDPR Right to data portability

16.2 APPENDIX: RETROSPECTIVE DATA SETS IN PROJECT LETHE

Within LETHE the following datasets will be used:

- ➢ Existing Retrospective data from medical partners
- ➢ Master data file of homogenized retrospective datasets in medical partners

- Datasets generated from questionnaires from partner Combinostics

Below is a categorization of retrospective datasets that the 4 medical centers / hospitals have and will provide for further processing. Categorization entails:

- Theme / dataset name – is the name of the dataset
- Variable / group of variables are a level below theme/dataset name
- Examples are the data of each variable group

And the theme / dataset name is categorized as follows:

- Clinical markers
- cognition,
- cognition-related markers,
- demographics,
- functional status,
- health status,
- lifestyle,
- mood,
- quality of life
- study visits
- blood analysis
- CSF information
- Psychometry
- Risk factors
- Volumetry
- Social Background
- WML and PET
- Genetic information

Not all medical partners have exactly the same data structure and datasets information, hence the need for homogenization prior to be stored on LETHE servers and further processing takes place to derive risk factors.

16.2.1 THL Retrospective Dataset Description (Finland)

THL Retrospective Dataset Description			
Theme / dataset name	**Variable/ group of variables**	**Examples**	**Data Type**
Clinical markers	Anthropometrics	e.g. BMI	number
Clinical markers	Blood markers	e.g. cholesterol, glucose, Metabolism, Vitamin B12	number

Clinical markers	Blood pressure and pulse		number
Clinical markers	Other		number
Cognition	Clinical Dementia Rating (CDR)		number
Cognition	CERAD, MMSE		number
Cognition	cognition composite scores		number
Cognition	cognition individual tests		number
Cognition	Subjective memory		number
Cognition	Subjective memory reported by close relative		number
Cognition	other		number
Cognition-related markers	MRI, PET		number
Cognition-related markers	other		number
Demographics	Characteristics	e.g. sex, age	number
Demographics	Living situation	e.g. marriage status, living at home or in an institution	number
Demographics	Sosioeconomic status	e.g. education, income	number

Demographics	other	e.g	number
Functional status	ADL and IADL		number
Functional status	care need, aid use		number
Functional status	falls		number
Functional status	Short Physical Performance Battery		number
Functional status	other		number
Health status	All medications - current, self reported		number
Health status	Current health status, self-reported		number
Health status	Health history, self-reported		number
Health status	Health care use, self-reported		number
Health status	Malnutrition status		number
Health status	Women's questions (reproductive health)		number
Health status	Other		number
Lifestyle	Alcohol use		number

Lifestyle	Cognitive & social activity		num ber
Lifestyle	Computer and mobile device use		num ber
Lifestyle	Nutrition		num ber
Lifestyle	Physical activity		num ber
Lifestyle	Sleep and related		num ber
Lifestyle	Smoking		num ber
Lifestyle	Self-evaluated lifetyle changes		num ber
Lifestyle	other		num ber
Mood and QOL	Quality of life		num ber
Mood and QOL	Stress		num ber
Mood and QOL	Depressive symptoms	e.g. zung scale	num ber
Mood and QOL	other		num ber
Process data	logistics data		num ber
Process data	other		num ber

16.2.2 GEDOC Huddinge Retrospective Dataset Description (Karolinska Hospital)

GEDOC Huddinge Retrospective Dataset Description			
Theme / dataset name	Variable/ group of variables	Examples / description	Data Type
General information and demographics	Age		
General information and demographics	Sex		
General information and demographics	Datefirstvisit		
Cognition	MedicalINVorderno	Diagnosis order number visits	
Cognition	MedicalMASSmmse_score	MMSE score	
Cognition	MedicalMASSmmse_date	MMSE date	
Cognition	MedicalMASSmoca_date	MOCA date	
Cognition	MedicalMASSmoca_score	MOCA score	
Health status	MedicalMASShight	Height m	
Health status	MedicalM08_diag	Diagnosis	
Health status	MedicalM08_diag_alt	Diagnosis additional information	
Health status	MedicalDateofdiagnose	Date of diagnosis	
Health status	MedicalMASSGDS_score	GDS score	
Health status	MedicalMASSGDS_date	GDS date	

Health status	MedicalMASSweight	Weight kg	
Health status	MedicalMASSCornell_score	Cornell score	
Health status	MedicalMASSCornell_date	Cornell date	
Health status	MedicalMASSbmi_index	BMI	
Health status	MedicalM06ass_BP_syst	Bloodpressure systolic	
Health status	MedicalM06ass_BP_diast	Bloodpressure diastolic	
Blood Analysis	BloodanalysisBLOOD_creatinin	Creatinine (umol/l)	
Blood Analysis	BloodanalysisBLOODalbum	Albumine (g/L)	
Blood Analysis	BloodanalysisBLOODcalcium	Calcium (mmol/l)	
Blood Analysis	BloodanalysisBLOOD_t3	T3 (pmol/l)	
Blood Analysis	BloodanalysisBLOOD_tyroxin	Tyroxin T4 (pmol/l)	
Blood Analysis	BloodanalysisBLOOD_tsh	TSH (mU/l)	
Blood Analysis	BloodanalysisBLOOD_B12	Vitamine B12 (pmol/l)	
Blood Analysis	BloodanalysisBLOOD_Folic acid	Folic acid (nmol/l)	
Blood Analysis	BloodanalysisBLOOD_homocystei	Homocysteine (micromol/l)	
Blood Analysis	BloodanalysisBLOOD_mma	Malonic acid (micromol/l)	
Blood Analysis	BloodanalysisBLOOD_Triglyceri	Triglycerides (mmol/l)	
Blood Analysis	BloodanalysisBLOOD_cholestero	Cholesterol (mmol/l)	

Blood Analysis	BloodanalysisBLOOD_LDL	LDL(mmol/l)	
Blood Analysis	BloodanalysisBLOOD_HDL	HDL(mmol/l)	
Sample Information	ProverProv_ID_remiss	Sample ID	
Sample Information	ProverVolym	Volume sample	
Sample Information	ProverProvstatus	Status sample	
Sample Information	ProverVolumsts	Status volume sample	
Sample Information	ProverProvtyp	Type of sample	
Sample Information	Biobank_remissRemissnr	Biobank ID	
Sample Information	Biobank_remissRegenhet	Location samples	
Sample Information	Biobank_remissProvtagningsdatu	Date blood sample	
Genetic information	GeneticsDateofinvestigation	ApoE date	
Genetic information	GeneticsGENapoe_a1	ApoE allel1	
Genetic information	GeneticsGENapoe_a2	ApoE allel2	
CSF Information	LPINVorderno	LP order number	
CSF Information	LPDateofinvestigation	LP date	
CSF Information	LPLP_laboratory		

