Deliverable Number: D7.7, version: 2.0

Data Management Plan - interim version

CAREGIVERSPRO MMD

CAREGIVERSPRO-MMD PROJECT

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

COOSS / Ricerca e formazione

Mobiles Dynamics

FUB UNIVERSITAT A MANRESA

UNIVERSITY OF Hull

CHU Hôpitaux de Rouen

"This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 690211"

Document information

Project Number	690211	**Acronym**	CAREGIVERSPRO-MMD
Full title	Self-management interventions and mutual assistance community services, helping patients with dementia and caregivers connect with others for evaluation, support and inspiration to improve the care experience		
Project coordinator	Universitat Politècnica de Catalunya- BarcelonaTech Prof. Ulises Cortés, ia@cs.upc.edu		
Project URL	http://www.caregiversprommd-project.eu		

Deliverable	**Number**	D7.7	**Title**	Data Management Plan - interim version
Work package	**Number**	WP7	**Title**	Dissemination, Communication, Exploitation and Business Planning

Date of delivery	**Contractual**	31/06/2017	**Actual**	30/04/2016
Nature	Report ☑ Demonstrator ☐ Other ☐			
Dissemination Level	Public ☑ Consortium ☐			
Keywords				

Authors (Partner)	Atia Cortés (UPC), Cristian Barrué (UPC), Ulises Cortés (UPC), Gabriel Verdejo (UPC), Paraskevi Zafeiridi (UHull), Anastasia Matonaki (QPL), Ioannis Pagliokas (CERTH), Isabelle Landrin (CHU), Rafa de Bofarull (MDA)			
Responsible Author	Cristian Barrué		**Email**	cbarrue@cs.upc.edu
	Partner	UPC	**Phone**	+ 34 93 413 40 11

Document Version History

Version	Date	Status	Author	Description
1.0	26-06-2016	Draft	Atia Cortés (UPC)	Integration of Ethics section to previous version (D7.3)
1.1	07-07-2016	Draft	Cristian Barrué (UPC)	Integration of anonymisation section
1.2	15-07-2016	Draft	Cristian Barrué (UPC), Atia Cortés (UPC)	Integration of PIA
1.5	30-11-2016	Draft	Cristian Barrué (UPC), all	Integration of Datasets
1.6	13-01-2017	Draft	UPC	Annex 2: description of datasets Document review
1.7	20-04-2017	Draft	Consortium, Cristian Barrue (UPC), Rafa de Bofarull (MDA)	Update of Annex 2, changes in the datasets
1.8	15-05-2017	Draft	UPC	Integration of new sections for security and data breach procedures
1.9	01-06-2017	Draft	UPC	Update of Annex 2, rework of section 3 accordingly
2.0	26-06-2017	Draft	Ioannis Pagliokas (CERTH), Anastasia Matonaki (QPL), Paraskevi Zafeiridi (UHull), Marco antomarini (COOS), Isabelle Landrin (CHU), Cristian Barrué (UPC)	Review of the document, dataset refinement. Integration.

Executive summary

This is a live document that describes the different processes regarding data management, storage and exploitation that are to be agreed and adopted by every member of the CAREGIVERSPRO-MMD Consortium. Over the course of the project this document will be reviewed and updated. Additional information on the data structure or the methodology, a change in responsibility for a task or in the budget, may be included in future versions of the Data Management Plan. This is the second deliverable of this document, including privacy impact assessment, ethical aspects of the data, data breach protocols, datasets definition at database level and more details on the sharing policies.

List of Acronyms

Acronym	Title
AB	Advisory Board
CERIF	Common European Research Information Format
CERTH	Center for Research and Technology Hellas
CHU	Centre Hospitalier Universitaire de Rouen
CNIL	Commission Nationale de l'Informatique et des Libertés
COOS	Cooperativa Sociale COOSS Marche
C-MMD	CAREGIVERSPRO-MMD
DMP	Data Management Plan
DoA	Description of Action
FUB	Fundació-Universitat del Bages
HONCode	Health On the Net Code
HUL	University of Hull
ICO	Information Commissioners' Office
MDA	Mobile Dynamics
PIA	Privacy Impact Assessment
PLWD	Person Living With Dementia
QA	Quality Assurance
QC	Quality Control
UPC	Universitat Politècnica de Catalunya
VM	Virtual Machine

List of Tables

List of Figures

Table of contents

1 Introduction

This document presents the second version of the Data Management Plan (DMP) for the CAREGIVERSPRO-MMD project. Projects funded by in the Horizon 2020 Open Research Data Pilot are required to develop several versions of a DMP, in which they will specify, among others, what data will be kept for the longer term. In the case of CAREGIVERSPRO-MMD, which is not participating in the Open Research Data Pilot, the DMP is presented as a tool that can improve pilot preparation and result analysis. The Consortium will follow the guidelines described in the OpenAire[1] platform and the document "*Guidelines on Data Management in Horizon 2020*". A DMP describes the data management life cycle for all datasets to be collected, processed or generated by a research project. It must cover:

- the management of research data during & after the project;
- what data will be collected, processed or generated;
- what methodology & standards will be applied;
- whether data will be shared /made open access & how;
- how data will be curated & preserved.

The Data Management Plan has been updated during the project lifetime since version presented in D7.3. New versions of the DMP are also developed whenever significant changes arise in the project (mainly subject to ethical approval) such as:

- new data sets;
- changes in consortium policies;
- external factors.

[1] https://www.openaire.eu/opendatapilot-dmp

2 Project Information

In this section we provide a brief fact sheet of the project details and associated data management requirements

Table 1 Project Fact Sheet

Project Title	CAREGIVERSPRO-MMD
Project Duration	36 months (01/01/16-31/12/18)
Partners	• Universitat Politècnica de Catalunya (UPC, Spain) • Mobile Dynamics (MDA, Spain) • University of Hull (HUL, UK) • Q-PLAN International LTD (QPL, Greece) • Cooperativa Sociale COOSS Marche (COO, Italy) • Fundació-Universitat del Bages (FUB, Spain) • Centre Hospitalier Universitaire de Rouen (CHU, France) • Center for Research and Technology Hellas (CERTH, Greece)
Brief Description	Self-management interventions and mutual assistance community services, helping patients with dementia and caregivers connect with others for evaluation, support and inspiration to improve the care experience
University Requirements for Data Management	UPC is responsible for allocating data in a safe environment, maintaining back-ups and processing the data generated
Funding Body	European Commission (Horizon2020 PHC-25-2105)
Grant Number	690211
Budget	4.087.198,75€
Funding Body Requirements for Data Management	For Open Data projects, the ones specified in Guidelines on Data Management in Horizon 2020[2].

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

3 Data, Materials, Resources Collection Information

The purpose of this section is to provide a full description of the data that will be generated and stored during this project. The information provided below might be adapted or updated in future versions of this document.

3.1 Description of the data

Most of the data will be generated through the use of the CAREGIVERSPRO-MMD online platform by different user groups, *i.e.* health and social professionals, caregivers and people living with dementia (PLWD). Each user category will have access to personalised content and will be able to generate different types of information according to the permissions granted.

For each user of the platform, different datasets described in this section may be generated. Additional datasets may be generated in the future. The data will also be collected before and after the pilot phase of the project at the screening and baseline research visits.

The platform will also provide means to assess and store data not directly produced by users *i.e.* the interaction among users and the evolution on their activity in the social network, which will also be subject to further analysis.

An open source surveying tool LimeSurvey[3] has been configured and deployed to store the results of the screening sessions that clinicians will perform every six months with pilot participants. This tool will only be accessed by authorised health and social professionals.

Personal Dataset

Table 2 Personal Dataset

Data set reference and name
C-MMD-Personal
Data set description
This data set contains all the personal, demographic, medical and social data captured through the registration tools integrated in the C-MMD platform for the dyad (PLWD and caregiver) and the health professionals. The registration tool collects standard personal information. i.e. as described in EU Data Protection Directive (95/46/EC)[4]: *"Personal data" shall mean any information relating to an identified or identifiable natural person ('Data Subject'); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or*

[3] http://www.limesurvey.org

[4] http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:en:HTML

social identity.

Therefore, the nature of the data corresponds to the values used to represent such concepts (*e.g.* text, integers). The specifics of the captured data is described in section 2.3.1 of document D1.3 Screening Strategy.

Details on this dataset can be found in Annex 2.

Standards and metadata

Data will be stored each time a user (be it a PLWD, caregiver or health professional) registers to the platform or modifies his/her profile. It is expected that data will be stored in a MySQL database, using noSQL database for complementary purposes. Records will also be related (and identified) with other datasets and the date when the data was recorded.

Metadata will include information about the profile creation time, range of possible values, etc. This metadata will be associated to each table and will follow the Common European Research Information Format (CERIF) metadata standard[5].

Data sharing

General

This dataset will not be shared outside of the Consortium boundaries for ethical and security reasons. Each dataset record belongs to the user and to the Consortium partner responsible for the user. Only the user, people authorised by him/her (*e.g.* caregiver) and authorised personnel of the Consortium partner responsible for the user, can access the record.

Platform

Data will be available to the user and people authorised by them through the C-MMD platform. A small part of the dataset will be openly accessible by platform users (i.e. name and profile picture or avatar) to enable social networking. Authorised personnel[6] of the pilot partner generating the data will be able to access aggregated data in periodic reports and will be able to access raw data dumped from the database in *csv* files or through a web service. Each access will be identifiable and traceable. The platform software will process the raw dataset to offer the planned services (see D3.1 Detailed system Architecture).

Consortium

Dataset records will be shared among defined Consortium partners anonymised for research purposes to be used for the tasks of the project. Anonymisation is the standard procedure followed to preserve confidentiality of participants.

Each participant (*e.g.* PLWD, caregiver, healthcare professional) will sign an informed

[5] http://www.eurocris.org/cerif/main-features-cerif
[6] Pilot responsible will be in charge of authorising personnel, employees of the Consortium.

consent at recruitment phase authorizing access to all of their data (raw, aggregated, anonymised) in the conditions described hereby. Users will agree to the anonymised and aggregated data being used for research and possibly commercial exploitation. The data repository will be in the C-MMD host in the UPC premises (more details are given in section 6). UPC, CERTH and MDA software components will process this dataset anonymised to offer the platform services. Specific data exchange agreements have been signed between the partners producing software and each pilot.
Archiving and preservation (including storage and backup)
See §7 and §8.

Screening Dataset

Table 3 Screening Dataset

Data set reference and name
C-MMD-Screening
Data set description
This data set contains all the clinical and social data captured through the screening tools integrated in the C-MMD platform for the dyad (PLWD and caregiver). The screening tools implement standard evaluation scales for different conditions (physical, psychosocial, neurological, functional, *etc*.). Therefore, the nature of the data corresponds to the values used to evaluate such scales. Details of the screening strategy can be found in the document D1.3 Screening Strategy, where the list of used scales is detailed Details on this dataset can be found in Annex 2.
Standards and metadata
The data will be stored following the standard numeric scales defined by each screening tool each time that a user (be it PLWD, caregiver or health professional) uses one of the screening tools. The data will be stored in a MySQL database, using noSQL database for complementary purposes. Records will also be related (and identified) with the user to which the recorded data belong and the date when the data was recorded. Metadata will include information about the scale recorded, range of possible

values, etc. This metadata will be associated to each table and will follow the CERIF metadata standard.

Data sharing

General

This dataset will not be shared outside of the Consortium boundaries for ethical and security reasons. Each dataset record belongs to the user and to the Consortium partner responsible for the user. Only the user, people authorised by him/her (*e.g.* caregiver) and authorised personnel of the Consortium partner responsible for the user, can access the record.

Platform

Data will be available to the user and people authorised by them through the C-MMD platform. A small part of the dataset will be openly accessible by platform users (i.e. name and profile picture) to enable social networking. Authorised personnel[7] of the pilot partner generating the data will be able to access aggregated data in periodic reports and will be able to access raw data dumped from the database in *csv* files or through a web service. Each access will be identifiable and traceable. The platform software will process the raw dataset in order to offer the planned services (see D3.1 Detailed system Architecture).

Consortium

Dataset records will be shared among defined Consortium partners anonymised for research purposes to be used for the tasks of the project. Anonymisation is the standard procedure followed to preserve confidentiality of participants.

Each participant (*e.g.* PLWD, caregiver, healthcare professional) will sign an informed consent at recruitment phase authorizing access to all his/her data (raw, aggregated, anonymised) in the conditions described hereby. Users will agree to the anonymised and aggregated data being used for research and possibly commercial exploitation.

The data repository will be in the C-MMD host in the UPC premises (more details are given in section 6). UPC, CERTH and MDA software components will process this dataset anonymised to offer the platform services.

Specific data exchange agreements have been signed between the partners producing software and each pilot.

Archiving and preservation (including storage and backup)

See §7 and §8.

[7] Pilot responsible will be in charge of authorising personnel, employees of the Consortium.

Adverse Event Dataset

Table 4 Adverse Events Dataset

Data set reference and name
C-MMD-Adverse Event
Data set description
This data set contains all the recorded adverse events for each user captured through the specific tool integrated in the C-MMD platform for that purpose. The adverse events dataset records the description of the event, the starting and end dates, the severity and outcomes. Details on this dataset can be found in Annex 2.
Standards and metadata
It is expected that data will be stored in a MySQL database, using noSQL database for complementary purposes. Records will also be related (and identified) with other datasets and the date when the data was recorded. Metadata will include information about the event creation time, range of possible values, etc. This metadata will be associated to each table and will follow the Common European Research Information Format (CERIF) metadata standard[8].
Data sharing
General This dataset will not be shared outside of the Consortium boundaries for ethical and security reasons. Each dataset record belongs to the user and to the Consortium partner responsible for the user. Only the user, people authorised by him/her (*e.g.* caregiver) and authorised personnel of the Consortium partner responsible for the user, can access the record. **Platform** Data will be available to the health care professional and people authorised by them through the C-MMD platform. Authorised personnel[9] of the pilot partner generating the data will be able to access anonymised data in periodic reports and will be able to access raw data dumped from the database in *csv* files or through a web service. Each access will be identifiable and traceable. The platform software will process the raw dataset to offer the planned services (see D3.1 Detailed system Architecture).

[8] http://www.eurocris.org/cerif/main-features-cerif
[9] Pilot responsible will be in charge of authorising personnel, employees of the Consortium.

Consortium
Dataset records will be shared among defined Consortium partners anonymised for research purposes to be used for the tasks of the project. Anonymisation is the standard procedure followed to preserve confidentiality of participants. Each participant (*e.g.* PLWD, caregiver, healthcare professional) will sign an informed consent at recruitment phase authorizing access to all his/her data (raw, aggregated, anonymised) in the conditions described hereby. Users will agree to the anonymised and aggregated data being used for research and possibly commercial exploitation. The data repository will be in the C-MMD host in the UPC premises (more details are given in section 6). UPC, CERTH and MDA software components will process this dataset anonymised in order to offer the platform services. Specific data exchange agreements have been signed between the partners producing software and each pilot.
Archiving and preservation (including storage and backup)
See §7 and §8.

Treatment Dataset

Table 5 Treatment Dataset

Data set reference and name
C-MMD-Treatment
Data set description
This dataset contains all the treatment information for each dyad. The treatment information will come from: (1) a specific toolset integrated in the platform for that purpose, (2) through the API to connect with national healthcare systems where possible. The nature of the data corresponds to medication descriptions, doses, schedules and follow-up of the adherence. More details on treatment adherence service and information to be gathered can be found on sections 10.2 and 10.3 of deliverable D1.1 Accessibility Report. Details on this dataset can be found in Annex 2.
Standards and metadata
The data will be stored following the numeric/text standards each time that a user (be it PLWD, caregiver or health professional) uses the treatment management interface to introduce or modify information about the pharmacological treatment being followed and the adherence regime to the treatment. The data will be stored

in a MySQL database, using noSQL database for complementary purposes. Records will also be related (and identified) with the user to which the recorded data belong and the date when the data was recorded.

