REPORT

Data Management Plan, M18 update

- project deliverable 1.2

Authors: Ville Ollikainen (VTT)
Carolina Goberna Caride (University of Passau)

Confidentiality: Public

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825585.

HELIOS Data Management Plan, M18 update	
Project name HELIOS	Grant agreement # 825585
Author(s) Ville Ollikainen, Carolina Goberna Caride	Pages 2+20
Reviewers Pau Pamplona Negre (UAB), Carlos Alberto Martín Edo (ATOS)	
Keywords data, data management, open data, gdpr	Deliverable identification D1.2

Summary

This document is the M18 update of HELIOS Data Management Plan, based in the initial version, including:

a) a description of the technical and organisational measures that will be implemented to safeguard the rights and freedoms of the research participants, possible consequences of profiling, and how their fundamental rights will be safeguarded;

b) how the data subjects will be informed of any possible existence of the profiling or data breaches; and

c) participation in Open Data Pilot.

The Data Management Plan will provide better understanding of the data produced, establish mechanism on how shared data will be exchanged and stored and secured with the collaboration tools of the project and what data will be publicly available and how.

Not all research data will be openly accessible: Specifically, no such data that can be considered personal data will be published. For other data, publishing as Open Data will be a default whenever the data is considered reusable.

This document addresses how the data published by HELIOS will follow FAIR principles – findable, accessible, interoperable and re-usable – including what data the project will generate, whether and how it will be made accessible for verification and re-use, and how it will be curated and preserved.

This document will be updated finally in D1.3 (M28).

Confidentiality	Public

Espoo, Finland 25.6.2020
Written by

Ville Ollikainen
Carolina Goberna Caride

Contact address
Ville Ollikainen, ville.ollikainen@vtt.fi, +358 400 841116

Distribution
HELIOS project partners, subcontractors, the Project Officer and HELIOS web site

Contents

History of Changes

Row#	Revision date	Revision description
1	2020-06-16	First version based on D1.1
2	2020-06-17	Internal review version
3	2020-06-24	To internal review
4	2020-06-25	Final version

1. Introduction

1.1 HELIOS in brief

HELIOS designs, implements and validates a state-of-the-art, decentralized peer-to-peer social media platform that:

- supports the ad-hoc creation and management of social graphs within the context and proximity of the user;
- provides all necessary means and features to ensure the highest level of trust and control is built into the platform (trust by design) which includes among others novel decentralization, content creation and sharing as well as new ways to control monetisation channels and means, and;
- ensures privacy, respecting ethical and legal requirements.

HELIOS also participates in the Open Research Data Pilot in Horizon 2020[1], which aims to improve and maximise access to and re-use of research data generated by Horizon 2020 projects. However, due to the nature of peer-to-peer approach in HELIOS, there is no centralized storage for user data: This fundamental design principle has an influence on, what data may or may not be available from the platform. Yet, research data may be collected from HELIOS platform by implementing special arrangements, while accumulating other research data that is subject for publishing.

1.2 Privacy overview

In its studies, HELIOS collects personal data that must be managed appropriately and excluded from any open data. For personal data, specific measures described in Chapter 4 will be reminded and enforced.

Data management relates to data minimisation, which in turn originates at the EU General Data Protection Regulation ('GDPR'). GDPR regulates the processing of personal data relating to individuals in the EU. To be compliant with the GDPR[2], HELIOS must ensure processing *only* personal data necessary for each specified purpose.

[1] Extended Pilot on Open Access to Research Data, in ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/open-access_en.htm

[2] GDPR Art.5 and Recital 39. Available at https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679

1.3 About this document

This is the M18 update of HELIOS Data Management Plan, including:

a) description of the technical (Chapter 4) and organisational (Chapter 3.4) measures that will be implemented to safeguard the rights and freedoms of the research participants, and;

b) explanation on how the data subjects will be informed of any possible existence of the profiling (Chapter 4.5).

This document is based on HELIOS deliverable D1.1 “Data Management Plan, initial version”, submitted on June 29, 2020. Only minor updates have been necessary, mainly concerning Open Data Pilot participation.

As in D1.1, this document will address measures for processing personal data as well as practises on publishing open data. Related to the topic, Data Minimisation guidelines, e.g. ICO guidance on data minimization[3], will be followed.