CSF Information	LPLP_spAlbumin	LP albumin (mg/l)	
CSF Information	LPLP_pAlbumin	Plasma albumine (g/l)	
CSF Information	LPLP_spdivpAlbumin	LP Plasma albumine ratio (x10^-3)	
CSF Information	LPLP_Incrpermbloodbrbarr	LP increased BBB permeability	
CSF Information	LPLP_sp_t_Tau	LP t Tau (ng/l)	
CSF Information	LPLP_spAbeta40	LP Ab40 (pM)	
CSF Information	LPLP_spAbeta42	LP Ab42 (pM)	
CSF Information	LPLP_fosfo_Tau	LP f Tau (ng/l)	
CSF Information	LPLP_betaamyloid	LP betaamyloid (ng/l)	
CSF Information	LPLP_abetakvot	LP Ab42 40 ratio (x10 ratio)	
CSF Information	LPLP_Neurofilamentlightprote	LP NFL (ng/l)	
CSF Information	LPINVorderno	LP order number	
Psychometry	PsykometriINVorderno	Psychometry order number	
Psychometry	PsykometriPsyR_ExamDate	Psychometry date	
Psychometry	PsykometriPsyR_FASF1 (& F2– F6)	FAS F 1 – F 6	
Psychometry	PsykometriPsyR_FASF	FAS F sum	
Psychometry	PsykometriPsyR_AVLT11 (& 21– 51)	RAVLT Immediate Recall column1-5 row1	
Psychometry	PsykometriPsyR_AVLT1	RAVLT Immediate Recall sum column1	
Psychometry	PsykometriPsyR_FASA1 (& A2– A6)	FAS A 1 – A 6	

Psychometry	BE	FAS A sum	
Psychometry	PsykometriPsyR_AVLT2	RAVLT Immediate Recall sum column2	
Psychometry	PsykometriPsyR_AVLT12 (& 12–52)	RAVLT Immediate Recall column1–5 row2	
Psychometry	PsykometriPsyR_FASS	FAS S sum	
Psychometry	PsykometriPsyR_AVLT3	RAVLT Immediate Recall sum column3	
Psychometry	PsykometriPsyR_FASS1 (& S2–S6)	FAS S 1 – S 6	
Psychometry	PsykometriPsyR_AVLT13 (& 23–53)	RAVLT Immediate Recall column1–5 row3	
Psychometry	PsykometriPsyR_FASTot	FAS Total	
Psychometry	PsykometriPsyR_AVLT4	RAVLT Immediate Recall sum column4	
Psychometry	PsykometriPsyR_AVLT14 (& 24–54)	RAVLT Immediate Recall column1–5 row4	
Psychometry	PsykometriPsyR_AVLT5	RAVLT Immediate Recall sum column5	
Psychometry	PsykometriPsyR_AVLTtot	RAVLT Immediate Recall sum total	
Psychometry	PsykometriPsyR_AVLT15 (&25–55)	RAVLT Immediate Recall column1–5 row5	
Psychometry	PsykometriPsyR_AVLTRet	RAVLT retention total	
Psychometry	PsykometriPsyR_WDigSym	Digit symbol	
Psychometry	PsykometriPsyR_WDigSpan	Digit span	
Psychometry	PsykometriPsyR_TMTACorr	TMT A Correct Connection	
Psychometry	PsykometriPsyR_DigitsForwMax	Digits Forward Max	

Psychometry	PsykometriPsyR_TMTA	TMT A Time	
Psychometry	PsykometriPsyR_DigitsBackwMax	Digits Backward Max	
Psychometry	PsykometriPSYR_TMTBcorr	TMT B Correct Connection	
Psychometry	PsykometriPSYR_TMTB	TMT B Time	
Psychometry	PsykometriPsyR_LuClockRead	Luria Clock Reading	
Psychometry	PsykometriPsyR_RCFTRe	RCFT recall	
Psychometry	PsykometriPsyR_WBlock	WAIS Block design	
Psychometry	PsykometriPsyR_RCFT	RCFT copy	
Psychometry	PsykometriPsyR_RCFTTime	RCFT copy time	
Psychometry	PsykometriPsyR_LuCopyCube	Luria Copy Cube 3D	
Psychometry	PsykometriPsyR_LuCopyCross	Luria Copy Cross 3D	
Psychometry	PsykometriPsyR_LuClockDraw	Luria Clock Drawing	
Psychometry	PsykometriPsyR_WInformation	General knowledge test	
Psychometry	PsykometriPsyR_WSimilar	Similarities sum	
Risk Factors	Smoking		
Risk Factors	M06narcotics	Narcotics	
Risk Factors	M06alco_lightbeer	Alcohol Lightbeer	
Risk Factors	M06alco_strongbeer	Alcohol strongbeer	

Risk Factors	M06alco_lightwine	Alcohol lightwine	
Risk Factors	M06alco_strongwine	Alcohol strongwine	
Risk Factors	M06alco_hardliquor	Alcohol hardliquor	
Risk Factors	M06alco_overconsumer	Alcohol overconsumtion	
Risk Factors	M06exposure_chem_solv	Exposure chemicals	
Risk Factors	Headinjury		
Risk Factors	M_heredity	Heredity	
Risk Factors	M06her_father	Heredity father	
Risk Factors	M06her_mother	Heredity mother	
Risk Factors	M06her_brother	Heredity brother	
Risk Factors	M06her_sister	Heredity sister	
Risk Factors	M06her_other	Heredity other	
Risk Factors	Heartdisease		
Risk Factors	M06_treat_heart_angina	Angina Pectoris	
Risk Factors	M06depression_diagnosed	Depression	
Risk Factors	M06_treat_heart_MI	Myocardial Infarction	
Risk Factors	M06depression_treated	Depression treated	
Risk Factors	M06_treat_heart_arrytmia	Arrhythmia	
Risk Factors	M06depression_episodes	Depression number episodes	
Risk Factors	M06_treat_heart_cardiacfailure	Heart failure	