Metadata will include information about the data recorded, range of possible values, etc. This metadata will be associated to each table and will follow the CERIF metadata standard.

Data sharing

General

This dataset will not be shared outside of the Consortium boundaries for ethical and security reasons. Each dataset record belongs to the user and to the Consortium partner responsible for the user. Only the user, people authorised by him/her (*e.g.* caregiver) and authorised personnel of the Consortium partner responsible for the user, can access the record.

Platform

Data will be available to the user and people authorised by them through the C-MMD platform. A small part of the dataset will be openly accessible by platform users (i.e. name and profile picture) to enable social networking. Authorised personnel[10] of the pilot partner generating the data will be able to access aggregated data in periodic reports and will be able to access raw data dumped from the database in *csv* files or through a web service. Each access will be identifiable and traceable. The platform software will process the raw dataset to offer the planned services (see D3.1 Detailed system Architecture).

Consortium

Dataset records will be shared among defined Consortium partners anonymised for research purposes to be used for the tasks of the project. Anonymisation is the standard procedure followed to preserve confidentiality of participants.

Each participant (*e.g.* PLWD, caregiver, healthcare professional) will sign an informed consent at recruitment phase authorizing access to all his/her data (raw, aggregated, anonymised) in the conditions described hereby. Users will agree to the anonymised and aggregated data being used for research and possibly commercial exploitation.

The data repository will be in the C-MMD host in the UPC premises (more details are given in section 6). UPC, CERTH and MDA software components will process this dataset anonymised to offer the platform services.

Specific data exchange agreements have been signed between the partners producing software and each pilot.

Archiving and preservation (including storage and backup)

[10] Pilot responsible will be in charge of authorising personnel, employees of the Consortium.

See §7 and §8.

Intervention Dataset

Table 6 Intervention Dataset

Data set reference and name
C-MMD-Intervention
Data set description
This data set contains all the intervention contents created by the consortium members during the lifetime of the project. These intervention contents include posts, articles, tips, multimedia, tutorials, webinars, cognitive games and any kind of educational content produced to support the caregiving process and the healthy ageing lifestyle following the strategy outlined in deliverable D1.3 Intervention Strategy and Contents. These intervention contents will be introduced in the platform through a specific tool designed for that purpose by the consortium. Standards in multimedia and text posts storage will be followed. To be noted that interactive interventions (e.g. Serious Games) are settled in this dataset as well. Those interventions follow a different route to upload their content and finally be available to users: they are authored using software development tools external to the C-MMD platform and finally they will become available through the standard content management tools used for other interventions by providing a link. Details on this dataset can be found in Annex 2.
Standards and metadata

The data will be stored following the standard text/media formats following best practices for data management (see section 6). The data will be stored in a MySQL database, using noSQL database for complementary purposes. Records will also be related (and identified) with the user authoring the contents and the date when the data was recorded. As explained in section 5.1 of DoA and later in this document in section 4, all contents created will follow the HONCode and will provide a traceable review board. If the intervention is not an original creation of the consortium, the original source will be properly cited and referenced. Metadata will include information about the intervention recorded and a list of tags or keywords that relate the content with specific symptoms, conditions or problems that the content refers to (*e.g.* a video about Alzheimer could have the tags *Alzheimer*, *dementia*, *cognitive decline*, etc.) This metadata will be associated to each table and will follow the CERIF metadata standard.
Data sharing
Each dataset record belongs to the Consortium partner responsible for creating it if it is original content. All the Consortium and suitable users[11] are authorised to access the recorded contents. Data will be available to users and people authorised by them through the C-MMD platform. Aggregated data about the amount of contents generated and specific metadata (*e.g.* tags) will be available as well as access to raw data dumped from the database in files to selected Consortium members. Dataset records, particularly aggregated data, will be shared among the Consortium partners for research purposes to be used in the tasks of the project. Original contents may be commercially exploited under internal Consortium agreements to be defined in the future D7.9. Business Plan – final version. The data repository will be allocated in the C-MMD host in the UPC premises (more details in §6). UPC, CERTH and MDA software components will process this dataset anonymised to offer the platform services. Specific data exchange agreements have been signed between the partners producing software and each pilot.
Archiving and preservation (including storage and backup)
See §7 and §8.

[11] In the case of patients or caregivers, contents should be available depending on their specific needs

Dissemination Dataset

Table 7 Dissemination Dataset

Data set reference and name
C-MMD-Dissemination
Data set description
This data set contains all the dissemination contents created by the consortium members during the lifetime of the project. These dissemination contents include scientific papers, newsletters, multimedia, press articles, lists of events, contact lists and any kind of dissemination content produced to support the communication activities of the project and dissemination of results. These contents created from different sources will be stored in a database/filesystem.
Standards and metadata
The data will be stored following the standard text/media formats following best practices for data management (see section 6). Records will also be related (and identified) with the user authoring the contents and the date when the data was recorded. Metadata will include information about the dissemination data recorded, the target audience, identifier (*i.e.* DOI, URI), authors, title of the publication, time of publication, related event (*e.g.* conference, forum, *etc.*) and a list of tags or keywords that relate the content with specific topics or results. This metadata will be associated to each table and will follow the CERIF metadata standard.
Data sharing
Each dataset record belongs to the Consortium partner/s responsible for creating it. These contents are open for access. The data repository will be allocated in the C-MMD host in the UPC premises (more details in §6).
Archiving and preservation (including storage and backup)
See §7 and §8.

User Interaction Dataset

Table 8 User Interaction Dataset

Data set reference and name

C-MMD-User Interaction
Data set description
This data set contains the aggregation of all the content created by the users while interacting with C-MMD's social network during the lifetime of the project. These user generated contents include number of posts, number of comments, numbers of likes, number of scales taken, etc. Details on this dataset can be found in Annex 2.
Standards and metadata
The data will be stored following the standard text/media formats following best practices for data management (see section 6). Records will also be related (and identified) with the user authoring the contents and the date when the data was recorded. Metadata will include information about the content recorded, the target audience (*e.g.* friends), context where it was published, date of publishing, etc. This metadata will be associated to each table and will follow the CERIF metadata standard.
Data sharing
Each dataset record belongs to the user responsible for creating it. These contents are open for access to the audience which the creator user has granted access to, and the members of the consortium for moderation and research purposes. The dataset is anonymised. The data repository will be allocated in the C-MMD host in the UPC premises (more details in §6). UPC, CERTH and MDA software components will process this dataset anonymised to offer the platform services. Specific data exchange agreements have been signed between the partners producing software and each pilot.
Archiving and preservation (including storage and backup)
See §7 and §8.

Medical Report Dataset

Table 9 Medical Report Dataset

Data set reference and name
C-MMD-Medical Report

Data set description
This data set contains all the content created by the system integrating data from the PLWD/caregiver's personal, screening and treatment datasets. These records contain aggregated data of the evolution of the user for health professional evaluation. These contents created from different sources will be stored in a database/filesystem. Details on this dataset can be found in Annex 2.
Standards and metadata
The data will be stored following the standard numeric scales defined by the aggregation of data coming from the screening tools as well as the treatment tool each time that the system periodically generates a report for a PLWD/caregiver. The data will be stored in a MySQL database, using noSQL database for complementary purposes. Records will also be related (and identified) with the user to which the recorded data belong and the date when the data was recorded. Metadata will include information about the scales aggregated, range of possible values, graphical representations, etc. This metadata will be associated to each table and will follow the CERIF metadata standard.
Data sharing
General This dataset will not be shared outside of the Consortium boundaries for ethical and security reasons. Each dataset record belongs to the user and to the Consortium partner responsible for the user. Only the user, people authorised by him/her (*e.g.* caregiver) and authorised personnel of the Consortium partner responsible for the user, can access the record. **Platform** Data will be available to the user and people authorised by them through the C-MMD platform. Authorised personnel[12] of the pilot partner generating the data will be able to access aggregated data in periodic reports and will be able to access raw data dumped from the database in *csv* files or through a web service. Each access will be identifiable and traceable. The platform software will process the raw dataset to offer the planned services (see D3.1 Detailed system Architecture). **Consortium** Dataset records will be shared among defined Consortium partners anonymised for research purposes in order to be used for the tasks of the project. Anonymisation is the standard procedure followed to preserve confidentiality of participants.

[12] Pilot responsible will be in charge of authorising personnel, employees of the Consortium.

Each participant (*e.g.* PLWD, caregiver, healthcare professional) will sign an informed consent at recruitment phase authorizing access to all his/her data (raw, aggregated, anonymised) in the conditions described hereby. Users will agree to the anonymised and aggregated data being used for research and possibly commercial exploitation. The data repository will be allocated in the C-MMD host in the UPC premises (more details in §6). UPC, CERTH and MDA software components will process this dataset anonymised in order to offer the platform services. Specific data exchange agreements have been signed between the partners producing software and each pilot.
Archiving and preservation (including storage and backup)
See §7 and §8.

User Gamification Model Dataset

Table 10 User Gamification Model Dataset

Data set reference and name
C-MMD-User Gamification Model
Data set description
This data set contains all the information related to the gamification model of each user (gamification profile). This profile may contain data records like role in the game, games enrolled, metrics, rules or earned rewards. An initial gamification profile should created -as an extension to the existing user profile- when a user enters in the system and evolves with the participation of the user in the platform. There is an additional short gamification model to be used for limited but fast references to the gamification status of a registered user. Details on this dataset can be found in Annex 2.
Standards and metadata
An initial profile will be stored each time a user (be it PLWD or a caregiver) registers to the platform or modifies their profile settings. This process of user registration and enrolment in one or more gamification proposals ('games') is made programmatically from the social platform using the gamification API. From a technical point of view, gamification can be applied only to registered users. Data will be stored in a MySQL database. Records will also be related (and identified) with other datasets and the date when the data was recorded. The profile and data recorded changes as the user participates in the gamified platform (*e.g.* getting point

and badges when following the wished behaviour, achieving goals, etc.). Metadata will include information about the profile creation time, the set og games a user is enrolled, etc. This model is internal to the gamification engine.
Data sharing
The user gamification model is defined by HCP and administrators and is not accessible openly in the platform to the users. Each dataset record belongs partially to the user and to the platform itself that auto-generates some recorded data. Only the user's HCP, people authorised by him/her, administrators of the platform and authorised personnel of the Consortium partner responsible for the user can access the record. Data will be available through the gamification component of the C-MMD platform. Dataset records will be shared among the Consortium partners anonymised for research purposes to be used in the tasks of the project. Anonymisation is the standard procedure followed to preserve confidentiality of participants. Each participant will sign an informed consent at recruitment phase authorizing access to all his/her data (raw, aggregated, anonymised). Users will agree to the anonymised and aggregated data being used for research and possibly commercial exploitation. The data repository will be allocated in the gamification host in the UPC premises (more details in section 6). CERTH and MDA software components will process this dataset anonymised to offer the platform services.
Archiving and preservation (including storage and backup)
See §7 and §8.

Backend Gamification Model Dataset

Table 11 Backend Gamification Model Dataset

Data set reference and name
C-MMD-Backend Gamification Model
Data set description
This data set contains all the information related to the backend gamification model for each 'game'. This profile may contain data records like game rules, actions related to the social platform, details of the awarding system, etc. An initial profile is created when a game-master (game-creator) enters the gamification front-end and creates a game for a group of CMMD platform users. Details on this dataset can be found in Annex 2.

Standards and metadata
An initial profile will be stored each time a game-master creates a new game. Data will be stored in a MySQL database. The game profile and data recorded changes as the game-creators make changes in the core elements of the game (number of points earned for each user action, details of quests, number and type of rules, etc.). Metadata will include information about the profile creation time, targeted user groups, etc. This model is internal to the gamification engine.
Data sharing
General This dataset will not be shared outside of the Consortium boundaries for ethical and security reasons. Each dataset record belongs to the user and to the Consortium partner responsible for the user. Only the user, people authorised by him/her (*e.g.* caregiver) and authorised personnel of the Consortium partner responsible for the user, can access the record. **Platform** Data will be available to the game-masters (normally one person per pilot site) and people authorised by them through the C-MMD platform. Authorised personnel[13] of the pilot partner generating the data will be able to access aggregated data in periodic reports and will be able to access raw data dumped from the database in *csv* files or through a web service. Each access will be identifiable and traceable. The platform software will process the raw dataset to offer the planned services (see D3.1 Detailed system Architecture). **Consortium** Dataset records will be shared among defined Consortium partners anonymised for research purposes to be used for the tasks of the project. Anonymisation is the standard procedure followed to preserve confidentiality of participants. Each participant (*e.g.* PLWD, caregiver, healthcare professional) will sign an informed consent at recruitment phase authorizing access to all his/her data (raw, aggregated, anonymised) in the conditions described hereby. Users will agree to the anonymised and aggregated data being used for research and possibly commercial exploitation. The data repository will be allocated in the C-MMD host in the UPC premises (more details in §6). CERTH and MDA software components will process this dataset anonymised to offer the platform services. Specific data exchange agreements have been signed between the partners producing software and each pilot.
Archiving and preservation (including storage and backup)
See §7 and §8.

[13] Pilot responsible will be in charge of authorising personnel, employees of the Consortium.

User Interface Dataset

Table 12 User Interface Dataset

Data set reference and name
C-MMD-User Interface
Data set description
This data set contains all the information related to the user interface configuration profile of each user. These profiles may contain data records like complexity degree, colour settings, brightness and alpha settings, etc. An initial profile is created when a user enters in the system and can be either re-configured or enriched by the user or by an administrator. Details on this dataset can be found in Annex 2.
Standards and metadata
An initial profile will be stored each time a user (be it patient, caregiver or health professional) registers to the platform or modifies his/her profile settings. Although at this moment the registering tool and profile management tool have not been defined yet, it is expected that data will be stored in a MySQL database, using noSQL database for complementary purposes. The profile and data recorded changes as the user re-configures it to fit their needs or preferences (*e.g.* changing theme, simplifying the UI, changing colour palette, etc). Metadata will include information about the profile creation time, range of possible values, etc. This metadata will be associated to each table and will follow the CERIF metadata standard.
Data sharing
This dataset is only available to the user owner of the data and the administrators of the platform (i.e. they are requested to modify some parameters of the user UI). No part of the dataset will be shared outside of the Consortium boundaries for ethical and security reasons. Each dataset record belongs to the user and to the platform itself that auto-generates some recorded data. Only the user, people authorised by him/her (*i.e.* caregiver), administrators of the platform and other authorised personnel of the Consortium partner responsible for the user can access the record. Data will be available to users and people authorised by them through the C-MMD platform. Dataset records will be shared among the Consortium partners anonymised for research purposes to be used in the tasks of the project. Anonymisation is the standard procedure followed to preserve confidentiality of participants. Each participant will sign an informed consent at recruitment phase authorizing

access to all his/her data (raw, aggregated, anonymised). Users will agree to the anonymised and aggregated data being used for research and possibly commercial exploitation. The data repository will be allocated in the C-MMD host in the UPC premises (more details in §6). CERTH and MDA software components will process this dataset anonymised to offer the platform services. Specific data exchange agreements have been signed between the partners producing software and each pilot.
Archiving and preservation (including storage and backup)
See §7 and §8.