This document will be updated over the course of the project whenever significant changes with impact on data arise, such as:

- new datasets identified,
- changes in consortium policies (e.g. new innovation potential, decision to file for a patent),
- changes in consortium composition and external factors (e.g. new consortium members joining or old members leaving).

The last scheduled update for this Data Management Plan is D1.3 in M28 (April 2021). It will further refine topics described in this document.

These documents adopt the guidelines of the Digital Curation Centre on how to implement the Data Management Plan[4] and the recommendations of the EC as published in the Guidelines on FAIR Data Management in Horizon 2020 version 3.0[5].

[3] The Information Commissioner’s Office (2018) Data Minimisation. 1.0.208, 02 August 2018. Accessed in July 2019 at https://iapp.org/media/pdf/resource_center/ICO-data-minimisation.pdf

[4] How to Develop a Data Management and Sharing Plan. Accessed in June 2019 at http://www.dcc.ac.uk/resources/how-guides/develop-data-plan

[5] Guidelines on FAIR Data Management in Horizon 2020. Accessed in June 2019 at http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

2. Data Summary

2.1 Purpose of the data

Purpose of the data that HELIOS will collect, process and/or generate data falls into three categories:

a. Deliverables: all captured (written or other forms) information regarding project's execution, together with descriptive tests.

b. Content: content used in the pilots and other media assets, created for the purposes of the project.

c. User data: the data gathered from users who take part in the validation activities (i.e. lab tests, trials, pilots) or are otherwise involved in the project.

2.2 Formats of the data

Data in these three categories vary in type and formats: "Deliverables" are expected to consist of text, image, software code and presentation data; "Content" includes pre-generated video and audio files that will be used in the pilots to showcase the functionalities realized within the project; "User data" includes both the subjective (questionnaires, evaluations, etc.) and objective (point cloud, mesh, voice recording, etc.) data which is anonymized or pseudonymized whenever applicable, otherwise treated in compliance with the GDPR.

HELIOS promotes using non-proprietary and open formats, preferring standards and best documented practises, which are commonly in use by the project partners and the relevant research community.

2.3 Re-using existing data

HELIOS will consider all data equally, no matter where the origin is, within the terms of original purpose and possible licensing. This data may originate e.g. from previous research or from external sources.

2.4 Origin of the data

Depending on the category, data can be generated by:

- project partners during its execution (deliverables and content)
- subcontractors (deliverables and content)
- third party organisations (deliverables and content)
- professional or end-users taking part in project activities, such as workshops, validation activities, requirements definition sessions and exhibitions (user data)

Project execution, experiments and pilots will be executed in different locations. Tasks will run concurrently, and in many cases, data of all three categories will be created.

2.5 Size of the data

Expected datasets are:

- Deliverable documents, expected size: small
- Focus reports, expected size: small
- Evaluation questionnaires, expected size: small
- Raw and edited video and audio content, expected size: large
- Content enrichments (indexing metadata, subtitles, sign language, audio description), expected size: large

2.6 Utilizing data

HELIOS aims at making as much (non-personal) data as possible publicly available, giving visibility for the project, maximizing the exploitation of its results, while looking towards long-term impacts and other opportunities. Whenever applicable, data generated in the project will follow the FAIR guidelines provided by the European Commission[5] and be published and distributed in relevant repositories. Furthermore, new knowledge will be published in local and international scientific conferences and academic publications, ensuring that all interested researchers are aware of the projects execution, status and results.

However, not all data will be published. Publishing data will be evaluated taking into account exploitation value of the data, methods for anonymization and pseudonymisation, risks associated with adverse merging of the data set to any known other data set or data set type, together with the Ethical considerations. A related procedure is addressed in Chapter 3.4:

As described in Chapter 3.4, the Project Coordinator will take the decision, about which data will be published as Open Data, on case-by-case basis, after consulting the Steering Committee of the project. The purpose for open datasets relates to disseminating project results, allowing scientific community to reuse the knowledge. The following table presents exemplary metadata for published information, to be filled in during the course of the project:

Responsible partner	Type	Description	Format	File Type	Owner	Open (Y/N)	Licensing	repository

In case third parties request access to restricted or private data sets, the request will be presented to the Project Coordinator, who will consult the Steering Committee of the project and forward the request to the responsible partner.