Risk Factors	M06depression_episodes_last	Depression last episode date	
Risk Factors	M06_treat_heart_bypassoperation	Heart bypass surgery	
Risk Factors	CerebInf_or_TIA	TIA	
Risk Factors	M06_treat_TIAepisods	TIA number episodes	
Risk Factors	M06_treat_TIAepisods_last	TIA last episode date	
Risk Factors	M06_stroke_when	Stroke date	
Risk Factors	M06_stroke_sequele	Stroke sequele	
Risk Factors	M06_stroke_ischemical	Stroke ischemical	
Risk Factors	M06_stroke_hemorragical	Stroke hemorragical	
Risk Factors	M06_treat_EP	Epilepsia	
Risk Factors	M06_stroke_unknown	Stroke unknown	
Risk Factors	M06_treat_Parkins	Parkinsons Disease	
Risk Factors	M06_treat_hypothyreos	Hypothyreosis	
Risk Factors	M06_treat_lowB12_folat	Deficiency B12 Folicacid	
Risk Factors	M06_hypertoni	Hypertension	
Risk Factors	M06_treat_hyperlipidemi	Hyperlipidemia	
Risk Factors	M06_treat_diabetes	Diabetes	
Risk Factors	MRISCvertigo	Vertigo	
Social Background	education_years		

Social Background	Typeofliving		
Social Background	education_type		
Social Background	Live_alone		
Social Background	Homeservice		
Social Background	P06_mainoccup	Main occupation	
Social Background	Workingpercent		
Social Background	Sickleave_retired		
Volumetry	VolumetryVOLmridate	MRI date	
Volumetry	VolumetryVOLrtl	Absolute brain volume right temporal lobe	
Volumetry	VolumetryVOLrmtl	Absolute brain volume right medial temporal lobe	
Volumetry	VolumetryVOLltl	Absolute brain volume left temporal lobe	
Volumetry	VolumetryVOLlmtl	Absolute brain volume left medial temporal lobe	
Volumetry	VolumetryVOLtotbrain	Absolute brain volume total brain	
Volumetry	VolumetryVOLintracram	Absolute brain volume intra cranial volume	
Volumetry	VolumetryVOLrelrtl	Relative brain volume right temporal lobe	

Volumetry	VolumetryVOLrelrtl	Relative brain volume right medial temporal lobe	
Volumetry	VolumetryVOLrelltl	Relative brain volume left temporal lobe	
Volumetry	VolumetryVOLrellmtl	Relative brain volume left medial temporal lobe	
Volumetry	VolumetryVOLreltotbrain	Relative total brain	
WML and PET	WMLINVorderno	WML order number	
WML and PET	WMLWML_MTA_höger_mri	WML MRI MTA right	
WML and PET	WMLWML_Fazekes_mri	WML MRI Fazekas	
WML and PET	WMLWML_Fazekes_score_mri	WML MRI Fazekas score	
WML and PET	WMLWML_MTA_vänster_mri	WML MRI MTA left	
WML and PET	WMLWMLmridate	WML MRI date	
WML and PET	WMLWML_DatScandate	WML DatScan date	
WML and PET	WMLWML_MTA_höger_ct	WML CT MTA right	
WML and PET	WMLWMLctdate	WML CT date	
WML and PET	WMLWML_Fazekes_ct	WML CT Fazekas	
WML and PET	WMLWML_Fazekes_score_ct	WML CT Fazekas score	
WML and PET	WMLWML_MTA_vänster_ct	WML CT MTA left	
WML and PET	WMLWML_PETdate	WML PET date	

WML and PET	WMLWML_PETtext	PET type	
WML and PET	WMLDateofinvestigation	WML date	
WML and PET	WMLWMLfrontwmlDX	WML WML Frontal lobe right	
WML and PET	WMLWMLfrontwmlSIN	WML WML Frontal lobe left	
WML and PET	WMLWMLnuclcaudDX	WML Basal ganglia nucleus caudatus right	
WML and PET	WMLWMLnuclcaudSIN	WML Basal ganglia nucleus caudatus left	
WML and PET	WMLWMLcentinfDX	WML Additional information Central infarct right	
WML and PET	WMLWMLcentinfSIN	WML Additional information Central infarct left	
WML and PET	WMLWMLoccipcapsDX	WML Caps Bands Occipital lobe caps right	
WML and PET	WMLWMLoccipcapsSIN	WML Caps Bands Occipital lobe caps left	
WML and PET	VolumetryWMLcerebDX	WML Infra tentorial cerebellum right	
WML and PET	WMLWMLcerebSIN	WML Infra tentorial cerebellum left	
WML and PET	WMLWMLsumtotscore	WML sum total	
WML and PET	WMLWMLparietwmlDX	WML WML Parietal lobe right	
WML and PET	WMLWMLparietwmlSIN	WML WML Parietal lobe left	
WML and PET	WMLWMLputamDX	WML Basal ganglia putamen right	
WML and PET	WMLWMLputamSIN	WML Basal ganglia putamen left	
WML and PET	WMLWMLcortinfDX	WML Additional information Cortical Infarct right	

WML and PET	WMLWMLcortinfSIN	WML Additional information Cortical Infarct left	
WML and PET	WMLWMLfrontcapsDX	WML Caps Bands Frontal lobe caps right	
WML and PET	WMLWMLfrontcapsSIN	WML Caps Bands Frontal lobe caps left	
WML and PET	WMLWMLmesenceDX	WML Infra tentorial mesencephalon right	
WML and PET	WMLWMLmesenceSIN	WML Infra tentorial mesencephalon left	
WML and PET	WMLWMLoccipwmlDX	WML WML Occipital lobe right	
WML and PET	WMLWMLoccipwmlSIN	WML WML Occipital lobe right	
WML and PET	WMLWMLglobpallDX	WML Basal ganglia globus pallidus right	
WML and PET	WMLWMLglobpallSIN	WML Basal ganglia globus pallidus left	
WML and PET	WMLWMLhemoDX	WML Additional information Hemorrhage right	
WML and PET	WMLWMLhemoSIN	WML Additional information Hemorrhage left	
WML and PET	WMLWMLlatventrDX	WML Caps Bands Lateral ventricle bands right	
WML and PET	WMLWMLlatventrSIN	WML Caps Bands Lateral ventricle bands left	
WML and PET	WMLWMLpons	WML Infra tentorial pons	
WML and PET	WMLWMLtempwmlDX	WML WML Temporal lobe right	
WML and PET	WMLWMLtempwmlSIN	WML WML Temporal lobe left	
WML and PET	WMLWMLthalamDX	WML Basal ganglia Thalamus right	
WML and PET	WMLWMLthalamSIN	WML Basal ganglia Thalamus left	