Recommender Dataset

Table 13 Recommender Dataset

Data set reference and name
C-MMD-Recommender
Data set description
This data set contains all the information related to the recommendation engine profile of each user that provides tailored educational contents to them. These profiles may contain data records such as preferences, past liked contents, content evaluations, visited contents, etc. An initial profile is created when a user enters in the system and evolves with the participation of the user in the platform. Details on this dataset can be found in Annex 2.
Standards and metadata
An initial profile will be stored each time a user (be it PLWD, caregiver or health professional) registers to the platform or modifies their profile settings. Data will be stored in a MySQL database, using noSQL database for complementary purposes. Records will also be related (and identified) with other datasets and ontologies and the date when the data was recorded. The profile and data recorded changes as the user participates in the platform (*e.g.* reading articles, commenting, liking proposed

contents, filling screening tasks that fire recommendations, etc). Metadata will include information about the profile creation time, range of possible values, etc. This metadata will be associated to each table and will follow the Common European Research Information Format (CERIF) metadata standard[14].
Data sharing
This dataset is internal to the recommender system and will not be shared with the users of the platform. Some aggregated information of this dataset may be shared for technical evaluation with the administrators and some will be integrated with the periodic report for the HCP. The whole dataset will not be shared outside of the Consortium boundaries for ethical and security reasons. Each dataset record belongs partially to the user and to the platform itself that auto-generates some recorded data. Dataset records will be shared among the Consortium partners anonymised for research purposes to be used in the tasks of the project. Anonymisation is the standard procedure followed to preserve confidentiality of participants. Each participant will sign an informed consent at recruitment phase authorizing access to all his/her data (raw, aggregated, anonymised). Users will agree to the anonymised and aggregated data being used for research and possibly commercial exploitation. The data repository will be allocated in the C-MMD host in the UPC premises (more details in section 6). UPC software components will process this dataset anonymised to offer the platform services. Specific data exchange agreements have been signed between the partners producing software and each pilot.
Archiving and preservation (including storage and backup)
See §7 and §8.

Pilot Data Dataset

Table 14 Pilot Data Dataset

Data set reference and name
C-MMD-PilotData
Data set description

[14] http://www.eurocris.org/cerif/main-features-cerif

This data set contains all the clinical and social data captured through the pilot tool (see 3.1) that collects the data that clinical practitioners introduce after a screening session with the PLWD/caregiver. The screening data introduced corresponds to the evaluations specified in the study protocol (physical, psychosocial, neurological, functional, *etc*.). Therefore, the nature of the data corresponds to the values used to evaluate such scales.

Details on this dataset can be found in Annex 2.

Standards and metadata

The data will be stored following the standard numeric scales defined by each screening tool each time that an authorised pilot's person introduces data of a user (be it PLWD or caregiver) captured during a face-to-face screening session. The data will be stored in a MySQL database. Records will also be related (and identified) with the user to which the recorded data belong and the date when the data was recorded. The Limesurvey database is not connected in any way with the C-MMD platform database.

Metadata will include information about the scales recorded, range of possible values, identity of the person introducing the data, etc. This metadata will be associated to each table and will follow the Common European Research Information Format (CERIF) metadata standard[15].

Data sharing

This dataset will not be shared outside of the Consortium boundaries for ethical and security reasons. Each dataset record belongs to the user and to the Consortium partner responsible for the user. Only authorised personnel of the Consortium pilot partner responsible for the user can access the record. Authorised personnel of the pilot partner generating the data will be able to access aggregated data in reports and will be able to access raw data dumped from the database in csv files or through a web service. Each access will be identifiable and traceable.

Dataset records will be shared among the Consortium partners anonymised for research purposes to be used in the tasks of the project. Anonymisation is the standard procedure followed to preserve confidentiality of participants.

Each participant will sign an informed consent at recruitment phase authorizing access to all his/her data (raw, aggregated, anonymised). Users will agree to the anonymised and aggregated data being used for research and possibly commercial exploitation.

The data repository will be allocated in the C-MMD host in the UPC premises (more details in section 6). The Limesurvey instance is managed by FUB, the clinical leader of the project.

[15] http://www.eurocris.org/cerif/main-features-cerif

Archiving and preservation (including storage and backup)
See §7 and §8.

Intervention Feedback Dataset

Table 15 Intervention Feedback Dataset

Data set reference and name
C-MMD-Intervention Feedback
Data set description
This data set contains all the information related to the feedback provided by a user to an intervention that has been proposed to him. These feedback references if the intervention has been shared, if has been viewed, time spent consuming it, etc. Details on this dataset can be found in Annex 2.
Standards and metadata
An initial feedback dataset will be created each time an intervention is provided to a user (be it PLWD, caregiver or health professional) and it will update as the user interacts with the intervention through time. The data will be stored in a MySQL database. Records will also be related (and identified) with other datasets and ontologies and the date when the data was recorded. Metadata will include information about the feedback creation time, range of possible values, etc. This metadata will be associated to each table and will follow the Common European Research Information Format (CERIF) metadata standard[16].
Data sharing
This dataset is internal to the platform and the recommender system and will not be shared with the users of the platform. Some aggregated information of this dataset may be shared for technical evaluation with the administrators and some will be integrated with the periodic report for the HCP. The whole dataset will not be shared outside of the Consortium boundaries for ethical and security reasons. Each dataset record belongs partially to the user and to the platform itself that auto-generates some recorded data. Dataset records will be shared among the Consortium partners anonymised for research purposes to be used in the tasks of the project. Anonymisation is the standard procedure followed to preserve confidentiality of participants. Each participant will sign an informed consent at recruitment phase authorizing

[16] http://www.eurocris.org/cerif/main-features-cerif

access to all his/her data (raw, aggregated, anonymised). Users will agree to the anonymised and aggregated data being used for research and possibly commercial exploitation. The data repository will be allocated in the C-MMD host in the UPC premises (more details in section 6). UPC , CERTH and MDA software components will process this dataset anonymised to offer the platform services. Specific data exchange agreements have been signed between the partners producing software and each pilot.
Archiving and preservation (including storage and backup)
See §7 and §8.

User Gamification Interaction History Dataset

Table 16 User Gamification Interaction History Dataset

Data set reference and name
C-MMD- User Gamification Interaction History
Data set description
This data set contains all the interactions made by a user in the gamification context while interacting with the C-MMD platform during the lifetime of the project. These user generated contents include pieces of information related to wished behaviours like type of actions performed, points awarded by these actions, enter/drop/win a game, etc. Details on this dataset can be found in Annex 2.
Standards and metadata
The data will be stored following the standard text/media formats following best practices for data management (see section 6). Records will also be related (and identified) with the user authoring the contents and the date when the data was recorded. Metadata will include information about the datetime an action took place, the data coming with this action and the results generated by this action (as output). This is an internal to the gamification engine data model.
Data sharing
Each dataset record belongs to the user responsible for creating it. These contents

are open for access to the audience which the creator user has granted access to, and the members of the consortium for moderation and research purposes. The dataset is anonymised. The data repository will be allocated in the C-MMD host in the UPC premises (more details in §6). UPC, CERTH and MDA software components will process this dataset anonymised in order to offer the platform services. Specific data exchange agreements have been signed between the partners producing software and each pilot.
Archiving and preservation (including storage and backup)
See §7 and §8.

Game History Dataset

Table 17 Game History Dataset

Data set reference and name
C-MMD- Game History
Data set description
This data set contains all the actions performed in a gamification proposal (‘game’) during the lifetime of the project. The contents of this dataset include information related to the changes performed in the rules of a game, the seasons (time periods the game was active), the time it was created and was set enabled/disabled, datetimes of ‘reset’ actions, etc. Details on this dataset can be found in Annex 2.
Standards and metadata
The data will be stored following the standard text/media formats following best practices for data management (see section 6). Records will also be related (and identified) with the user authoring the contents and the date when the data was recorded. Metadata will include information about the datetime an action took place, the data coming with this action and the results generated by this action (as output). This is an internal to the gamification engine data model.
Data sharing
Each dataset record belongs to the user responsible for creating it. These contents

are open for access to the audience which the creator user has granted access to, and the members of the consortium for moderation and research purposes. The dataset is anonymised. The data repository will be allocated in the C-MMD host in the UPC premises (more details in §6). UPC, CERTH and MDA software components will process this dataset anonymised in order to offer the platform services. Specific data exchange agreements have been signed between the partners producing software and each pilot.
Archiving and preservation (including storage and backup)
See §7 and §8.

Notifications Dataset

Table 18 Notifications Dataset

Data set reference and name
C-MMD- Gamification Notification
Data set description
This data set contains all the information regarding the gamification notification system and will store in the database all the notifications automatically generated during the lifetime of the project. The contents of this dataset include information related to the title and text of the notification, the duration (time period to be in the 'hot topics' list, etc.). Details on this dataset can be found in Annex 2.
Standards and metadata
The data will be stored following the standard text/media formats following best practices for data management (see section 6). Records will also be related (and identified) with the user authoring the contents and the date when the data was recorded. Metadata will include information about the datetime an action took place, the data coming with this action and the results generated by this action (as output). This is an internal to the gamification engine data model.
Data sharing
Each dataset record belongs to the user responsible for creating it. These contents

are open for access to the audience which the creator user has granted access to, and the members of the consortium for moderation and research purposes. The dataset is anonymised. The data repository will be allocated in the C-MMD host in the UPC premises (more details in §6). UPC, CERTH and MDA software components will process this dataset anonymised in order to offer the platform services. Specific data exchange agreements have been signed between the partners producing software and each pilot.
Archiving and preservation (including storage and backup)
See §7 and §8.

Dataset Summary

Table 19 Dataset Summary

Dataset	Who	Ownership	Access[17]
Personal Dataset	User	Yes	Yes, full
	Partner (recruiting)	Yes	Yes, full to authorised personnel
	Others in the platform	No	Only authorised by the user
	Rest of Consortium	No	Yes, only anonymised and aggregated data
	World	No	No
Screening Dataset	User	Yes	Yes, full
	Partner (recruiting)	Yes	Yes, full to authorised personnel
	Others in the platform	No	Only authorised by the user

[17] Users can grant Access to some or all of their profile information to the platform users they like to under their own responsibility. This access can only be granted inside the platform.

	Rest of Consortium	No	Yes, only anonymised and aggregated data
	World	No	No
Adverse Events Dataset	User	Yes	Yes, full
	Partner (recruiting)	Yes	Yes, full to authorised personnel
	Others in the platform	No	Only authorised by the user
	Rest of Consortium	No	Yes, only anonymised and aggregated data
	World	No	No
Treatment Dataset	User	Yes	Yes, full
	Partner (recruiting)	Yes	Yes, full to authorised personnel
	Others in the platform	No	Only authorised by the user
	Rest of Consortium	No	Yes, only anonymised and aggregated data
	World	No	No
Intervention Dataset	User	No	Yes, depending on their needs
	Partner (authoring)	Yes	Yes, full to authorised personnel
	Rest of Consortium	No	Yes, full to authorised personnel
	World	No	Limited and depending on project needs and exploitation policies
User Interaction	User	Yes	Yes, full
	Partner	Yes	Yes, full to authorised

dataset	(recruiting)		personnel
	Others in the platform	No	Only authorised by the user
	Rest of Consortium	No	Yes, only anonymised and aggregated data
	World	No	No
Dissemination Dataset	User	No	Yes
	Partner (authoring)	Yes	Yes
	Rest of Consortium	No	Yes
	World	No	Yes
Medical Report Dataset	User	Yes	Yes
	Partner (dyad manager)	Yes	Yes, full to authorised personnel
	Others in the platform	No	Only authorised by the user
	Rest of Consortium	No	Yes, only anonymised and aggregated data
	World	No	No
User Gamification Model Dataset	User	Yes	No
	Partner (admin)	Yes	Yes, full to authorised personnel
	Others in the platform	No	No
	Rest of Consortium	No	Yes, only anonymised and aggregated data
	World	No	No
Backend	User	Yes	Yes

Gamification Model Dataset	Partner (admin)	Yes	Yes, full to authorised personnel
	Rest of Consortium	No	Yes, only anonymised and aggregated data
	World	No	No
	Others in the platform	No	Only authorised by the user
User Interface Dataset	User	Yes	Yes
	Partner (admin)	No	Yes, full to authorised personnel
	Others in the platform	No	Only authorised by the user
	Rest of Consortium	No	Yes, only anonymised and aggregated data
	World	No	No
Recommender Dataset	User	Yes	No
	Partner (admin)	No	Yes, full to authorised personnel
	Others in the platform	No	No
	Rest of Consortium	No	Yes, only anonymised and aggregated data
	World	No	No
Pilot Data Dataset	User	Yes	No
	Partner (authoring)	Yes	Yes, full to authorised personnel
	Rest of Consortium	No	Yes, only anonymised and aggregated data
	World	No	No

Intervention Feedback Dataset	User	Yes	No
	Partner (authoring)	Yes	Yes, full to authorised personnel
	Rest of Consortium	No	Yes, only anonymised and aggregated data
	World	No	No
User Gamification Interaction History Dataset	User	Yes	No
	Partner (authoring)	Yes	Yes, full to authorised personnel
	Rest of Consortium	No	Yes, only anonymised and aggregated data
	World	No	No
Game History Dataset	Game creator	Yes	Yes
	Partner (authoring)	Yes	Yes, full to authorised personnel
	Rest of Consortium	No	No
	World	No	No
Notifications Dataset	Users	Yes	Yes
	Partner (admin)	Yes	Yes, full to authorised personnel
	Rest of Consortium	No	No
	World	No	No

3.2 Quality Assurance Process

Data collection process is susceptible to contamination in the absence of adequate preventive measures. Data contamination results from a process or phenomenon, other than the one of interest, which can affect the variable values. Data contamination results in erroneous values in the data set. In general, there are two types of errors that can occur in a data set. Firstly, **errors of commission** are the result of incorrect or inaccurate data being included in the data set. This may happen because of a malfunctioning instrument that produces faulty results, data that are mistyped during entry, or other problems.

Errors of omission are the second type of errors. These result from data or metadata being omitted. Situations that result in omission errors occur when data are inadequately documented, when there are human errors during data collection or entry, or when there are anomalies in the field that affect the data.

Quality assurance/quality control (QA/QC) activities should be an integral part of any inventory development processes as they improve transparency, consistency, comparability, completeness and accuracy.

Quality control (QC) is defined as a system of checks to assess and maintain the quality of the data inventory being compiled. Quality control procedures are designed to provide routine technical checks to measure and control the data consistency, integrity, correctness and completeness; and to identify and address errors and omissions. Quality control checks should cover everything from data acquisition and handling, application of approved procedures and methods, and documentation. Examples of general quality control checks include:

- checking for transcription errors in data input, introducing twice each variable;
- checking that scale measures are within the range of acceptable values;
- checking that proper conversion factors are used;
- revisiting introduced data

In future versions of this document we will provide more details on the QC protocols adopted during the project lifetime.

Quality assurance (QA) is a planned system of review procedures conducted outside the actual inventory compilation by personnel not directly involved in the inventory development process. It is a non-biased, independent review of methods and/or data summaries that ensures that the inventory continues to incorporate correctly the scientific knowledge and data generated. Quality assurance procedures may include expert peer reviews of data summaries and audits to assess the quality of the inventory and to identify where improvements could be made. If deemed necessary, selected members of the Advisory Board (AB) may perform this task in the course of the project lifecycle.

4 Privacy and Security of the data

4.1 Infrastructure

The purpose of this subsection if to provide an overview of the resources and security mechanisms involved in the hosting of the IT system for the CAREGIVERSPRO-MMD EU funded project that performs processing of sensitive personal information in the premises of Universitat Politècnica de Catalunya – BarcelonaTech (UPC).

This IT system is hosted in a dedicated hardware purchased specifically for the project:

1 x Dell Networking N4032	(network connectivity)
2 x PowerEdge R320	(hardware redundancy)

The server is located in the UPC campus Data Center (CPD). This data center is a dedicated 250m^2 facility with controlled access, personal ID cards for authorized staff and video surveillance 24x7. The server has dedicated bandwidth and backup power system to guarantee availability. Our Data Center has two different sensor systems to detect and respond electrical or fire problems. The first monitor system is an optical-heat detection scheme and the second one is laser based to extinguish fire using HFC227 gas.

This hardware hosts different software modules developed by three partners of the CAREGIVERSPRO-MMD project, among them UPC.