2.7 Data stakeholders

Data stakeholders within the project can come from many different backgrounds and with different intentions. Depending on their importance to the project's execution decisions are made in order to increase ease of access. The most important stakeholders are:

- Project partners
- Subcontractors
- European Commission
- European Social Media research cluster
- Research community
- Virtual Reality specialists
- Media content producers
- Media industry

3. FAIR data

3.1 Notes on open data

3.1.1 Open Access in H2020

Open Access (OA) refers to the practice of providing to any user, free of charge, online access to all peer-reviewed scientific information and research data generated in a project. As indicated in Article 29.2 of the European Commission Grant Agreement[6] the members of the consortium must:

a. As soon as possible and at the latest on publication date, deposit a machine-readable electronic copy of the published version or final peer-reviewed manuscript accepted for publication in a repository for scientific publications; Moreover, the beneficiary must aim to deposit at the same time the research data needed to validate the results presented in the deposited scientific publications.

b. Ensure open access to the deposited publication — via the repository — at the latest:
   - On publication, if an electronic version is available for free via the publisher,
   - Within six months of publication (twelve months for publications in the social sciences and humanities) in any other case.

c. Ensure open access — via the repository — to the bibliographic metadata that identify the deposited publication.

To ensure open access it may be agreed in the project to use a repository containing the stored research publications as well as the public repository as for example Zenodo[17] or the one at UAB[7].

3.1.2 Open Data in H2020

The project HELIOS will be part of the Research Data Pilot. In consequence and regarding the digital research data generated in the action ('data'), the beneficiaries must:

a. deposit in a research data repository and take measures to make it possible for third parties to access, mine, exploit, reproduce and disseminate — free of charge for any user — the following:
   i. the data, including associated metadata, needed to validate the results presented in scientific publications as soon as possible;
   ii. other data, including associated metadata, as specified and within the deadlines laid down in the subsequent versions of this Data Management Plan;

b. provide information — via the repository — about tools and instruments at the disposal of the beneficiaries and necessary for validating the results (and — where possible — provide the tools and instruments themselves).

[6] H2020 Model Grant Agreements: H2020 General MGA - Multi: v5.0. Accessed in July 2019 at http://ec.europa.eu/research/participants/data/ref/h2020/mga/gga/h2020-mga-gga-multi_en.pdf

[7] Dipòsit Digital de Documents de la UAB, public repository at UAB. Accessed at https://ddd.uab.cat

3.1.3 Open Data in HELIOS

Open data published by HELIOS will be summarized in the following format:

Data set name	HELIOS-[Partner-abbreviation]-[number]-[version][8]
Date	yyyy-mm-dd
Responsible partner	
Summary of the data	
How the data is generated	[equipment, software, research infra…]
Format	[Format or document name]
Volume of the data	[x]GB
User consent, if applicable	[Consent, document name or N/A]

Possibility to re-use any existing open research data will be examined carefully during the project.

Whenever use of specific software is mandatory or recommended for processing the data, it will be presented in the summary. Potential re-utilization will be enabled and quality of the data ensured by careful documentation of data collection methods as well as the contents of the datasets.

List of existing and planned HELIOS Open Data datasets are presented in the Annex: Dataset descriptions (June 24th, 2020).

3.1.4 Quality assurance

Quality control measures will be taken to maintain the accuracy of data throughout the project.

Project's deliverables, audio-visual content and scientific publication will be peer-reviewed by project partners and in some cases by external reviewers as well. Through this approach we expect to ensure high data quality within the project, promoting project data reuse and sharing.

3.2 Notes on personal data

3.2.1 What is 'personal data'?

It is essential for any EU project to understand if some data is potentially personal, because of regulations and ethical guidelines. In the sense of the GDPR[9], personal data is any data that *can* (even indirectly) identify an individual.

Example: IP addresses The word 'can' above has a broad sense. For instance, consider if you have a web site and you keep log of visitors' IP addresses; you will collect 'personal data'.

[8] This is an example. More descriptive names are preferred for published datasets.
[9] The General Data Protection Regulation. Available at http://data.europa.eu/eli/reg/2016/679/2016-05-04

Why?

Let us think the opposite: *If* IP addresses would *not* be personal data, they could be shared freely. Now, someone who would receive the logs from another party could merge the records with their own records e.g. login information associated with same IP addresses, disclosing identities in the received data and users' actions there.

Therefore, IP addresses are personal data. No matter, if you share them or keep them in your own records, they are.