WML and PET	WMLWMLcontusDX	WML Additional information Contusion right	
WML and PET	WMLWMLcontusSIN	WML Additional information Contusion left	
WML and PET	WMLWMLmedulobl	WML Infra tentorial medulla oblongata	
WML and PET	WMLWMLsum1	WML sum Caps Bands	
WML and PET	WMLWMLsum2	WML sum WML	
WML and PET	WMLWMLsum3	WML sum Basal ganglia	
WML and PET	WMLWML_NPH	WML Additional information Normal pressure hydrocephalus	
WML and PET	WMLWMLsum4	WML sum Infra tentorial	
WML and PET	WMLWMLsum5	WML sum Additional information	

16.2.3 GEDOC Solna Retrospective Dataset Description (Karolinska Hospital)

GEDOC Solna Retrospective Dataset Description			
Theme / dataset name	**Variable/ group of variables**	**Examples / description**	**Data Type**
General information and demographics	Age	Age (years)	
General information and demographics	Sex	Sex	
General information and demographics	Diagnos_grupp	Diagnosis group	

General information and demographics	Rokare-tobak	Smoking currently or has ever been smoking	
General information and demographics	Social_status	Cohabitation status	
General information and demographics	Modersmal___svenska__	Patients first language Swedish	
General information and demographics	Education	Education (years)	
Health Status	Stroke	Ever been diagnosed for stroke	
Health Status	Parkinsons_disease	Ever been diagnosed for Parkinson's disease	
Health Status	MI	Ever been diagnosed for myocardial infarction	
Health Status	DM_1	Ever been diagnosed for diabetes type 1	
Health Status	DM_2_	Ever been diagnosed for diabetes type 2	
Health Status	Thyroid_disease	Ever been diagnosed for hypothyreosis, hyperthyreosis or other thyroid disease	
Health Status	Somnproblem	Ever been diagnosed for sleep disorder	
Health Status	OSAS	Ever been diagnosed for obstructive sleep apnea syndrome	
Health Status	Depression	Ever been diagnosed for depression	
Health Status	PHQ9	Patient Health Questionnaire-9	
Health Status	angest_	Ever been diagnosed for anxiety disorder	

Health Status	PTSD	Ever been diagnosed for PTSD	
Health Status	Malignancy	Ever been diagnosed for malignancy	
Health Status	Utmattning	Ever been diagnosed for burn-out according to Swedish ICD-10 code F43.8A	
Health Status	Operation	Ever had an operation	
Health Status	Nr_medication	Total number of medication during hospital visit	
Health Status	S_BP	Systolic blood pressure (mmHg)	
Health Status	D_BP	Diastolic blood pressure (mmHg)	
Health Status	BMI	Body mass index (kg/m2)	
Health Status	Phase_Angle	Phase angle measurement	
Health Status	PA_Percentile	Phase angle, percentile	
Health Status	GH_UU	10-meter walk test (m/s)	
Health Status	GH_MU1	10-meter walk test, with assignment 1 (m/s)	
Health Status	GH_MU2	10-meter walk test, with assignment 2 (m/s)	
Health Status	GH_MU3	10-meter walk test, with assignment 3 (m/s)	
Health Status	GS	Hand grip strength test (kg), dominant hand	
Health Status	30s_UPP	The 30-second chair stand test, total	
Health Status	Balans	The four stage balance test, total	
Health Status	Fall_	Has ever fallen	

Blood and CSF Samples	Sparade_prover	Saved blood and/or CSF samples	
Blood and CSF Samples	Sparad_serum	Saved blood serum sample	
Blood and CSF Samples	Sparad_plasma_	Saved plasma sample	
Blood and CSF Samples	Sparad_CSF	Saved CSF sample	
Blood and CSF Samples	Homocystein	Plasma homocysteine (µmol/l)	
Blood and CSF Samples	Vitamine_D	Serum 25-OH-vitamin D (nmol/l)	
Blood and CSF Samples	TSH	Serum TSH (mE/l)	
Blood and CSF Samples	T4	Serum free T4 (pmol/l)	
Blood and CSF Samples	T_Chol	Plasma total cholesterol (mmol/l)	
Blood and CSF Samples	HDL_Chol	Plasma HDL cholesterol (mmol/l)	
Blood and CSF Samples	LDL_Chol	Plasma LDL cholesterol (mmol/l)	
Blood and CSF Samples	HbA1c	Hemoglobin A1c, HbA1c (IFCC, mmol/mol)	
Blood and CSF Samples	Apo_grupp	ApoE group	
Blood and CSF Samples	APO_E	ApoE alleles	

Blood and CSF Samples	csv_Tau_KUL	CSF Total Tau (ng/l)	
Blood and CSF Samples	csv_Ptau_KUL	CSF Phosphorylated tau (ng/l)	
Blood and CSF Samples	csv_AB42_KUL	CSF beta-amyloid 42 (ng/l)	
Blood and CSF Samples	csv_AB42_40_KUL	CSF beta-amyloid 42/40	
Blood and CSF Samples	csv_AB42_Ptau_KUL	Beta-amyloid 42/phosphorylated tau ratio	
Blood and CSF Samples	csv_NFL_KUL	CSF neurofilament light chain (ng/l)	
Cognition	MMSE	Mini-Mental State Examination (MMSE), total	
Cognition	MoCA_total	Montreal Cognitive Assessment (MoCa), total	
Cognition	MoCA_MIS	Montreal Cognitive Assessment (MoCa), Memory Index Score (MIS)	
Cognition	R30_tot	Rey Auditory Verbal Learning Test (RAVLT), learning	
Cognition	R30	Rey Auditory Verbal Learning Test (RAVLT), free immediate recall	
Cognition	H1	Hagman's visual test, first administration	
Cognition	H2	Hagman's visual test, second administration	
Cognition	RCF	Rey Complex Figure (RCF), recall	

Cognition	KOD	Wechsler Adult Intelligence Scale 4th Edition (WAIS-IV), coding	
Cognition	NS_Frontal	Ragnar Åstrand Cognitive Impairment Questionnaire, frontal total	
Cognition	NS_Par__Temp	Ragnar Åstrand Cognitive Impairment Questionnaire, parieto-temporal total	
Cognition	NS_Subk	Ragnar Åstrand Cognitive Impairment Questionnaire, subcortical total	
Cognition	NS_Minne	Ragnar Åstrand Cognitive Impairment Questionnaire, memory total	
Cognition	NS_Ass	Ragnar Åstrand Cognitive Impairment Questionnaire, associated symptoms total	
Cognition	FAS	F-A-S Verbal Phonemic Fluency Test, total	
Cognition	Djur	Category Fluency Test, animals	
Neuroimaging	Rontgen	Medical visual imaging available, CT or MRI	
Neuroimaging	MTA_va	Medial temporal lobe atrophy (MTA) score, left	
Neuroimaging	MTA_ho	Medial temporal lobe atrophy (MTA) score, right	
Neuroimaging	GCA	Global cortical atrophy (GCA) scale, total	
Neuroimaging	PA	Posterior atrophy (PA) score or Koedam score, total	
Neuroimaging	Fazekas	Fazekas, total	
Neuroimaging	PET	Medical visual imaging available, PET	