The company MobilesDynamics develops the core software module that gathers, processes and stores highly sensitive information from users of different EU countries (United Kingdom, Spain, Italy, France). The data includes:

- personal data (e.g. name, age, postal code, etc.)
- medical data (e.g. diseases, comorbities, allergies, etc.)
- treatment data (e.g. medications, schedules, etc.)
- Other non-sensible data

The UPC develops a software module that processes anonymised data that comes from the aforementioned software developed by MobilesDynamics.

The Centre for Research and Technology Hellas (CERTH), develops a software module that processes anonymised data that comes from the aforementioned software developed by MobilesDynamics.

4.2 Adopted Security Measures

Security rules had been adopted at different system levels to provide safeguards to the operating system, web applications, databases and user data.

- Operating System (OS) level: Our Linux systems have an IP firewall configured in order to monitor and grant access only to authorised protocols and verified IP addresses. We provide only Secure Shell (SSH[18]) protocol to make cypher communications mandatory end-to-end.
  DenyHosts service has been configured to detect "brute force" password attacks. When a wrong user log-in password is used three times, the system will block any connection from the source IP address and logs the action.

- Web server application level: We use only the *Hyper Text Transfer Protocol Secure* (*HTTPS*) for user connections to encrypt data interchange between the users and the web application.

- Database level: The database server is isolated from the Internet and is connected only to the application server. Furthermore, the server has its own IP firewall service to allow only connections from the dedicated network between the web server and the database server.

- Physical access level: Our servers are placed in a controlled environment at UPC Data Center. An authorised personal card is required to access to the facilities and a CCTV system provides video surveillance.

We use a Linux based system (Ubuntu 16.04 LTS x64) as an operating system for the applications and database servers. This Unix-like system ensures password encryption[19] services for every system user. Moreover, a strict file and directory permission system[20] is implemented as a system default behaviour to ensure data protection between different system users.

[18] https://en.wikipedia.org/wiki/Secure_Shell

[19] http://www.tldp.org/LDP/lame/LAME/linux-admin-made-easy/shadow-file-formats.html

[20] https://wiki.archlinux.org/index.php/File_permissions_and_attributes

Legal regulations about data privacy (LORTAD) require to identify and keep available for a two year period of time the accounting files of the server's access services. We have identified this required information in the following log files:

/var/log/syslog

/var/log/authlog

/var/log/denyhosts

/var/log/nginx/error.mmd.log

The Security Chief (see 4.3: Overview Of roles) will be notified monthly about the selected log files and the backup procedures.

The system administrator (root) can manage the whole system and change/revoke user's access rights if requested. It also creates and delete system users. A yearly password change policy is adopted for security regulations.

The different services are monitored though a plethora of complementary systems providing alarms and reactive solutions (e.g.):

- Cloud services (virtualization): OpenNebula
- Servers: Pandora FMS
- Data network: Cacti/RRDtool

4.3 Overview of Roles

The machines where software is hosted and data is stored can be accessed by the next stakeholders:

- ***System administrators***: they have full access to the system and their role is to keep the system running, update the operative system and other software stack pieces as required, perform backups, ensure the security of the system.

- ***Security Chief***: audit at least monthly the access control logs, the backup procedures and other security measures and elaborate a report for the Data Officer at UPC Mr. Victor Huerta.

- ***Programmers***: they have partial access to the machines, in repositories where they can upload/download code.

- ***System owner:*** This role is filled by the staff member or management member who has responsibility for the business function performed by the system. In this case the system owner of the software that processes sensible data is MobilesDynamics.

- ***Data owner***: This role is filled by the staff member or management member who has responsibility for the data stored in the system. In this case, it can be one responsible from each CAREGIVERSPRO-MMD pilot site.

- ***End users***: Only the software developed by MobilesDynamics has end users. They access to this website through a specific static URL, introduce their data and receive services.

These stakeholders will be notified with a form about their (different) responsibilities with regards system security as well as the consequences of not being compliant with them.

UPC

The management of the CPD (refrigeration, access control, electrical maintenance and wiring, etc.) is performed by personnel from the UPCnet S.L. company created by UPC for the ICT services management. The CPD is monitored 24/7 by UPCnet personnel.

The server and services management is performed by RDLab personnel, belonging to UPC. Only authorized RDLab System administrators and Security Chief will have physical access to the hardware stored in the data center. RDLab personnel provides services on workdays (9-14h and 15-17h Monday to Thursday and Friday from 9-14h) and intensive timetable on mornings of summer season.

Service requests can be performed 24/7 through the rdlab.cs.upc.edu website or the email address rdlab@cs.upc.edu.

4.4 Information System Architecture and Data

The platform server is executed in a virtual machine (VM1) that has access to the Internet and where end users will connect to the static platform URL **through encrypted https connections**. The databases are stored in a different virtual machine (VM2) that has no access to the Internet, is allocated in a private network which only accessible by VM1. Consequently, data is isolated from the Internet and only the C-MMD platform has secured access to it.

As we can see in blue in Figure 1, only the Personal Dataset (PD) dataset/table contains personal data and is not anonymised because the C-MMD platform needs to access these data. On the other hand, the rest of datasets (in green) related with C-MMD application will be pseudo-anonymised just in case of intrusion into these particular datasets. Keys to re-identification are kept in PD.

Virtual Machines

Recommender system

The recommending system developed by UPC that processes C-MMD data, receives anonymised data from the platform from personal datasets, screening datasets, intervention datasets and stores processed profiles in the recommender dataset. More details on these datasets can be found in the annexed document D7.3 Data Management Plan. The software is hosted in a virtual server running a linux setup as described in section 2. The person responsible for these services is Luis Oliva (loliva@cs.upc.edu).

Gamification and UI personalization system

The gamification server is hosted by UPC and managed by CERTH. It is installed in a virtual machine with Windows 10 (x64) Professional with firewall installed, automatic updates and windows defender active. It has disabled all the windows shared services and can only be accessed remotely via VNC with a specific IP range belonging to CERTH machines. This system provides different services to the c-mmd platform processing anonymous data gathered via API. More details on these datasets can be found in the annexed document D7.3 Data Management Plan. The person responsible for managing this service is Ioannis Paliokas (ipaliokas@iti.gr).

The Pilot Tool

As in the previous case, the Pilot Tool will run in a dedicated virtual server VM3 with access to the internet and data will be stored in a different virtual machine VM4 in a private subnetwork with no internet access where only VM3 can access. In this case, all datasets in PDD (in blue) will be pseudo-anonymised (*i.e.* using keyed-hash functions with stored secret key) with no personal identifications stored. The keys for re-identification will be kept by the CRO of each pilot site following the security protocols stablished by them.

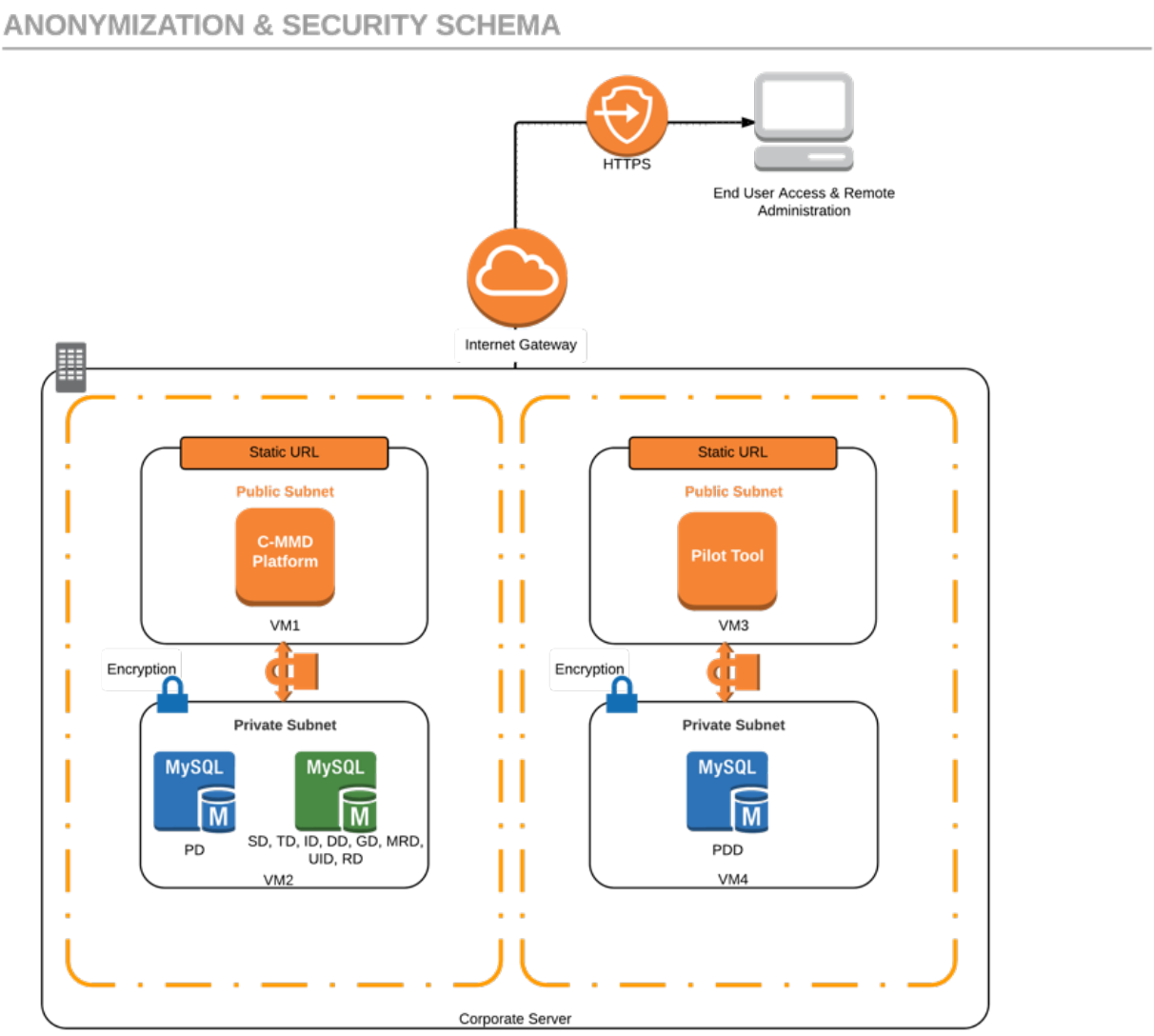

Figure 1 Anonymisation & Security Schema

4.5 Privacy Impact Assessment

The Privacy Impact Assessment is a tool used by organizations aiming to identify possible risks during the processing of personal data and to minimize its impact. According to the Article 35 of the 2016/679 Regulation, *"where a type of processing in particular using new technologies, and taking into account the nature, scope, context and purposes of the processing, is likely to result in a high risk to the rights and freedoms of natural persons, the controller shall, prior to the processing, carry out an assessment of the impact of the envisaged processing operations on the protection of personal data. A single assessment may address a set of similar processing operations that present similar high risks."* Privacy

impact assessment will help the C-MMD Consortium to identify and reduce the privacy risks of the project while allowing the goals of the project to be achieved[21]. This PIA will be used during the development and implementation of the project through its management processes.

PIAs are an integral part of the privacy *by design* approach, which is also one of the principles of the Data Privacy Directive[22] and it is recommended to implement them from the early phase of a project. PIAs aim to promote good practices for personal data processing, improve organizations' transparency and increase the public's understanding of how their information is used.

According to the ICO PIAs code of practice[21], a PIA should incorporate the following steps:

- Identify the need for a PIA
- Describe the information flow
- Identify the privacy and related risks
- Identify and evaluate the privacy solutions
- Sign off and record the PIA outcomes
- Integrate the outcomes into the project plan
- Consult with internal and external stakeholders as needed throughout the process

One of the most critical points is the identification of privacy risks. The ICO has identified three main categories of possible risks: (i) risks to **individuals**; (2) **corporate** risks; and (3) **compliance** risks.

The CNIL[23] provides a method for estimating the risk level in terms of **severity** (or magnitude of the risk) and **likelihood** (or the possibility for a risk to occur). Moreover, they propose a cyclic approach, where the PIA is evaluated after a 4-step process until the PIA is accepted. The approach is structured as follows:

1. **Context**: presentation of the project and its objectives, stakeholders, processing of personal data
2. **Controls**: description of the legal control following the Data Privacy Directive and the risk management plan defined by the organizations involved in the project
3. **Risks**: detailed description of the potential risks and threats that may occur during the project and determine the risk level.
4. **Decision**: to validate the PIA according to the preceding steps and, if accepted, prepare and action plan for all the planned controls.

The C-MMD PIA is fully described in Annex 1.

[21] Conducting privacy impact assessments code of practice 20140225 (Information Commissioners' Office) https://ico.org.uk/media/for-organisations/documents/1595/pia-code-of-practice.pdf

[22] Directive 95/46/EC and Regulation 2016/679

[23] Comission National de l'Informatique et des libertés: https://www.cnil.fr/fr/node/15798

4.6 Security/Data Breach Management

Organisations which process personal data must take appropriate measures against unauthorised or unlawful processing and against accidental loss, destruction of or damage to personal data[24]. One of those measures can be the design and adoption of a protocol to deal with a data security breach.

A data security breach can happen for several reasons, e.g.:

- Loss or theft of data or equipment on which data is stored
- Inappropriate access controls allowing unauthorised use
- Equipment failure
- Human error
- Unforeseen circumstances such as a fire or flood
- Hacking attack
- Social engineering where information is obtained by deceiving the organisation who holds it.

The C-MMD consortium will follow the data breach management protocol described in Annex 3.

4.7 Anonymisation

Directive 95/46/EC on the protection of individuals with regard to the processing of personal data and on the free movement of such data defines "*a personal data shall mean any information relating to an identified or identifiable natural person ("data subject"); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity;*"

Anonymisation and pseudonymization concepts rise from the need to protect the privacy of the data subjects while their data is being processed or transferred by telecommunication networks. Anonymisation is the process of turning data into a form that does not identify individuals and where identification is not likely to take place. This allows for a much wider use of the information[25]. For instance, in the C-MMD context where different researchers must process pilot participant's data in order to extract scientific conclusions, anonymisation is required in order to ensure data privacy of the pilot users. Among the personal data, we distinguish Personally Identifiable Information (PII), information that can be used on its own or with other information to identify, contact, or locate a single person, or to identify an individual in context[26].

[24] https://ico.org.uk/media/for-organisations/documents/1562/guidance_on_data_security_breach_management.pdf
[25] https://ico.org.uk/for-organisations/guide-to-data-protection/anonymisation/
[26] https://en.wikipedia.org/wiki/Personally_identifiable_information

In order to be able to process and research with collected personal data, the so-called PII is either being deleted (anonymisation) or replaced by neutral identifiers (pseudonymization).

According to ICO's 'Anonymisation: managing data protection risk'[27], there are many advantages of using anonymisation:

- it protects against inappropriate disclosure of personal data; fewer legal restrictions apply.
- it can be easier to use anonymised data in new and different ways because the purpose limitation rules do not apply;
- it allows organisations to make information public while still complying with their data protection obligations; and
- the disclosure of anonymised data is not a disclosure of personal data – even where the data controller holds the key to allow re-identification to take place.

It must be kept in mind that in the last years, technology has made it possible to re-identify anonymised data, matching anonymised data back with individual persons whose data was extracted from or even when the original data did not belong to the dataset as such:

- Sweeney[28] demonstrated that 87% of all Americans could be uniquely identified using three pieces of data: birthdate, sex and ZIP code.
- Horvát et al[29] published a study on social networks proving that "using machine learning one can reach a 85% prediction rate whether two non-members known by the same member of the social network are connected or not. Thus showing that the seemingly innocuous combination of knowledge of confirmed contacts between members on the one hand and their email contacts to non-members on the other hand provides enough information to deduce a substantial proportion of relationships between non-members."