Example: pseudonymising data.

It is easy to think that if you somehow convert a user identifier into some gibberish, the data is no longer personal data. Unfortunately, it is not as simple as that. Let me explain:

First, if a security specialist (acting here as an adversary) gets hold on the data, also she can do the same conversion (if known) or try out a limited set of well-known conversions. For instance, let's assume that 'Willem' is replaced by a Blowfish HASH of it:

$2y$10$s.bePR/qn86Ls09bZj9jrexM9qSs51T7muBg5P3FZnM2h6CHAjYse

When the adversary knows to look for 'Willem', she can go through all well-known HASHing algorithms, and once she applies Blowfish - *voilà* Willem data is breached! This is related to a so-called Rainbow Attack method.

(You could use some proprietary algorithm, sure, but then you would sail into truly dangerous waters with insecurity-by-obscurity.)

As a countermeasure we could 'salt' the HASH… pre-convert 'Willem' into something different before HASHing, and previously discussed attack becomes futile (unless the adversary gets to know the 'salt', of course, which may happen as well).

However, even with salting, if the same 'salt' is used in different occasions, the adversary can see that the gibberish ID repeats. Now it needs to disclose the identity only once, and *voilà* the user becomes totally exposed. Therefore, it is still very vulnerable.

We should not blindly count on pseudonymization (HASHes and so forth), as the data may still stay personal. When going for pseudonymization, it is highly recommended to consult cybersecurity experts.

Under no circumstances will HELIOS publish personal data of any kind, unless there are legal grounds for it, until further notice is provided after updating this document.

3.2.2 Minimizing personal data

HELIOS has produced Data Minimisation Guidelines as a confidential (internal) deliverable D1.5.

As a summary, HELIOS must collect only such personal information that is directly relevant and necessary to accomplish a *specified purpose*. The owners of the data should be asked for consent, and the specified purpose should be mentioned there.

The data should be retained only for as long as is necessary to fulfil that purpose. In other words, it is advisable to get rid of any personal data promptly.

When minimising the data, it should be considered, what data is actually needed and why. If there is no reason to collect some data, it should not be collected.

Example: Travel card[10]

A town council offers a chip card to regular users of the town's public transport system for a certain fee. The card carries the name of the user in written form on the card's surface and also in electronic form in the chip. Whenever a bus or tram is used, the chip card must be passed in front of the reading devices installed, for example, in buses and trams. The data read by the device are electronically checked against a database containing the names of the people who have bought the travel card.

This system does not adhere to the data minimisation principle in an optimal way: checking whether an individual is allowed to use transport facilities could be accommodated without comparing the personal data on the card's chip with a database. It would suffice, for instance, to have a special electronic image, such as a bar code, in the chip of the card which, upon being passed in front of the reading device, would confirm whether the card is valid or not. Such a system would not record who used which transport facility at what time. This would be the optimal solution in the sense of the minimisation principle, as this principle results in the obligation to minimise data collection.

The ICO (Information Commissioner's Office, the UK's independent body set up to uphold information rights) has published an excellent briefing for Data Minimization[11]. Briefly:

You must ensure the personal data you are processing is:

- adequate – sufficient to properly fulfil your stated purpose;
- relevant – has a rational link to that purpose; and
- limited to what is necessary – you do not hold more than you need for that purpose.

3.2.3 Using public datasets containing personal data

There are publicly available datasets, which contain data that may be considered personal, for instance databases collected from Facebook or Twitter platforms (hereinafter referred as Social Networking Platforms, or SNP's), and which can be used by researchers.

In Europe, when analysing or processing data within SNP's, it is done under specific, explicit and legitimate research purposes. The legitimacy of the purpose requires that processing must be based in any of the lawful requirements for it established in Article 6 of the GDPR. In this case,

- researchers may rely on the GDPR Art. 6.1(f) or Art. 9.2(j)[9] considering that the data is necessary for the legitimate interest pursued or,

[10] The Handbook on European data Protection Law, FRA (Fundamental Rights Agency) 2018, p. 126

[11] The Information Commissioner's Office (2018) Data Minimisation. 1.0.208, 02 August 2018. Accessed in March 14, 2019, at https://iapp.org/media/pdf/resource_center/ICO-data-minimisation.pdf

- if researchers do not have the data subjects' consent, they should apply Art. 6.4 to justify the compatibility of the criteria with the purpose.
- Recital 50[12] in connection with Art. 6.4 establishes that scientific research processing purposes are to be considered compatible lawful processing operations.