Neuroimaging	Brain_tissue___all_regions___tot	MRI Variable	
Neuroimaging	3rd_ventricle_total_volume	MRI Variable	
Neuroimaging	4th_ventricle_total_volume	MRI Variable	
Neuroimaging	5th_ventricle_total_volume	MRI Variable	
Neuroimaging	Accumbens_area_right_volume	MRI Variable	
Neuroimaging	Accumbens_area_left_volume	MRI Variable	
Neuroimaging	Amygdala_right_volume	MRI Variable	
Neuroimaging	Amygdala_left_volume	MRI Variable	
Neuroimaging	Brain_stem_total_volume	MRI Variable	
Neuroimaging	Caudate_right_volume	MRI Variable	
Neuroimaging	Caudate_left_volume	MRI Variable	

16.2.4 Austria Clinical Retrospective Dataset Description

Austria Clinical Retrospective Dataset Description			
Theme / dataset name	**Variable/ group of variables**	**Examples**	**Data Type**

Clinical markers	Anthropometrics	e.g. BMI	
Clinical markers	Blood markers	e.g. cholesterol, glucose, Metabolism, Vitamin B12	
Clinical markers	Blood pressure and pulse		
Clinical markers	Other		
Cognition	Clinical Dementia Rating (CDR)		
Cognition	CERAD, MMSE		
Cognition	cognition composite scores		
Cognition	cognition individual tests		
Cognition	Subjective memory		
Cognition	Subjective memory reported by close relative		
Cognition	other		
Cognition-related markers	MRI, PET		
Cognition-related markers	other		
Demographics	Characteristics	e.g. sex, age	
Demographics	Living situation	e.g. marriage status, living at home or in an institution	
Demographics	Socioeconomic status	e.g. education, income	

Demographics	other	e.g.	
Functional status	ADL and IADL		
Functional status	care need, aid use		
Functional status	falls		
Functional status	Short Physical Performance Battery		
Functional status	other		
Health status	All medications - current, self-reported		
Health status	Current health status, self-reported		
Health status	Health history, self-reported		
Health status	Health care use, self-reported		
Health status	Malnutrition status		
Health status	Women's questions (reproductive health)		
Health status	Other		
Lifestyle	Alcohol use		
Lifestyle	Cognitive & social activity		
Lifestyle	Computer and mobile device use		

Lifestyle	Nutrition		
Lifestyle	Physical activity		
Lifestyle	Sleep and related		
Lifestyle	Smoking		
Lifestyle	Self-evaluated lifestyle changes		
Lifestyle	other		
Mood and QOL	Quality of life		
Mood and QOL	Stress		
Mood and QOL	Depressive symptoms	e.g. Zung scale	
Mood and QOL	other		
Process data	logistics data		
Process data	other		

16.2.5 Austria Insurance Retrospective Dataset Description

Austria Clinical Retrospective Dataset Description			
Theme / dataset name	**Variable/ group of variables**	**Examples**	**Data Type**
Clinical markers	Anthropometrics	e.g. BMI	
Clinical markers	Blood markers	e.g. cholesterol, glucose, Metabolism, Vitamin B12	
Clinical markers	Blood pressure and pulse		

Clinical markers	Other		
Cognition	Clinical Dementia Rating (CDR)		
Cognition	CERAD, MMSE		
Cognition	cognition composite scores		
Cognition	cognition individual tests		
Cognition	Subjective memory		
Cognition	Subjective memory reported by close relative		
Cognition	other		
Cognition-related markers	MRI, PET		
Cognition-related markers	other		
Demographics	Characteristics	e.g. sex, age	
Demographics	Living situation	e.g. marriage status, living at home or in an institution	
Demographics	Socioeconomic status	e.g. education, income	
Demographics	other	e.g.	
Functional status	ADL and IADL		
Functional status	care need, aid use		

Functional status	falls		
Functional status	Short Physical Performance Battery		
Functional status	other		
Health status	All medications - current, self-reported		
Health status	Current health status, self-reported		
Health status	Health history, self-reported		
Health status	Health care use, self-reported		
Health status	Malnutrition status		
Health status	Women's questions (reproductive health)		
Health status	Other		
Lifestyle	Alcohol use		
Lifestyle	Cognitive & social activity		
Lifestyle	Computer and mobile device use		
Lifestyle	Nutrition		
Lifestyle	Physical activity		
Lifestyle	Sleep and related		
Lifestyle	Smoking		

Lifestyle	Self-evaluated lifestyle changes		
Lifestyle	other		
Mood and QOL	Quality of life		
Mood and QOL	Stress		
Mood and QOL	Depressive symptoms	e.g. Zung scale	
Mood and QOL	other		
Process data	logistics data		
Process data	other		

16.2.6 University of Perugia pilot user

University of Perugia Retrospective Dataset Description			
Theme / dataset name	**Variable/ group of variables**	**Examples**	**Data Type**
Clinical markers	Anthropometrics	e.g. BMI	
Clinical markers	Blood markers	e.g. cholesterol, glucose, Metabolism, Vitamin B12	
Clinical markers	Blood pressure and pulse		
Cognition	Clinical Dementia Rating (CDR)		
Cognition	CERAD, MMSE		
Cognition	cognition individual tests		

Cognition	Subjective memory	If SCI is diagnosed	
Cognition-related markers	MRI, PET	MRI available for most subjects, PET for a subgropup	
Demographics	Characteristics	e.g. sex, age	
Demographics	Living situation	e.g. marriage status, living at home or in an institution	
Demographics	Socioeconomic status	e.g. education, income	
Functional status	ADL and IADL		
Functional status	care need, aid use		
Functional status	falls		
Functional status	Short Physical Performance Battery	since 2021. Before the Elderly mobility Scale has been used	
Health status	All medications - current, self-reported		
Health status	Current health status, self-reported		
Health status	Health history, self-reported		
Health status	Malnutrition status	Mini Nutritional Assessment	
Health status	Women's questions (reproductive health)		
Lifestyle	Alcohol use		

Lifestyle	Nutrition	MNA	
Lifestyle	Sleep and related		
Lifestyle	Smoking		
Lifestyle	Self-evaluated lifestyle changes		
Mood and QOL	Depressive symptoms	Geriatric Depression Scale	
Process data	logistics data	Only UPG data	

16.3 APPENDIX: Combinostics Tools to provide Data

cCOG is a web-based cognitive test battery consisting of seven tasks [1]. Within LETHE it will be used to quantify different domains of cognition: episodic memory, processing speed and executive function, and attention and reaction time. The performance in each task is characterized using a specific digital biomarker and a percentile value which is obtained by contrasting the biomarker value with age and education normalized reference data. In addition, a composite score reflecting the overall cognitive performance is quantified.