Consequently, the project consortium will carefully consider re-identification risks in the PIA and implement the anonymisation procedures accordingly, performing re-identification tests.

Anonymisation Implementation in C-MMD

In C-MMD case, we are handling an open social network that manages several datasets with different access levels. From the application point of view, data will not be anonymised as it would oppose the social network philosophy and where access control is managed by its own users. Thus, C-MMD will provide all end users the possibility to manage the access level of all the generated data related to them, be it personal data, questionnaires completed, information about treatment, etc. More details on default settings can be found in Annex 1, where data flow is described.

[27] https://ico.org.uk/media/for-organisations/documents/1042731/anonymisation_code_summary.pdf
[28] Nate Anderson. Anonymised data really isn't—and here's why not. September 2009. Available at http://arstechnica.com/tech-policy/2009/09/your-secrets-live-online-in-databases-of-ruin/
[29] Emöke-Ágnes Horvát, Michael Hanselmann, Fred A. Hamprecht, Katharina A. Zweig. One Plus One Makes Three (for Social Networks). April 2012. Available at http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0034740#pone.0034740-Jernigan1

When we leave behind the application level and we concentrate in the low level (database, table, server, etc.), the proposed configuration can be seen in Figure 1, where we can see a clear separation between C-MMD platform and the Pilot Tool[30] in terms of execution and data storage.

Data Dissemination

Data that has to be shared outside of the Consortium boundaries for scientific dissemination (*e.g.* aggregated data in a scientific paper) will require stronger anonymisation measures than pseudo-anonymisation, as it could be the case that C-MMD disseminated data can be crossed with other existing records and user identity could be deduced. In order to avoid this, we will follow anonymisation recommendations of Article 29 Data Protection Working Party[31], combining several strategies in order to avoid the weaknesses of each approach (i.e. randomization and generalization techniques). Data dissemination will not happen until the end of the pilot experimentation, so the Consortium has time to prepare detailed and tailored anonymisation strategies to ensure that re-identification is not possible. Specific technical details of these measures will be detailed in future versions of this document. The objectives of anonymising data to be openly disseminated will be to avoid:

- *Singling out*, the possibility to isolate some or all records which identify an individual dataset
- *Linkability*, the possibility to link, at least, two records concerning the same data subject or a group of data subjects

Inference, the possibility to deduce with significant probability the value of an attribute from the values of a set of other attributes

5 Ethics, Intellectual Property, Citation

5.1 Ethics

The lack of ethical principles standardization at international level may potentially lead to the abuse of data collection, use and storage by exploiting differences between societies with regard to established ethical standards. Ethics of data collection, and data use and storage in medical applications, is of growing importance since the quality and quantity of medical data usage is growing quickly both in Europe and worldwide. Great concerns are raised about data protection and privacy issues in the area of biometric and health applications with growing markets that might be affected by insufficiently protected sensitive information.

The success of the C-MMD project depends greatly on having all project partners being aware of the ethical challenges involved in the inception and implementation of the

30 see 3.1

31 Opinion 05/2014 on anonymisation Techniques. http://ec.europa.eu/justice/data-protection/index_en.htm

proposed platform and services. The C-MMD Consortium will respect the ethical guidelines described in:

- Charter of Fundamental Rights of the European Union (2012/C 326/02);
- Convention for the Protection of Human Rights and Fundamental Freedoms;
- World Medical Association Declaration of Helsinki. Ethical Principles for Medical Research Involving Human Subjects;
- Directive 2001/20/EC of the European Parliament and of the Council on the approximation of the laws, regulations and administrative provisions of the Member States relating to the implementation of good clinical practice in the conduct of clinical trials on medicinal products for human use;
- Convention for the Protection of Human Rights and Dignity of the Human Being with regard to the Application of Biology and Medicine: Convention on Human Rights and Biomedicine.

Hereby we describe a summary of the fundamental concepts relevant for C-MMD:

On data privacy

- Clinical trials with human subjects must ensure methods of protecting the individual's dignity and identity.
- The protocol presented to the Ethic Committee shall include the process of anonymisation of personal data for later analysis.

On the right of being informed

- The persons undergoing research have been informed of their rights and the safeguards prescribed by the law for their protection
- Human subjects have the right to be informed about the outcomes of the clinical trial.

On informed consents

- People participating in clinical trials have the right to be informed about the risks and benefits of the study
- People shall not participate in a clinical study without signing a formal consent
- People have the right to withdraw the clinical trial once it has commenced.
- In case of incapacitated adults or people with dementia, a legal representative must consent to participate.

On the access to health care

- The main priority for clinicians shall be to guarantee their clients wellbeing and intervene in case of risk of adverse effects.
- An individual has the right to continue with its regular medical treatment even if they have withdrawn consent to participate in the study.

A full review on the above-mentioned list is provided in **D8.3 Report on Legal and Regulatory Framework**.

5.2 Intellectual Property

Regarding property and ownership of medical data and records, there are two distinct views. From the standpoint of practitioners (*i.e.*, healthcare providers, hospitals), patient medical records are the practitioner's property because they are the ones who write, compile and produce the records (data producers). At the same time, patients tend to believe that medical records belong to them as they provide the relevant information.

Nevertheless, the project will produce data assets that do not correspond to medical records. For instance:

- Intervention contents and guidelines;
- Gamification reports;
- Treatment adherence reports;
- Aggregated medical data reports; and
- Reports and statistics of platform usage.

The resulting ownership agreements will be compliant with corresponding legislation (i.e. Data Protection Act, Copyright, Freedom of Information Act, *etc*.).

5.3 Citation

An article, paper or presentation that refers to, or draws, information from a data set should cite the data set, just as it would cite other sources such as books and articles. A citation gives appropriate credit to the data set creator(s), and allows interested readers to find the data set so they can confirm the data is being correctly represented, or can use it in their own work. There is no universal standard for formatting a data set citation.

There are many different styles for formatting citations, such as APA and Chicago Manual of Style. In addition, most scientific publications have their own style, either unique to themselves or based on an existing style. A few of these styles, such as APA 6th edition, specify how to cite data sets. However, most citation style manuals do not currently cover citing data sets. Consequently, adaptation of the styles' general format can be applied to the needs of data sets.

At this early stage, the information used to cite C-MMD data sets could be:

- Author(s) (*the principal investigator can be used as the "author" of a data set*)
- Title
- Year of Publication
- Publisher (*partner producing the dataset*)

- Version
- Access information (*doi or url*)

6 Access and Use of Information

One of the objectives of the CAREGIVERSPRO-MMD project is to develop the solution into a commercial product. This is the main reason why the Consortium has decided that selected, potentially publishable summary data will not be available for open access until the end of the project, once the exploitation paths have been defined.

However, results of the pilot execution and platform evaluation will be made publicly available through the deliverables **D6.1 – Mid-Pilot preliminary analysis report, D6.2 – Final Pilot analysis report** and **D6.3 – User feedback and usability report.**

More details on specific dataset access regimes are defined in § 3.1.

7 Storage, Backups and Data Recovery

In order to safeguard the appropriate preservation of the data, a portion of the budget has been allocated to data storage and backups during the lifespan of the project and at least for the two years[32] following the grant duration.

The data will be stored in databases installed on the same server that holds the CAREGIVERSPRO-MMD platform. These Databases are only accessible locally (i.e. only available to the server itself) to prevent any connection from outside. The system and server configuration have been arranged to support local data encryption to avoid physical access to the hard disk drive. This measure would prevent access to the data if the physical storage was stolen or accessed directly.

The server has a local firewall that only allows secure web connections to the Internet and verified IP addresses for development/updates of the C-MMD application. Every access to the server is recorded at a local log file.

A daily backup procedure has been designed to ensure data integrity and recovery. This backup has two main subsystems:

1. File system backup: A daily copy of every file in the file system is stored in compressed format.
2. Database backup: A daily dump of every database/table is stored in a single file in the same server.

[32] In the case that any of the Consortium pilot members require to keep the data stored for a longer period, UPC will transfer securely the corresponding data to the partner so it can continue the curation.

The backup system that stores and holds a daily copy of file system and databases is an autonomous system to ensure security integrity of data. On one hand provides an independent service with disk redundancy (RAID[33]), thus is not affected for any failure or malfunction of the main storage system. On the other hand, it controls and restrict access to proper IT staff because it has a specific rights management.

A 25-30-day window backup system has been programmed and enough disk space has been reserved for a monthly operation. The access to the backup data is only available for UPC system administrators and is also logged. The recovery time depends on the amount of data to be recovered, in the frame of workday, can be from 1 to 4 hours.

Our Data Centre management staff provide a pay-per-use service for external backup placement if needed for legal regulations, privacy or security issues.

8 Archiving and Future Proofing of Information

The national legislation (European compliant) of the server site (Spain) compels UPC to preserve all data and access records for **two years**[34] after the project completion. The server will remain in the same safe location to preserve physical and logical access. Consequently, the data will be kept in the server and will be accessible under the same terms that will be agreed among partners during the project lifespan.

All public project deliverables will be available at least for **five years** after the project completion at the project portal.

Selected datasets, databases, standalone documents, and even software may be made public or open for exploitation at the end of the project if that fact is compliant with ethics and data protection guidelines described in this document. These resources may prove useless without explanatory notes (metadata) accompanying them. Metadata will be clearly linked to the materials so that they can adequately inform any future user about the material. For example, a published dataset will typically be accompanied by a metadata document that explains the various fields, their usefulness and summarises the purpose of the dataset in general. These documents will be stored along with the dataset and made accessible in the same manner as the dataset (*e.g.* online, or download). Contact information will be provided accordingly in case that the future user needs further clarification.

[33] https://en.wikipedia.org/wiki/RAID

[34] In the case that any of the Consortium pilot members require to keep the data stored for a longer period, UPC will transfer securely the corresponding data to the partner so it can continue the curation.

8.1 Best Practices for File Formats

The file formats used have a direct impact on the ability to open those files later and on the ability of other people to access those data.

Proprietary vs Open Formats

Data should be saved in a non-proprietary (open) file format when possible. If conversion to an open data format will result in some data loss from the files, it should be considered saving the data in both the proprietary format and an open format. Having at least some of the information available in the future is better than having none.

When it is necessary to save files in a proprietary format, it will be included a readme.txt file that documents the name and version of the software used to generate the file, as well as the company who made the software.

Guidelines for Choosing Formats

When selecting file formats for archiving, the formats should ideally be:

- Non-proprietary;
- Unencrypted;[35]
- Uncompressed;
- In common usage by the research community;
- Adherent to an open, documented standard:
    - Interoperable among diverse platforms and applications
    - Fully published and available royalty-free
    - Fully and independently implementable by multiple software providers on multiple platforms without any intellectual property restrictions for necessary technology
    - Developed and maintained by an open standards organization with a well-defined inclusive process for evolution of the standard

Some Preferred File Formats[36][37]

- Containers: TAR, GZIP, ZIP
- Databases: XML, CSV
- Geospatial: SHP, DBF, GeoTIFF, NetCDF
- Moving images: MOV, MPEG, AVI, MXF
- Sounds: WAVE, AIFF, MP3, MXF
- Statistics: ASCII, DTA, POR, SAS, SAV
- Still images: TIFF, JPEG 2000, PDF, PNG, GIF, BMP
- Tabular data: CSV
- Text: XML, PDF/A, HTML, ASCII, UTF-8
- Web archive: WARC

[35] Data will be encrypted in the UPC server for security reasons

[36] http://www.digitalpreservation.gov/formats/

[37] http://www.loc.gov/preservation/resources/rfs/data.html

9 Audits

An internal audit procedure will be performed at least every 18 months to verify that all the aforementioned measures are implemented and are compliant with regulations.

10 Resourcing of Data Management

This section outlines the staffing and financial details of the data management within the CAREGIVERSPRO-MMD project. The former aspect provides information about the role and responsibilities of the partners that generate the data and those who control it. The latter aspect describes the financing process for data management and data storage.

10.1 Roles in Data Management

Each pilot partner (HUL, COO, FUB, CHU) is responsible for the data generated in their own pilots by the different stakeholders of the platform as **data producers**. Each pilot partner will assign a responsible person from his or her institution for this task to be designed for the next version of this document.

The UPC is responsible for all the aspects related with data storage and backup as **data processor**.

MDA and CERTH as the main developers of the C-MMD platform will be responsible as **data processor** and **service provider** of all the aspects related with data gathering, data integrity, access logging, *etc.* , related to the software components they develop.

As specified in §5.1.3 of DoA, specific agreements will be signed among partners in order to grant access to the different datasets for the different uses (data storage, data processing, service provision).

10.2 Financial Data Management Process

As mentioned before, the Consortium has reserved a portion of the project budget for data hosting and backup.

11 Review of Data Management Process

The follow-up of this plan will be reported in future versions of this document, where detailed protocols and measures will be described to ensure the compliance with the plan along with preliminary results on the observed evolution. UPC as main contributor to this plan, supported by the roles described in section 8.1, will perform the follow-up.

External reviewers of the Consortium as well as selected members of the AB will support the peer-review process.

Annex 1 - C-MMD Privacy Impact Analysis

A1.1 Identifying the need for a PIA

CAREGIVERSPRO-MMD project aims to build a digital platform tailored for people living with dementia and their caregivers, considering this "*dyad*" as the unit of care. The platform will offer both a variety of advanced, personalized services that will improve the quality of their lives and enable them to live well in the community for as long as possible.

The services offered by the CAREGIVERSPRO-MMD platform will be based on the collection and analysis of data provided by the end-users or generated through their interaction with the system. Accessible through user-friendly and easy-to-use interfaces for smartphones, tablets and web browsers, these services will include:

- clinical and psychological screening;
- treatment adherence services;
- educational interventions tailored to each user's symptoms;
- social networking with other people living with dementia, caregivers and clinicians;
- a service for reporting to doctors and medical staff about treatment adherence levels and other important clinical information.

In all, the C-MMD platform is an IT system that will be storing sensible personal and health data of its users. Some of this data will be shared between some users (i.e. PLWD-caregiver or PLWD-doctor). The bulk of the captured data will also be used for research studies and some part of this data might be shared with the research community and society.

Thus, we have a system that will be managing and storing very sensible information subject to ethical and legal considerations and potentially subject to privacy risks. Consequently, the Consortium has undergone the task of writing a PIA that depicts possible risks and mitigation measures so it can be used through the development and implementation of the C-MMD project.

A1.2 Describing information flows

A1.2.1 Information Overview

In previous sections we have described the different datasets that compose the C-MMD ecosystem and in order to perform a more detailed PIA, we provide a description of how the information that enriches C-MMD is obtained, used and retained. In **Error! Reference source not found.** we can see a schema of the different stakeholders interacting with the C-MMD platform and the Pilot Tool, which datasets are they generating and finally the access role of each stakeholder on those datasets.