In LETHE, an additional questionnaire will be implemented for quantifying the risk of dementia. **The CAIDE risk score** [2] will be implemented consisting of the following factors: age, education time, sex, systolic blood pressure, body mass index, total cholesterol level and level of exercise. The risk score provides a value between 0 (low risk) and 15 (high risk). Another relevant and validated risk scores may be added to cCOG if so decided during the project. Below the CAIDE risk score and dataset description that will be used in the pilot cases.

CAIDE		
Age (years)	<47	0
	47-53	3
	>53	4

Education time (years)	≥10	0
	7-9	2
	0-6	3
Sex	Female	0
	Male	1
Systolic blood pressure	≤140	0
	≥140	2
Body mass index	≤30	0
	≥30	2
Cholestrol level	≤6.5	0
	≥6.5	2
Excercise	Active	0
	Non-active	1
Points	**Risk (%)**	
0-5	1	
6-7	1.9	
8-9	4.2	
10-11	7.4	
12-15	16.4	

[1] H. Rhodius-Meester, T. Paajanen, J. Koikkalainen, S. Mahdiani, M. Bruun, M. Baroni, A. Lemstra, P. Scheltens, S-K. Herukka, M. Pikkarainen, A. Hall, T. Hänninen, T. Ngandu, M. Kivipelto, M. van Gils, S.

Hasselbalch, P. Mecocci, A. Remes, H. Soininen, W. van der Flier and J. Lötjönen. cCOG: a web-based cognitive test tool for detecting neurodegenerative disorders. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring, 2020. https://doi.org/10.1002/dad2.12083

[2] M. Kivipelto, T. Ngandu, T. Laatikainen, B. Winblad, H. Soininen, J. Tuomilehto. Risk score for the prediction of dementia risk in 20 years among middle aged people: A longitudinal, population-based study. Lancet Neurol. 2006;5:735–741.

16.4 APPENDIX: Technical and organizational measures for EGI

This document describes the technical and organizational measures implemented by EGI Foundation to meet legal and contractual requirements when processing personal data.

The measures described in numbers 1 to 13 serve the purpose …

- to encrypt or pseudonymize personal data where necessary (see, inter alia, 6 to 8),
- to ensure the confidentiality, integrity, availability and resilience of systems and services in connection with the processing of personal data (see, inter alia, 1 to 10),
- to restore the availability of and access to personal data in the event of a physical or technical incident in a timely manner (see 11), and
- to regularly review, assess and evaluate the effectiveness of all technical and organizational measures to ensure the security of processing (see, inter alia, 12 and 13).

The following measures apply to all data processing activities that are under control of EGI Foundation, or where EGI Foundation is a subcontracted data processor on behalf of another data controller.

In situations where EGI Foundation is the data controller and another organization is the data processor on behalf of EGI Foundation, EGI Foundation aims at ensuring that the technical and organizational measures implemented by the subcontracted processor equals at minimum the processing security level indicated by following measures.

Please note: In federated service delivery scenarios, one or more data controllers and one or more subcontracted data processors may be entrusted with or involved in processing personal data.

(1) Access control

All access rights (both for access to IT systems and data and for access to buildings and rooms) are assigned according to the principle that employees and third-party users are only granted the level of access they need to perform their activities (need-to-know principle).

Access rights are granted according to defined (role-based) permission profiles. The access rights granted are reviewed regularly. Rights that are no longer required are withdrawn immediately.

Access to networks and network services is restricted by technical and physical measures. Access to wireless corporate networks that allow access to personal data is protected by personalized authentication (PKI, IEEE

802.1x). This applies in an analogous manner to wired access, unless it is from a secure area that is sufficiently protected and controlled by physical access control measures.

(2) Physical access control

Physical secure areas (zones) are defined on the basis of information security and data protection requirements and protected against unauthorized access by appropriate physical safeguards. The physical security concept distinguishes between public areas, controlled areas, restricted/internal areas, and high-risk zones. Secure zones are defined based on the protection needs of the information assets housed or made accessible within them.

Depending on the specific zone classification, selected or all of the following security features are implemented: Access restriction through personalized access medium, video surveillance and door-open sensors at access points, motion detection, privacy screens or view guards on potentially confidential information, and no photography policy.

For dealing with visitors and deliveries, procedures are used to prevent unauthorized persons from accessing security areas.

(3) Logical access control to processing systems

All data processing systems are equipped with a secure authentication mechanism (X509 certificate or password protection).

Defined procedures are used to authorize access to information, taking into account the need-to-know principle. Special procedures are in place for granting access rights to privileged systems (e.g., systems or applications used to control or administrate critical processes or to manage access rights for other systems).

For authentication on data processing systems (IT systems), secure passwords are used that have sufficient length, are robust against dictionary attacks, do not contain strings of consecutive letters or digits and are not based on facts that could be easily be guessed by others. Passwords must be changed whenever there is an indication that the password has been compromised. A changed password must not match or contain a password that has been used in the past. Where technically possible, the use of two-factor-authentication is enforced.

A " clear desk & screen policy" is implemented: When leaving the workplace, all computers in use must be locked (screen lock). In case of inactivity, the screen lock is automatically activated after a maximum of 10 minutes. Documents that may contain confidential information must not be kept open and unattended on desks or in other freely accessible storage areas.

(4) User activity control

All employees must attend mandatory basic training on information security and data privacy on an annual basis. Participation in this training is recorded. New employees are familiarized with the main regulations on information security and data privacy relevant to them at the start of their employment or assignment.

User activities, including logon attempts to data processing systems (IT systems), are logged to the extent required.

User accounts via which personal data can be accessed as part of processing activities must be personalized and must not be shared by more than one person.

Administrative activities on IT systems (such as changes to system configurations) are logged. Configuration files are historized, backed up and checked regularly and as required.

(5) Segregation control

It is ensured that personal data collected for different purposes are not mixed in their processing. To this end, multitenant systems are used where necessary, or systems are physically or logically separated.