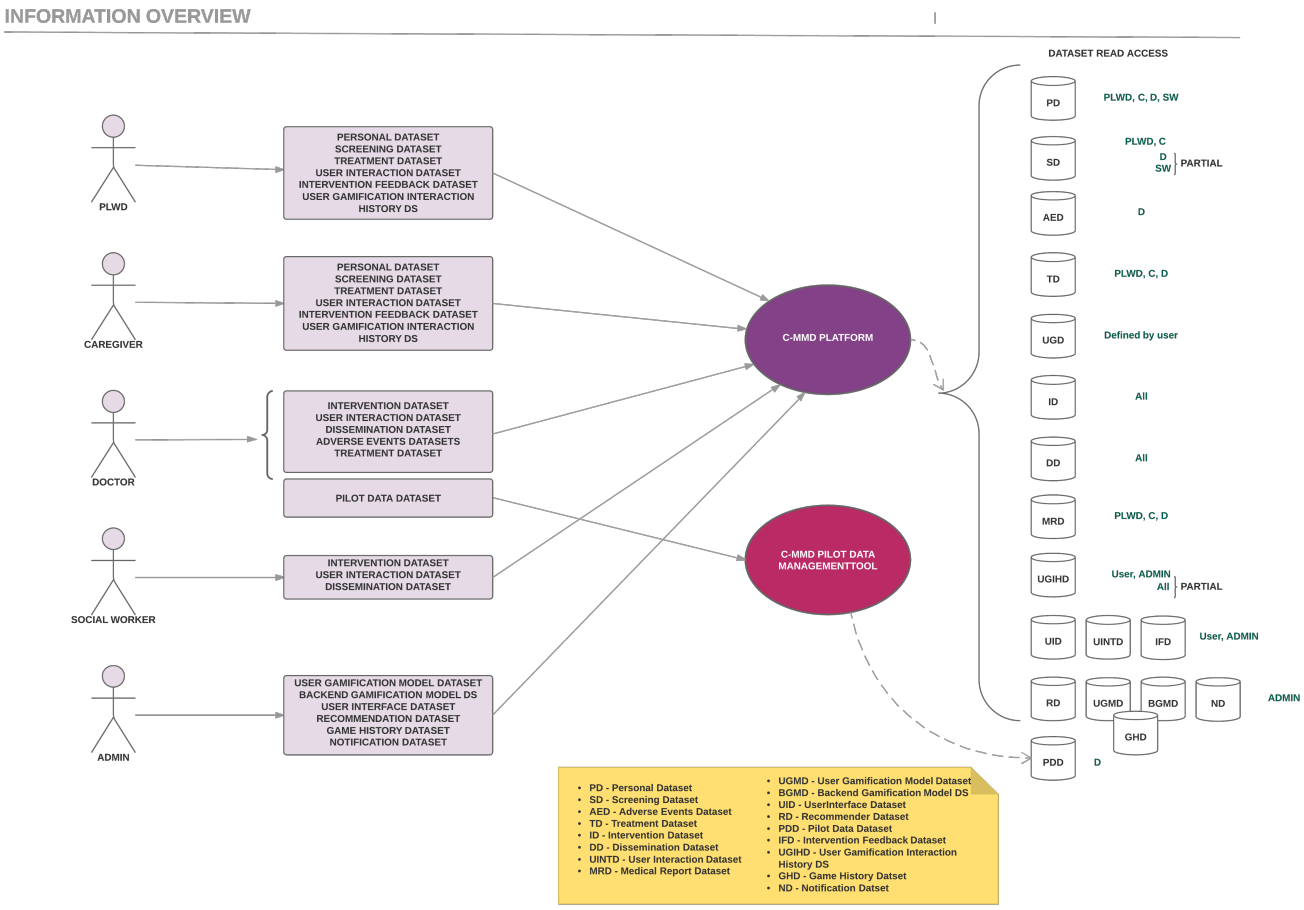

Figure 2 C-MMD Information Overview

We can observe that the main stakeholders interacting and introducing data in the C-MMD platform are PLWD, their caregivers, health professionals and social workers. It is assumed that as in any IT system, administrators have access to the system and may set up initial configurations for some services. In section 3.1 we can find a description of the different datasets and a broad description of ownership and access to data depending if it is a user (PLWD-caregiver), a Consortium partner co-owning the data (health or social professional, or researcher working with the users in each pilot site), a member of the rest of the Consortium or an external person of the project. In Figure 2 we can see which stakeholder produces each dataset and which stakeholder can access each of them. Where is stated 'Partial' access, means that not all the dataset is available (i.e. a social worker will only be able to access screening information related to social aspects, but not clinical). In Figure 3 we can see the data flow diagram of C-MMD, where stakeholders (in red) interact with the platform, introduce data that composes datasets (boxes in light red) and fire processes (purple circles) that can generate new datasets, logs or IT data repositories.

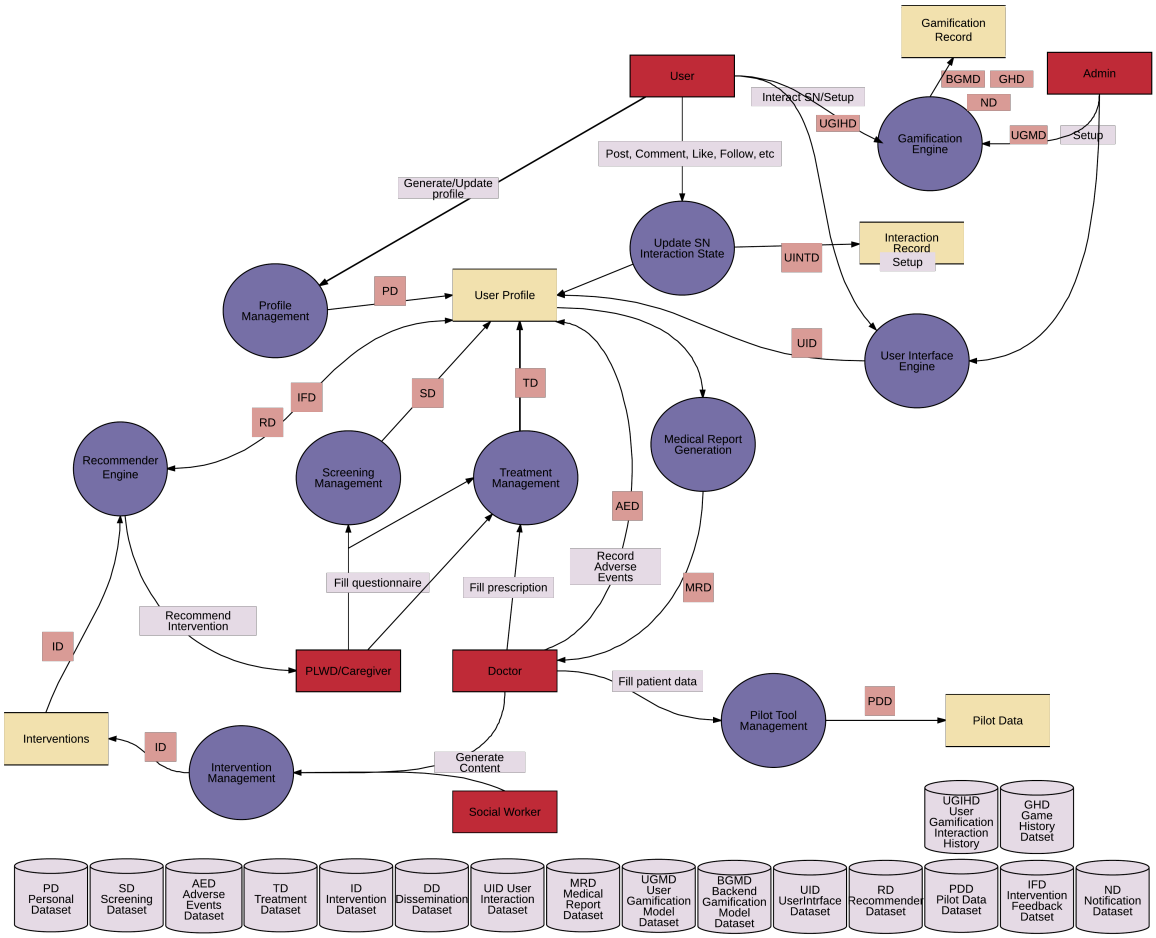

Figure 3 C-MMD Data Flow diagram

A1.3 Identifying privacy and related risks

At this stage, we have described the kind of information exchanged among stakeholders and its relevance. Within this scenario, we must identify which threats are more likely to occur as well as their possible source and severity. Since each organization will evaluate the risks and the possible solutions according to their own characteristics and resources the outcome of this study is necessarily subjective.

Severity[38] represents the magnitude of a risk. It is primarily estimated in terms of the extent of potential impacts on data subjects, taking account of existing, planned or additional controls. The severity level obtained may be raised or lowered in relation to (1) the level of identification of personal data; (2) the nature of the risk source; (3) the number of

[38] Definition from CNIL

interconnections (especially with foreign sites); (4) the number of recipients (which facilitates the correlation between originally separated personal data).

Likelihood[39] represents the feasibility of a risk to occur and is estimated in terms of vulnerabilities of the supporting assets involved and the capabilities of the risk source to exploit them. The justification of the likelihood is provided by the proposed control solutions presented in A1.4.

According the guidelines provided by CNIL, there are three potential data breaches: (i) illegitimate access to personal data, (ii) unwanted modification of personal data and (iii) disappearance of personal data. The following table describes the possible threats that might be present in the C-MMD project categorized within the aforementioned data breaches.

[39] Definition from CNIL

DATA BREACH	#	TYPE OF SUPORTING ASSET	THREATS	ACTION	RISK SOURCES	SEVERITY	LIKELIHOOD	JUSTIFICATION (of likelihood)
Illegitimate access to personal data	1	HW in pilot site	Uncontrolled access to data by unauthorized personal	Observed	Internal/external human	Limited	Depends on each pilot	Negligible if personal data is stored in a room protected by access code. Maximum in case of being in unprotected rooms or in public areas
	2	HW in pilot site	Use of USB flash drives or disks that are ill-suited to the sensitivity of the information	Used inappropriately	Internal human	Negligible	Negligible	No data export means outside the protected DBs or secured protocols will be allowed
	3	HW at UPC server	Lost of equipment (theft)	Lost	Internal/external human	Limited	Negligible	Protected room with controlled access. Protected server with physical lock. Back up and encryption strategies for data protection
	4	SW at UPC server	Infection by malicious code, hacking DB	Altered	Internal/external human or non-human source	Maximum	Negligible	Protected access
	5	SW at pilot sites	Infection by malicious code, hacking DB	Altered	Internal/external human or non-human source	Significant	Depends on each pilot	Security policy of each pilot site will determine the likelihood
	6	SW pilots & server	Errors during updates, configuration or maintenance; replacement of components	Altered	Internal/external human or non-human source	Limited	Limited	Access restricted to authorized trained staff

"This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 690211"

	7	People in pilots, UPC server controllers and technical developers	Abuse of use of personal data for other purposes (transactional, navigation or geo-localization data; behaviour monitoring, profiling and decision making)	Manipulated	Internal human	Significant	Negligible	Previous training on ethical and regulatory dispositions makes unlikely the use of information for unplanned and illegitimate purposes
	8	People in pilots, UPC server controllers and technical developers	Breach of confidentiality of personal data by employees of the organization	Observed	Internal human	Significant	Negligible	Previous training on ethical and regulatory dispositions makes unlikely the disclosure of personal information
	9	People in pilots, UPC server controllers and technical developers	Unauthorized access to personal data	Used inappropriately	Internal human	Limited	Negligible	Access restricted to authorized trained staff
	10	People in pilot site	Obtaining an informed consent doubtful, corrupt or invalid for the treatment of transfer of personal data; hinder withdrawal of consent or opposition to treatment or disposal	Used inappropriately	Internal human	Limited	Negligible	Clinical partners are committed to the protocol presented and approved by an Ethical Committee following the Fundamental Human Rights
Unwanted modification of personal data	11	HW at UPC server	Uncontrolled access to data, no back-up strategy, abuse of storage period	Used inappropriately	Internal/external human	Limited	Negligible	Servers stored in a room with controlled access. Backup protocol
	12	Computer channels at UPC server	Man-in-the-middle attack or other attacks	Used inappropriately	Non-human source	Significant	Negligible	All communication channels are encrypted

	13	SW	Unwanted modifications on DB, erasure of files	Used inappropriately	Internal/external human	Limited	Limited	Backup protocol
Disappearance of personal data	14	SW pilots & server	Exceeding DB size, injection of data outside the normal range of values, denial of service attack	Overload	Internal human or non-human source	Negligible	Negligible	All DB modifications are checked before commit
	15	HW pilots & server	Storage unit full, power outage, processing capacity overload, overheating, denial of service attack	Overload	Internal human	Negligible	Negligible	Server allocated in data centre that covers all these cases
	16	Paper documents in pilot sites	Replacement of original files (falsifications)	Altered	Internal/external human, non-human source	Significant	Limited	Protected rooms with badge reader and/or access code.
	17	Paper documents	Theft of files from offices, mail from mailboxes or retrieval of discarded documents	Lost	Internal/external human, non-human source	Significant	Negligible	Protected rooms with badge reader and/or access code.
	18	HW and SW	Technology obsolescence	Overload	Non-human source	Negligible	Negligible	Tools selected to be valid during the project lifetime

A1.4 Identifying and evaluating privacy solutions

RISK	CONTROLS	RESULT
1	1. Avoid public locations inside hospitals (*e.g.* corridors, shared working spaces)	Reduced
2	1. Controlled access to data 2. Activity log	Eliminated
3	1. Restricted and controlled physical access 2. Encryption 3. External back up 4. Physical HW protection	Eliminated
4, 5, 6	1. Security policy 2. Disciplinary sanctions 3. Encryption 4. Data disassociation & anonymisation 5. Activity log	Reduced
7	1. Define a privacy policy visible and accessible 2. Transparent reporting on the use and purpose of cookies 3. Dissuasive sanctions	Reduced
8	1. Training on duties and responsibilities regarding information confidentiality 2. Data disassociation & anonymisation 3. Disciplinary sanctions for staff members who breach the duty of secrecy and confidentiality policies of the organization	Reduced

"This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 690211"

9	1. Promote awareness on the obligation of professional secret regarding personal data 2. Disciplinary sanctions 3. Official communication channels with those workers informing about the responsibilities and consequences of accessing personal data 4. Official communication channel with authorities reporting any confidentiality breach.	Eliminated
10	1. Promote awareness of good practices 2. Disciplinary sanctions for those who breach the right of rejection or withdrawal of human subjects participating in medical research	Reduced
11	1. Identification of authorized users 2. Assignation of security responsible & security policy 3. Activity log	Eliminated
12	1. Security policy 2. Official communication with authorities reporting the attack	Reduced
13	1. Activity log 2. Recovery strategy	Eliminated
14	1. Initial estimation of requirements 2. Emergency plan	Accepted
15	1. Initial estimation of requirements 2. Emergency plan	Accepted

16	1. Security policy (non-transferable personal data among pilots) 2. Disciplinary sanctions 3. Data disassociation & anonymisation	Accepted
17	1. Physical security 2. Dissuasive and disciplinary sanctions	Reduced
18	1. Add criteria when considering HW/SW options in the architecture design	Eliminated