(6) Data carrier and mobile device control

Data carriers containing personal data are stored in secure locations that prevent access to these carriers by unauthorized persons.

Personal data stored on mobile devices and data carriers (including laptops, smartphones, USB sticks) are required to be encrypted. The use of any type of private Internet/Cloud storage for the (temporary) storage of such data is prohibited. Confidential data will never be stored on private storage media or end devices.

Personal data that are no longer required are deleted. Electronic storage media and paper documents that are no longer required will be disposed of or destroyed / made unusable in such a way that it is no longer possible to gain knowledge of the data stored or contained on them.

The use of mobile devices is restricted and controlled. If personal data are accessed via mobile devices, suitable measures are taken to ensure that the devices cannot be used by unauthorized persons, for example in the event of loss or theft. All mobile devices used for business purposes are configured in such a way that they are protected by a query for a secret (e.g., PIN, pattern or biometric information) in the lock screen. The lock screen is automatically activated during inactivity. The corresponding mobile devices must never be left unattended. Modifications to the operating system software / firmware are prohibited. Security-relevant updates and patches are applied automatically. The devices are subject to comprehensive mobile device management (MDM), which technically implements these and other restrictions, policies, and measures.

(7) Pseudonymization and anonymization

Measures for pseudonymization or anonymization of personal data are implemented to the extent necessary. Data in development environments used for testing purposes is anonymized or pseudonymized wherever possible. Data on the usage of websites that is evaluated to generate usage statistics is anonymized.

(8) Transfer and dissemination control

Mechanisms for securing data traffic and communication connections, as well as for monitoring and logging activities in networks, have been established to the required extent. As appropriate, firewalls and intrusion detection and prevention systems (IDS / IPS) are in place.

When personal data is transmitted via public communication networks, secure end-to-end encryption of the communication is ensured. When establishing secure connections (VPN tunnels) offering access to IT resources via public networks, two-factor authentication is used as a matter of principle. If the exchange of confidential authentication information is required, this is done via a different communication path than the actual data transmission.

When transporting personal data stored on data carriers, the use of encryption, among other things, ensures that the data is protected against unauthorized access, manipulation or loss. After transport, the data is deleted from the storage media used for transport if it is no longer required on them.

Paper printouts and exports of confidential data from their source system are avoided whenever possible. Hard copies and electronic exports of confidential information leaving the business premises are handled with special care, taking into account the relevant confidentiality level - with the aim of preventing disclosure, loss and unauthorized copying. As soon as a paper printout is no longer required, it is destroyed. Electronic data exports that are no longer required are deleted again from the respective storage location and any transport data carrier used.

(9) Input control

Measures for subsequent verification of whether and by whom data has been entered, changed or removed (deleted) are implemented to the extent necessary. In systems used to collect and process personal data, access is categorized and automatically recorded. The integrity of log information is ensured.

(10) Availability control

A redundant design of communication and data processing systems (IT systems) and supporting facilities has been implemented to the required extent. An uninterruptible power supply (UPS) and high-availability Internet connection with automatic failover have been implemented at all relevant locations. Server and storage systems are designed redundantly (including redundant power supply units, disk mirroring). As appropriate, load balancing and failover are implemented for virtualized server systems.

(11) Recoverability

Data backups of databases and operating system images are taken to the extent required and with the aim of preventing the loss of personal data in the event of a technical malfunction or human error. Backups are performed for network drives and servers in productive operation, and the performance is recorded (logged) and monitored. The recovery of data backups is tested.

Processes or procedures for handling disruptions to IT systems and for restoring systems after a disruption have been established to the extent required.

Business continuity management (BCM) includes activities for business process impact analysis (BIA), definition and application of measures to ensure business continuity, taking into account information security and data protection aspects, as well as tests and reviews of the effectiveness of the measures implemented. A business process impact analysis is prepared or reviewed at least annually on the basis of the key business processes and services.

(12) Job control and subcontracting

The selection of subcontractors is carried out with the objective of ensuring that there is no increased risk to compliance with data protection objectives.

Depending on their role and the scope of access to confidential or personal data, subcontractors must, among other things, acknowledge and comply with regulations on secrecy / confidentiality as well as data protection (e.g., confidentiality / non-disclosure agreement), as well as an information security policy for suppliers.

In the case of security-critical subcontractors, service providers or suppliers, the following reporting and audit requirements are implemented: evaluation of contractually agreed reports (e.g., security events/incidents, availability statistics) as well as supplier audits using a self-assessment questionnaire, with an additional on-site inspection as necessary.

(13) Review, assessment and evaluation

Information on potential technical vulnerabilities or errors in data processing systems (IT systems) is evaluated at regular intervals and appropriate measures are initiated. Critical patches are deployed for both operating systems and software applications in use.

Data processing systems (IT systems) are checked regularly to the extent required and after changes to ensure that they are functioning properly.

An internal audit program is in place that covers regular system audits, process audits, IT security audits and data protection audits and controls.

16.5 APPENDIX: Data Object Characteristics and Object Identifiers

Please note that sections B–E are heavily based on the current DataCite (The Metadata Working Group, 2015) metadata specification and so are dealt with, relatively briefly. Elements referenced (e.g. F.6.2, F.7.1) in the following paragraphs can be found in original paper (Canham & Ohmann, 2016) and DataCite (The Metadata Working Group, 2015).

B.1 Data object identifier

Data objects available publicly, such as journal articles, plus some of the data sets and protocols in repositories, should have a DOI (in line with the DataCite specification). As discussed in the Methods section above, non-public data objects should also, wherever possible, also have a DOI. If a DOI is not possible, or has not yet been assigned, then the object should be identified either by an accession number from a metadata repository system or by using the object's name and version code, coupled with a unique identifier for the source study. The data object identifier (like study identifiers) therefore needs to be a composite, indicating its type and source as well as its value.

B.2 Other object identifiers

'Other object identifiers' refers to other unique identifiers that have been assigned to the data object in addition to its primary identifier. Again, such IDs would be composite and include the identifier type and assigning organization, as well as its value, and optionally the identifier scheme and date of assignment. The lists used for identifier type and assigning organization could be common with the lists used for study identifiers.

B.3 Object title and B.4 Additional titles

Within the context of the associated study or studies, the default name of the data object should be unique. Additional names can also be provided. If given, they are composite: the title plus one of 'title type' (e.g., translated title, alternative title, subtitle).

B.5 Version

Optionally, the version code for the data object is used. Many versions of a particular data set or document may have been created in the course of a clinical study, but the focus here is on the version or versions that are made available for sharing. The data generators will need to make that selection, though the normal expectation would be that the final version of a data object (e.g., a protocol) would be the one that was shared with others.

In some cases, multiple versions of the same document or data set could be made available, or they might be specifically requested. For instance, data sets used for the primary analysis should normally be available, as well as possible later data sets that have additional follow-up data. A protocol published before the trial began may need to be differentiated from the protocol as it existed at study end. Assuming the data objects have similar names, they will therefore need to be clearly differentiated using version codes (and relevant dates [see D.2 below] and possibly descriptions [see E.3 below]). E.6 describes how the relationship to previous or next versions can be made explicit. The form of the version coding would be as created and applied by the data generators.

C.1 Creators

The creators are the main personnel involved in producing the data, or the authors of a publication. It may be a set of institutional or personal names. Each name in the list is a composite element and can contain optional identifiers (e.g., Open Researcher Contributor Identification [‘ORCID’] identifiers and/or organizational affiliations, as well as the name itself).

C.2 Contributors

Optionally, other institutions and/or persons responsible for collecting, managing, distributing or otherwise contributing to the development of the data object can be included. If given, any contributor record is composite, with the same structure as the Creator data above, plus an additional data point specifying contributor type. The latter may need extending in the context of clinical research to include, for example, drug supplier, drug distributor, device manufacturer, central laboratory, sponsor contact, recruitment contact, principal and chief (or co-ordinating) investigator.

D.1 Creation year

The creation year is the year in which the object was created, expressed as four digits. Its precise definition will vary with the nature of the data object. For data sets, it will be the year of their extraction; for published documents, the year of their initial publication; and for internal documents, the year of their approval for use. Note that ‘creation year’ is intended only to provide an indicator of the time something was created (e.g., in an on-screen listing). It is not a date, which is collected and stored separately (see D.2 below).

D.2 Dates

None, one or more dates or date ranges that are relevant to the data object, in the standard ISO 8601 format (ISO Organization, 2019), are used. Each date should be accompanied by a date type value that indicates what the date represents, such as accepted, available, copyrighted, collected, created, issued, submitted, updated, valid. This list (from DataCite) may need extending to better span the clinical research domain.

E.1 Resource type general

Resource type is one of the existing DataCite controlled list. In most cases, for clinical research data objects, the type will be ‘text’ or ‘data set’.

E.2 Resource type

Resource type is a description of the resource. The format is open, but the preferred format is a single term, so that a pair can be formed with the ‘resource type general’ described above (e.g., data set/census data or text/conference abstract). Existing types will need extending by a list of standard resource types for clinical research (e.g., protocol, patient information sheet, final analysis data set, quality of life data set). In practice, an expandable list would be needed (i.e., one where a user could supplement the supplied controlled vocabulary terms by free text, as and when necessary).

E.3 Description

The description comprises none, one or more pieces of additional general information. The format is open, but any description should be accompanied by a description type to further characterise the data: one of abstract, methods, series information, table of contents, other.

E.4 Subjects

Subjects comprise none, one or more subject names or phrases, keywords, classification codes describing the resource. In general, however, the recommendation is to include any subject or topic descriptors, keywords, and so forth, with the study data rather than the individual data objects (see A.3 above).

E.5 Language

The language is the primary language of the resource, using the International Organisation for Standardisation (ISO) language codes (e.g., EN, DE, FR).

E.6 Related identifiers

Related identifiers are the identifiers of related resources, which must be globally unique identifiers. Related resources will normally be data objects themselves. The record is composite and must include the identifier itself, the related identifier type and the relation type. Relation types include IsCitedBy, Cites, IsSupplementTo, IsSupplementedBy, IsContinuedBy, Continues, IsNewVersionOf, IsPreviousVersionOf, IsPartOf, HasPart, IsIdenticalTo, IsDerivedFrom and IsSourceOf.

A particularly important relationship for clinical study data is the pairing of HasMetadata-IsMetadata. Metadata in clinical research can include, for example, an ODM file or data dictionary that provides the metadata for a data set. The metadata in this context is itself a file, and as a data object in its own right, it is a ‘study data metadata data object’. This is quite distinct from the type of metadata used to describe it and all the other documents and data sets, as a data object, which is ‘data object metadata’.

F. Identifying location, ownership and access

The other area where the existing DataCite schema needs to be extended is in providing a full description of the access arrangements for any data object. The following data points are proposed.

F.1 Publisher

In this schema, this is the organisation that manages access to the document, including making the overall decision about access type (see F.3). For data, this would be the name of the organisation managing the repository. For journal papers, it is the name of the company that publishes the journal and which would normally run the primary website on which it can be accessed.

F.2 Other hosting institutions

Other hosting institutions are any organisations other than the publisher identified in F.1 that also host the data object within their IT infrastructure.

F.3 Access type

Access type is one of ‘public download’, ‘public on-screen access’, ‘restricted download’, ‘restricted on-screen access’, ‘case-by-case download’ or ‘case-by-case on-screen access’. Restricted means access would be dependent on membership of a predefined group, usually as determined by an authentication mechanism (e.g., username with password), such as is the case with subscription to a journal. Case-by-case means that there is no predefined access, but that applications for access to the data object will be considered by the object owners. On-screen access means that a researcher can view and process data within a specified environment but cannot download a file of the raw data, though export of the results of re-analysis would be allowed.

F.4 Access details (mandatory for any of the non-public access types)

Access details refers to a textual description of the access being offered, such as identifying the groups to which access is granted, the criteria on the basis of which a case-by-case decision would be based, or any further restrictions on on-screen access.

F.5 Access contact (mandatory for any of the non-public access types)

Access contact is a link to a resource that explains how access may be gained, such as how a group can be joined, and/or how application can be made for access on an individual basis. This could include an email address but more normally would be a link to a web page on the publisher’s site that would explain access procedures or provide an application pro forma.

F.6 Resources

Resources comprise the web-based resources that represent this data object. This is mandatory for public or restricted access objects when at least one resource should be listed. Each record would be composite and include the F.6.1 resource URL and, if downloadable, the F.6.2 resource file type (e.g., file extension or Multipurpose Internet Mail Extension [‘MIME’] type) and the F.6.3 resource size, usually in kilobytes or megabytes (Canham & Ohmann, 2016). The resource host would usually be obvious from the URL.

F.7 Rights

Rights include any intellectual property rights information for the data object, as a textual statement of the rights management associated with the resource. The URI for the specific rights management should also be given (see original element F.7.1) (Canham & Ohmann, 2016).