Annex 2 - C-MMD Datasets

PERSONAL DATASET (personal data, demographics, medical, social)				
Personal data				
Property	**Description**	**Data Type**	**Nullable**	**Unique**
User_permalink (auto generated) *	Unique alphanumeric code representing the user	String	False	TRUE
User_Id (auto generated)*	A unique Identifier (auto-increment)	Longint	FALSE	TRUE
NickName*	A user identifier to appear in public	String	FALSE	TRUE
Email*	The primary email address of user validated by the RFC 5322 Section 3.2.3	String	FALSE	TRUE
City	Name of city	String	FALSE	FALSE
Province / County	Name of province	String	FALSE	FALSE
ZIPCode	Postal code	String	False	False
CountryCode	International country code validated by the ISO 3166-1	String	FALSE	FALSE
SHA256* (technical, auto-generated)	Secure Hash Algorithm Code	String	FALSE	TRUE
SALT (technical, auto-generated)	The salt key for the one-way hashing of password	String	FALSE	FALSE
ConfirmationDate	Date the user account was confirmed validated by ISO 8601	DateTime	FALSE	FALSE
RegistrationDate	Date the user account was created validated by ISO 8601	DateTime	FALSE	FALSE
Role_Id*	Integer expression of the user role: [0: normal user, 1: PWLD, 2: caregiver; 3: social professional; 4: health professional]	Integer	FALSE	FALSE
Photo	Image or avatar	Image	FALSE	FALSE
Demographics				
Gender**	'M' for Males and 'F' for Females. 'O'	Char	FALSE	FALSE

	used for other or null.			
Birth date**	Date of birth validated by ISO 8601	DateTime	FALSE	FALSE
Preferred Language	Preferred language of the interface validated by ISO 639-1 from the list of available languages	String	FALSE	FALSE
Education Level	**[ISCED 2011]** Early childhood education Primary education Lower secondary education Upper secondary education Post-secondary non-tertiary education Short-cycle tertiary education Bachelor's or equivalent level Master's or equivalent level Doctoral or equivalent level	Byte	TRUE	FALSE
Living status (where)	**[Only for PLWD]** Own house Relative's house Social Housing Sheltered accommodation Nursing home	Byte	TRUE	FALSE
Living status (with whom)	**[For PLWD]** alone with the main caregiver (only as 2) with other family members). **[For CG]** alone with the carereceiver (only as 2) with the carereceiver(with others)	Byte	TRUE	FALSE
Working condition	**[Only for CG]** retired housekeeper full time employed part time employed	Byte	TRUE	TRUE

	free-lancer unemployed (looking for a job) student			
Working benefit	**[Only for working CG]** Do you benefit of any measure for those providing care (e.g. authorised working permits for caring)	Boolean	TRUE	TRUE
Other support	**[Only for CG]** No support Formal and professional carer (service provided by public health and social system) Formal and professional carer (service provided by private health and social provider) Volunteers from community charities and associations Other relatives	Byte	TRUE	FALSE
ICT or Technological devices	*Subjective estimation of computer driving skills* Willing to use Casual Expert	Byte	TRUE	TRUE
Hobbies	Animals/pets/dogs, Arts, Astrology, Astronomy, Baseball, Basketball, Beach/Sun tanning, Bird watching, Boating, Bonsai Tree, Cake Decorating, Calligraphy, Camping, Casino Gambling, Ceramics , Church/church activities, Collecting, Music, Computer activities, Cooking, Crafts, Crossword Puzzles, Dancing, Photography, Dominoes, Drawing, Eating out, Educational Courses, Electronics, Exercise (aerobics, weights), Falconry, Fishing, Floorball, Floral Arrangements, Football, Games, Gardening, Cinema, Golf, Guitar, Home Repair, Internet, Puzzles, Macramé, Painting, Photography, Piano, Reading, Shopping, Spending time with family/kids, Stamp Collecting, Swimming, Tennis, Traveling, TV watching, Video Games, Violin,	Array of Strings	TRUE	FALSE

	Volunteer, Walking, Writing, Yoga			
Health literacy	Ability to access information on medical or clinical issues Ability to understand medical information and derive meaning Ability to interpret and evaluate medical information Ability to make informed decisions on medical issues	Byte	TRUE	TRUE
Interests (self)	D1.4 Domains (binary tree, multiple selection) (p.35-39). See sample in Domain field of interventions. These interests are selected by the PLWD him/herself	Array	FALSE	FALSE
Interests (caregiver)	D1.4 Domains (binary tree, multiple selection) (p.35-39). See sample in Domain field of interventions. These interests are selected by the PLWD's caregiver.	Array	FALSE	FALSE
Interests (doctor)	D1.4 Domains (binary tree, multiple selection) (p.35-39). See sample in Domain field of interventions. These interests are selected by PLWD's doctor.	Array	FALSE	FALSE
Medical				
Cognitive disorder	MCI / Dementia	String	TRUE	FALSE
Type of dementia	**[If dementia]** Alzheimer Disease Frontotemporal Dementia Vascular Dementia Unknwon type of Dementia	Byte	TRUE	FALSE
Diagnosis date	Diagnosis date validated by ISO 8601	Date	TRUE	FALSE

Disorder	**[Only HCP]**
	Belongs to disorder group
	• Blood Condition
	  ○ Anemia
	• Cardiovascular diseases
	  ○ Atrial fibrillation
	  ○ Arterial Hypertension
		ArrayList	TRUE	FALSE

	○ Heart failure ○ Chronic myocardic ischemia ○ Recent myocardic infarction ○ Cerebral vascular disease ○ Peripherial arterial disease • Endocrine and metabolism disorders ○ Diabetes with insulin treatment ○ Diabetes II with oral treatment ○ diabetes with complications ○ Hypercholesterolemia ○ Hypothyroidism ○ Obesity or overweight ○ Malnutrition ○ Dehydration • Genitourinary disorders ○ Urinary tract infection ○ Urinary incontinence • Liver and gastrointestinal disorders ○ Chronic hepatitis ○ Gastritis or ulcer ○ Constipation • Nervous System disorders ○ Parkinson disease ○ Epilepsy ○ Stroke sequellae • Psychologic and psychiatric disorders ○ Insomnia ○ Anxiety ○ Depression ○ Psychosis ○ Confusion ○ Apathy ○ Behaviour disturbance • Rheumatologic disorders ○ Arthrosis ○ Joint pain • Sensory impairment ○ Deafness ○ Visual impairment			

	• Toxic
	  ○ Alcoholism and other addiction
	• Nephrology
	  ○ Chronic renal insuffiency
	• Respiratory disease
	  ○ Chronic pulmonary disease
	  ○ Respiratory infection
	• Cancer
	  ○ Malignancy
	Date of diagnosis
	UI impact (graded by its impact severity)
	• Audio impairment (influencing the access to platform services)
	• Visual impairment (influencing the access to platform services)
	• Physical impairment (influencing the caring activities)
	• Pharmacological therapy			

Weight	In kg / pound	int	TRUE	FALSE
Height	in cm / feet	int	TRUE	FALSE
Social				
Relationships	Ids of relationships with other users. (primary caregiver id is included too).	ArrayList	TRUE	FALSE
Privileges	Privileges of a relationship • Read posts • Read scales • Read events • Read medical info • Read activity	ArrayList	TRUE	FALSE
PrimaryCaregiver_Id (initial phase --> only one caregiver per PLWD)	Id of the user with primary responsibility of the caregiving	Longint	TRUE	FALSE
Groups_Id	Array of group names the user participates in	Array of Strings	TRUE	FALSE

ConnectionDegree	The number of nodes directly connected to this node (personal circle's size)	Integer	TRUE	FALSE
Centrality	Array of social network centrality metrics as triplet vectors of: [GroupName, CentralityType, CentralityValue] (See below for details).	Array	TRUE	FALSE

SCREENING DATASET *In order to develop the survey engine for making surveys online, the data structure used is described here.*				
Survey Id	Id of each survey	Integer		
Participant surveys	Info about each time a participant fulfills a survey			
Location	Place where the survey has been taken			
Notes	Comments about the survey			
Score	Final score of the survey, in case it is performed offline			
Status	undone: 0, in_course: 1, finished: 2			
Participant Survey Partial Scores	Info about partial scores of one survey, if needed			
Name/domain	Name to identify the partial score			
Score	Score for a fragment of a survey			
Participant answers	Answers of each participant to each proposed survey answer			
Participant answer id	Belongs to a participant survey			
Question_id	Belongs to a question			
Answer id	Belongs to an answer			
Content	Free text for a free text answer			
Is_answered	True if it is answered, false otherwise			

ADVERSE EVENTS DATASET				
Description	Description of adverse event (a diagnosis is preferred)	String	FALSE	FALSE
Start date	Date validated by ISO 8601	DateTime	FALSE	FALSE
End date	Date validated by ISO 8601	DateTime	FALSE	FALSE
Severity	Severity of the event: YES/NO	Boolean	FALSE	FALSE
Related to investigational product	Relationship to the investigational product	Boolean	FALSE	FALSE
Has many Counter measures:	Outcomes • Ongoing • Resolved • Not resolved Serious: YES/NO	ArrayList	TRUE	FALSE

TREATMENT DATASET				
Medication	Name of the drug	String	FALSE	TRUE
Prescribed date	Date for treatment ending. Ask for duration instead.	DateTime	FALSE	FALSE
End_date	Date of treatment end	DateTime	FALSE	FALSE
Dosage_array	7x4 matrix indicating number of pills / quantity for each weekday x 4 times a day.	ArrayList	TRUE	FALSE
Condition	To which disorder this treatment is intended for (see list in Personal dataset)	String	FALSE	FALSE
atc	Anatomical, Therapeutic, Chemical classification system	String	FALSE	FALSE
Administration route	Oral Intravenous Nasal Respiratory (inhalation) Transdermal Other https://www.fda.gov/drugs/developmentapprovalprocess/formssubmissionrequirements/el	Byte	FALSE	FALSE

	ectronicsubmissions/datastandardsmanualmonographs/ucm071667.htm			
Adherence	each row is a dosage (taken, not taken, missing info)	Byte	TRUE	FALSE

INTERVENTION DATASET				
Intervention_ID	A unique Identifier (auto-increment)	Longint	FALSE	TRUE
Date of creation	Automatic	Timestamp	FALSE	FALSE
Author	Name of content creator	String	FALSE	FALSE
Author_contact_details	Author contact details	Srring	FALSE	FALSE
Author_qualification		String	FALSE	FALSE
Language	Codes of Languages	String	FALSE	FALSE
Intended audience	Role-based content	Array of strings	FALSE	FALSE
Social exchange	Preventive/Palliative content	String	TRUE	FALSE

DOMAIN	D1.4 Domains (binary tree, multiple selection) (p.35-39)
	• Understanding dementia
	  ○ Symptoms
	  ○ Diagnosis
	  ○ Anxiety and confusion
	  ○ Memory loss
	  ○ Physical changes
	  ○ Progression
	  ○ Behaviour that challenges
	  ○ Medication
	  ○ Recognising I am a carer
	• Daily life
	  ○ Eating and drinking
	  ○ Living at home
	  ○ Using public toilets
	  ○ Washing and dressing
	  ○ Shopping
	  ○ Driving
	  ○ Going to the toilet and continence
	  ○ Days out and holidays
	  ○ Safety
	  ○ Communication
	  ○ Hiding things
		Array	FALSE	FALSE

	○ Coping with my reactions • Who can help? ○ Dementia support and groups ○ Other helpful organisations ○ Carers' support and groups ○ Paid carers ○ Care homes ○ GPs and other medical people ○ Hospital ○ Money ○ Social care ○ Day care ○ Planned respite ○ Emergency respite • Looking after myself ○ Appreciate the present ○ My health ○ Having a laugh ○ Frustration ○ Staying positive ○ "Me" time ○ Calm ○ Sleep ○ Someone to talk to • My Relationship ○ Seeing the person ○ Power and control ○ Maintaining independence ○ Maintaining a relationship ○ Arguing ○ Caring at a distance ○ Living better with dementia ○ Deciding things together ○ Doing things together • Friends and Family ○ Faith and community ○ Support groups ○ Friends ○ Enjoying a family life ○ Helping others to understand ○ Support from family members ○ Enjoying a social life ○ Keeping friendships going ○ Dementia friendly communities • Planning for the future ○ End of life ○ Helping my loved one to accept changes ○ Decisions			

	○ Getting help ○ Managing care when I am unwell ○ Preparing for the future ○ Anticipating changes ○ Knowing my limits ○ Planned respite			
[TYPE OF]	• Dementia Advisors • Post Diagnostic Groups • Signposting • Peer Support Groups • Stress/Anxiety management • Reminiscence • Assistive Technology: advice and support • Cognitive Training (CT) • Physical Exercise therapy • Fall prevention • Home modification • Music Therapy • Other contents	Byte	FALSE	FALSE
format	• Information about local services • Personal advice • Practical suggestions • Seminar/conference • Training event • Personal experience • Monitoring tools • External Resources • Entertainment solutions • Reward system	Array	TRUE	FALSE
delivery	• Infographic • Video • How-to • Guideline • Case study • Tips and tricks • Studies • Apps and games	Array	TRUE	FALSE
frequency	Frequency in days (e.g., every 3 days)	Int	TRUE	FALSE
repetitions	Amount of times (e.g., once)	int	TRUE	FALSE
Version	Id of each version of the intervention.	String	FALSE	TRUE
Is_active	Only one of the different versions of the intervention is published.	Boolean	FALSE	FALSE

Editor_Notes	Comments and notes about the intervention	String	TRUE	FALSE
Tags	Defined tags (disorders)	Array	TRUE	FALSE
Sources	Source of the content, referrals or mentions	String	FALSE	FALSE
content	path to raw file, information structure of rich content	String	FALSE	TRUE

USER INTERACTION DATASET				
NoOfPosts	Number of message posts	Integer	TRUE	FALSE
NoOfLikes	Number of Likes	Integer	TRUE	FALSE
NoOfReviews	Number of Reviews	Integer	TRUE	FALSE
NoOfArticleViews	Number of articles viewed by the user	Integer	TRUE	FALSE
NoOfArticleAuthored	Number of articles authored by the user	Integer	TRUE	FALSE
NoOfScalesTaken	Number of Scales taken by the user	Array	TRUE	FALSE
NoOfInvitationsReceived	Number of invitations received	Integer	TRUE	FALSE
NoOfInvitationsSent	Number of invitations sent	Integer	TRUE	FALSE
[InteractionHistory]	[Array of interactions-log file]	Array	TRUE	FALSE

USER GAMIFICATION MODEL DATASET				
User	The users table: user_id, nickname, role_id	ArrayList	TRUE	FALSE
Role	The Roles table stores a description of the user's id according to the roles in physical life and the health conditions: id, description	ArrayList	TRUE	FALSE
Games	This is the table containing the games the user participates in: game_id, title, description, etc.	ArrayList	TRUE	FALSE
Details	All the information mentioned in the short gamification profile, but organized per game title (points, badges, tangible objects, privileges).	ArrayList	TRUE	FALSE

GAME MODEL DATASET				
Game_Id	A unique identifier for each new game	int	TRUE	FALSE
Metrics	The table of metrics lists all kinds of metrics used to monitor user's activity and awarding back: metric_id, name_metric, type_metric, description	ArrayList	TRUE	FALSE
Actions	Contains the list of action and basic descriptors: actionId, title, description	ArrayList	TRUE	FALSE
Quests	List of quest Ids (timed objectives for the players proposed to them in order to gain a specific reward) with descriptors like questId, title, startDate, endDate, rewardId.	ArrayList	TRUE	FALSE
Leaderboards	List of leaderboards, their details and their range in players and team: leaderboard_Id, leaderboard_name, leaderboard__description, leaderboard_entity_type (players/team)	ArrayList	TRUE	FALSE
Rewards	The types of rewards to be applied in the game: Reward_id, reward_type, reward_verb, reward_condition	ArrayList	TRUE	FALSE
CreationDate	The datetime the game-master created this game	Datetime	TRUE	FALSE

SHORT GAMIFICATION MODEL DATASET				
[Game_Ids]	Array of game IDs the user participates in	Array of IDs	TRUE	FALSE
TotalPoints	Vector of four elements: The sum of all points earned in all games (like a wallet with points earned by social networking, communication activities, treatment adherence and education/training).communication activities, points earned by treatment adherence and points earned by education/training	Array	TRUE	FALSE
Badges	Array of all badges earned in all games	Array	TRUE	FALSE

TangibleObjects	Array of pairs: object name and quantity	Array	TRUE	FALSE
Privileges	For future use	Array	TRUE	FALSE

RECOMMENDER DATASET				
USER				
USER_ID	A unique Identifier	Longint	FALSE	TRUE
FEATURE VECTOR	Represents the user profile	Array of doubles	FALSE	FALSE
INTERVENTION				
INTERVENTION_ID	A unique Identifier	Longint	FALSE	TRUE
FEATURE VECTOR	Represents the item	Array of doubles	FALSE	FALSE

PILOT DATA DATASET (content introduced by clinic partners)				
ScalesScores	Array of pairs [psychological, medical and behavioral scales and scores]	Array	TRUE	FALSE
Additional notes	Notes associated to a pilot's visit	String	TRUE	FALSE

INTERVENTIONS FEEDBACK DATASET				
userId	Identifies the user that has provided feedback	Longint	FALSE	TRUE
interventionId	Identifies the intervention that has received some feedback from the given user (including interactive interventions like Serious Games for cognitive training)	Longint	FALSE	TRUE
timeStamp	Datetime of the intervention	Datetime	FALSE	TRUE
shared	Stores if the user has shared the intervention or not	Boolean	TRUE	FALSE

Views	Views (or tries) is the amount of times the user has consumed the intervention	Integer	FALSE	FALSE
links_used_counter	Number of times of usage for the links inside the interventions	Integer	FALSE	FALSE
Consumption time	Time spent by the user to consume this intervention. In case of interactive interventions this is the total time of playing a game.	Integer	TRUE	FALSE
Consumption time for subtask	Time spent by the user to perform each subtask of the intervention	Array	TRUE	FALSE
Success rate	Numeric descriptor of the level of success. In interactive interventions this will be the main performance indicator (e.g. 75% completed). For an article read with will be either 0% (not consumed) or 100% (fully consumed), or if the intervention has subsections then the rate of the completed subsections to the total number of subsections (e.g. 75% for 3 out of 5 things done).	Float	TRUE	FALSE
Score	Numeric expression of the success rate if any. In the case of games this is the score (number of points earned). In case of a questionnaire or an interview (given as intervention) this number could be the number of questions answered or the degree of user participation.	Float	TRUE	FALSE
Other	General purpose field related to the type of the intervention. For example, in some games we may need store the Average Response Time.			

USER GAMIFICATION INTERACTION HISTORY MODEL DATASET				
UserGamificationHistoryId	The id of an interaction performed by the user	Int	FALSE	TRUE
Timestamp	The date and time the interaction performed	DateTime	FALSE	TRUE
userId	The id of the user (player) who performed	Int	FALSE	FALSE

	the interaction			
gameId	The id of the game this interaction was performed	int	FALSE	FALSE
actionId	The id of the registered action related to this interaction	int	FALSE	FALSE
scoreEarned	The change in the sum points as a result of the interaction	int	TRUE	FALSE
awardId	The id of an award –if any- gained by the user at the time of this interaction	int	TRUE	FALSE
levelId	The id of the level the user was playing at the time of the interaction event	int	TRUE	FALSE
questId	The id of the quest possibly this interaction was related to	int	TRUE	FALSE
questStatus	An indicator if the interaction resulted in the enrolment of the user in a quest (value 1), or the drop-out of the quest by the user (value 1) or the winning of the quest (value 2)	int	TRUE	FALSE
gameStatus	An indicator if the game status resulted by this interaction event, enrolment in the game (value 1), or the drop-out of the game by the user (value 1) or the winning of the game (value 2)	int	TRUE	FALSE

GAME HISTORY DATASET				
gameHistoryId	The id of an event performed on the game	int	FALSE	TRUE
Timestamp	The datetime an event occurred in a game	Datetime	FALSE	TRUE
GameEventId	A short description of the game event like: Enable game, disable game, notification, reset, additions, updates and deletions of game elements (e.g. rules, actions, awards, etc.).	ArrayList	FALSE	FALSE
userId	The id of the game-master who made this change on the game	Int	FALSE	FALSE
Details	More information depending on the type of game event (e.g. content of the notification,	String	TRUE	FALSE

	name of the rule, etc.).			

NOTIFICATIONS MODEL DATASET				
notificationId	The users table: user_id, nickname, role_id	ArrayList	TRUE	FALSE
notificationCategory		ArrayList	TRUE	FALSE
nbotificationTitle		ArrayList	TRUE	FALSE
notificationText				
lifecycle	Time			
notificationStatus		ArrayList	TRUE	FALSE
userIds				

A type is said to be nullable if it can be assigned a value or can be assigned null, which means the type has no value whatsoever.

Annex 3 - Security/Data Breach Management Protocol

Identification

The identification phase of incident response has as its goal the discovery of potential security incidents and the assembly of an incident response team that can effectively contain and mitigate the incident:

a. Identify a potential incident. The incident handler may do so through monitoring of security sensors. System owners or system administrators may do so by observing suspicious system behaviour. Any user of the system may identify a potential security incident though external complaint/notification, or other knowledge of impermissible use or disclosure of Restricted Data.

b. Notify: users of the system that suspect an IT system has been accessed without authorization must immediately report the situation to rdlab@cs.upc.edu. Once the incident handler is aware of a potential incident, s/he will alert local system administrators.

c. Quarantine: The incident handler will quarantine compromised hosts at the time of notification unless they are on the Quarantine Whitelist. If they are on the Quarantine Whitelist, the incident handler will promptly reach out to the system administrator or system owner to create a plan to contain the incident. Note that the incident handler may notify on suspicious behaviour when they are not confident of a security compromise; in these cases they do not quarantine the host immediately, but wait 24-48 hours and quarantine only if the registered contact is unresponsive.

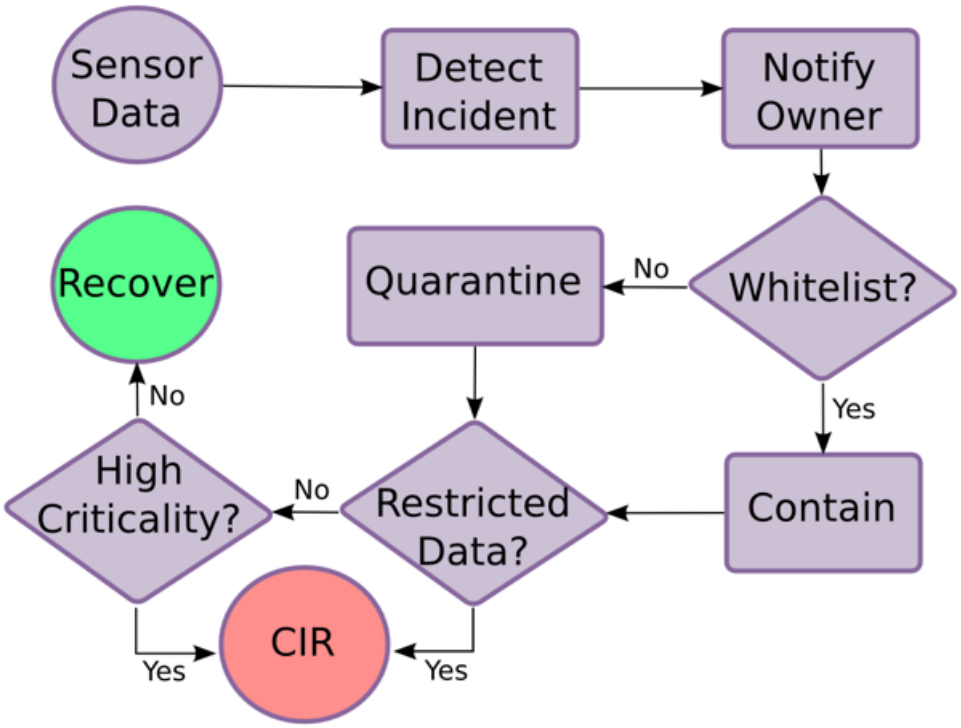

Figure 4 If Restricted Data is present on the compromised system, the Critical Incident Response (CIR) is followed.

Verification

This phase also precedes Critical Incident Response (CIR), and has the primary goal of confirming that the compromise is genuine and presents sufficient risk to engage the CIR process:

a. Classify: The CIR must be initiated if…
   i. The system owner or system administrator indicates that the system is a high-criticality asset
   ii. OR the system owner or system administrator asserts that the system contains Restricted Data.
   iii. OR someone of appropriate authority (for example, the Rector) with input from a cognizant UPC officer determines that the system poses a unique risk that warrants investigation.

b. Verify: The CIR process should be initiated ONLY if…

i. The incident handler verifies that the triggering alert is not a false positive. The incident handler will double-check the triggering alert, and correlate it against other alerting systems when possible.
ii. AND the type of data or system at risk is verified to be of an appropriate classification, as determined above. The system owner or system administrator should provide a detailed description of the data at risk, including approximate numbers of unique data elements at risk, and the number, location, and type of files it is stored in.

The order of the steps above can vary from incident to incident, but for the CIR process to be initiated the criticality of the asset must be confirmed, and it must be confirmed that the triggering event is not a false positive. In cases where the CIR process is not required, the incident handler can resolve the case as follows:

a. Obtain a written (email is acceptable and preferred) statement from the system owner or system administrator documenting that the system has no Restricted Data and is not a high-criticality asset.
b. Obtain a written statement from the system owner or system administrator that the system has been reinstalled or otherwise effectively remediated before quarantine is lifted.
c. For incidents involving an unauthorized wireless access point, obtain a written statement that the access point has been disabled.

Containment

The containment phase represents the beginning of the CIR workflow and has the following goals:

a. If the host cannot immediately be removed from the network, the incident handler will initiate a full-content network dump to monitor the attacker's activities and to determine whether interesting data is leaking during the investigation.
b. Eliminate attacker access: Whenever possible, this is done via the incident handler performing network quarantine at the time of detection AND by the system administrator unplugging the network cable. In rare cases, the incident handler may request that network operations staff implement a port-block to eliminate attacker access. In cases where the impact of system downtime is very high, the incident handler will work with system administrators to determine the level of attacker privilege and eliminate their access safely.

c. The incident handler will collect data from system administrators in order to quickly assess the scope of the incident, including:
    i. Preliminary list of compromised systems
    ii. Preliminary list of storage media that may contain evidence
    iii. Preliminary attack timeline based on initially available evidence
d. Preserve forensic evidence:
    i. System administrators will capture first responder data if the system is turned on. The incident handler will provide instructions for capturing this data to the individual performing that task.
    ii. The incident handler will capture disk images for all media that are suspected of containing evidence, including external hard drives and flash drives.
    iii. The incident handler will dump network flow data and other sensor data for the system.
    iv. The incident handler will create an analysis plan to guide the next phase of the investigation.

This is the most time-sensitive and the most contextually dependent phase of the investigation. The actions that need to be taken will depend on the uptime requirements of the compromised system, the suspected level of attacker privilege, the nature and quantity of data at risk, and the suspected profile of the attacker. The most important goals of this phase are to eliminate attacker access to the system(s) as quickly as possible and to preserve evidence for later analysis.

Additionally, this is the phase where the incident handler works most closely with system administrators and system owners. During this phase, they are expected to take instruction from the incident handler and perform on-site activities such as attacker containment, gathering first response data, and delivering host-based analysis if required.

Analysis

The analysis phase is where in-depth investigation of the available network-based and host-based evidence occurs. The primary goal of analysis is to establish whether there is reasonable belief that the attacker(s) successfully accessed Restricted Data on the compromised system. Secondary goals are to generate an attack timeline and ascertain the attackers' actions.

All analysis steps are primarily driven by the incident handler, who coordinates communications between other stakeholders, including system owners, system

administrators, and relevant compliance officers. Questions which are relevant to making a determination about whether data was accessed without authorization include:

a. Suspicious Network Traffic: Is there any suspicious or unaccounted for network traffic that may indicate data exfiltration occurred?
b. Attacker Access to Data: Did attackers have privileges to access the data or was the data encrypted in a way that would have prevented reading?
c. Evidence that Data was Accessed: Are file access audit logs available or are file system mactimes intact that show whether the files have been accessed post-compromise?
d. Length of Compromise: How long was the host compromised and online?
e. Method of Attack: Was a human involved in executing the attack or was an automated "drive-by" attack suite employed? Did the tools found have capabilities useful in finding or exfiltrating data?
f. Attacker Profile: Is there any indication that the attackers were data-thieves or motivated by different goals?

In the case of a potential Breach of sensible data (SD), this analysis will include the C-MMD Coordinator, the UPC Data Officer and the RDLab Security Manager. They will conduct a risk assessment to determine the probability that the security or privacy of the C-MMD machines has been compromised based on an evaluation of the elements above in addition to the following four factors:

a. the nature and extent of the SD involved, including the types of identifiers and the likelihood of re-identification,
b. the unauthorized person who used the SD or to whom the disclosure was made,
c. whether the SD was actually acquired or viewed, and
d. the extent to which the risk to the SD has been mitigated.

Using these factors, UPC Data Officer will determine the degree of technical probability that the security or privacy of the SD has been compromised, but the final determination belongs to the affected C-MMD Pilot Partner. In order to make this determination, the Data Officer at the affected Pilot Partner will document each impermissible use and disclosure and the risk assessment conducted for each.

Exceptions to the definition of a Breach of SD are:

a. Any unintentional acquisition, access, or use of protected health information by a workforce member or person acting under the authority of a covered entity or a business associate, if such acquisition, access, or use was made in good faith and

within the course and scope of authority and does not result in further access, use or disclosure.

b. Any inadvertent disclosure by a person who is otherwise authorized to access protected health information at a covered entity or business associate to another person authorized to access protected health information at the same covered entity or business associate, or organized health care arrangement in which the covered entity participates, and the information received as a result of such disclosure is not further accessed, used or disclosed in a manner not permitted under LOPD.
c. A disclosure of protected health information where a covered entity or business associate has a good faith belief that an unauthorized person to whom the disclosure was made would not reasonably have been able to retain such information.

At the conclusion of the analysis, but before the final report is written, a peer review should be requested of the other RDLab technical staff. Complete the write-up of the notes, including conclusions, and archive processed source materials (e.g., grep-results, file-timelines, and filtered flow-records). The peer review may result in some issues that must be addressed and some issues that may optionally be addressed. All recommendations should be resolved or acknowledged and deferred. The incident handler's role is to determine, from a technical perspective, whether there is a reasonable belief that Restricted Data, including SD, was available to unauthorized persons. The determination of whether the circumstances warrant a Breach notification will be made jointly by the UPC Data Officer convened upon review of the results of the investigation and the technical opinion of RDLab.

Recovery

The primary goal of the recovery phase is to restore the compromised host to its normal business function in a safe manner.

a. The system administrators will remediate the immediate compromise and restore the host to normal function. This is most often performed by reinstalling the compromised host; although if the investigation confirms that the attacker did not have root/administrator access other remediation plans may be effective.
b. The system administrators will make short-term system, application, and business process changes to prevent further compromise and reduce operating risk.

Data Retention

a. The incident handler will archive the final report in case it is needed for reference in the future; reports must be retained for six (6) years.
b. Incident notes should be retained for six (6) months from the date that the report is issued. This includes the confluence investigation page, processed investigation materials like grepped file-timelines and filtered network-flows, etc.
c. Raw incident data should be retained for thirty (30) days from the date that the report is issued. This includes disk-images, unfiltered netflow-content, raw file-timelines, and other data that was collected but deemed not relevant to the investigation.

User Notification

a. If the risk assessment determines that a Breach has occurred, the UPC will provide written notice without unreasonable delay and in no event later than sixty (60) days from incident discovery, to the user or:
    i. If the user is deceased, the next of kin or personal representative.
    ii. If the user is incapacitated/incompetent, the personal representative.
    iii. If the user is a minor, the parent or guardian.
b. Written notification will be in plain language at an appropriate reading level with clear syntax and language with no extraneous materials.
c. Written notification will be sent by first-class mail to the last known address of the patient or, if deceased, the next-of-kin, or if specified by the user, by encrypted electronic mail.
d. Written notification will contain:
    i. A brief description of what occurred with respect to the Breach, including, to the extent known, the date of the Breach and the date on which the Breach was discovered;
    ii. A description of the types of compromised data that were involved in the Breach;
    iii. A description of the steps the affected individual should take in order to protect himself or herself from potential harm resulting from the Breach;
    iv. A description of what the UPC is doing to investigate and mitigate the Breach and to prevent future Breaches; and
    v. Contact procedures for individuals to ask questions or learn additional information, which will include a telephone number, an email address, Web site or postal address.

e. If the UPC determines the user should be notified urgently of a Breach because of possible imminent misuse of compromised data, the UPC may, in addition to providing notice as outlined in steps b-d above, contact the user by telephone or other means, as appropriate.