Furthermore, regarding personal data, which has not been obtained directly from the data subject, researchers may proceed according to Article 14 and Recital 62[13]. When providing the data subject with information about the controller involves a disproportionate effort, and the purpose of processing entails scientific research purposes (among others), the safeguards and conditions established in Article 89.1 and 2 apply and Recital 156[14] and 159[15] must be considered. Under Article 89.2 the following rights of data subjects may be derogate:

- Art. 15 right of access by data subject
- Art. 16 right to rectification
- Art. 18 right to restriction of processing
- Art. 21 right to object

Appropriate measures to protect the data subject´s rights (such as anonymization or pseudonymisation) must still be applied.

Before collecting any data from any 3rd party platform, ”terms of service” and ”community standards” of the platform must be respected. Due to complexity of the topic, legal consultation is mandatory, preferably from a dedicated project partner.

Under no conditions will HELIOS provide any personal data as Open Data. It should be noted that while this is self-evident alone, there is also a statement in OpenAIRE instructions regarding a Data Management Plan[16]: “Scientific research data should be”…“ Useable beyond the original purpose for which it was collected”. This simply excludes any personal data, since under GDPR, user consent only applies for specific purposes.

[12] Recital 50 of the GDPR: Further processing of personal data. Accessed in June 2019 at https://gdpr.eu/recital-50-further-processing-of-personal-data/

[13] Recital 62 of the GDPR: Exceptions to the obligation to provide information. Accessed in June 2019 at https://gdpr.eu/recital-62-exceptions-to-the-obligation-to-provide-information/

[14] Recital 156 of the GDPR: Processing for archiving, scientific or historical research or statistical purposes. Accessed in June 2019 at https://gdpr.eu/recital-156-processing-for-archiving-scientific-or-historical-research-or-statistical-purposes/

[15] Recital 159 of the GDPR: Processing for scientific research purposes. Accessed in June 2019 at https://gdpr.eu/recital-159-processing-for-scientific-research-purposes/

[16] OpenAIRE. What is a Data Management Plan (DMP). Accessed in June 2020 at https://www.openaire.eu/what-is-a-data-management-plan-dmp

3.3 Making data findable

3.3.1 Data discovery

All such data produced and/or used in the project that is aimed for sharing, will be associated with publicly available metadata. Use of identifiable and locatable standard identification mechanisms (e.g. persistent and unique identifiers such as Digital Object Identifiers) will be considered in the course of the project.

Discipline compliant metadata elements will be used describing the data to aid data discovery and potential re-use. List of metadata elements and metadata standards used are provided in a separate spreadsheet. Metadata of opened data will be made available for research and re-use after project closure.

3.3.2 Naming conventions and version numbers

Each set of data produced (dataset, deliverables, video footage…) will be named in a uniform way and will include a table with a version control.

- For deliverables: Dx.y - [Name of the deliverable as described in the DoA] being x - work package assigned to the deliverable y - the number of deliverable within the work package, e.g. D1.5 “Data Minimisation Guidelines”
- For datasets[8]: WP [Work Package number] P [Validation activity or pilot number] - [description of the activity] i.e.: WP7 P9 - VTT_summer_trainee_usage_data
- For Video footage: WP7 T7.1 MEDIA T0x (short description).

Names *may* be prefixed by the project name HELIOS, and they *may* be postfixed by versioning information, such as date (in the format yyyy-mm-dd).

The table for controlling the version will include the following fields:

- Revision: number (starting from 0.1) of the version. Always following the order.
- Date: date in which that version was available, in the format yyyy-mm-dd.
- Author: who prepared the version.
- Organization: entity in charge of that version.
- Description: summary of main changes in that version.

For the video footage, it is not expected to use versioning, but in case it is necessary, clear versioning will be provided, as well as a “readme text” stating what are the improvements regarding the previous version.

3.4 Making data openly accessible

Due to strict privacy regulations in the European Union, no HELIOS datasets are openly accessible by default. Having said this, HELIOS will maintain record of its data sets, and actively seek for opportunities to prepare any data set for open access.

Focus in data sharing will be on the data underlying prospective scientific publications ensuring the validation of results presented in publications.

3.4.1 Personal data

Before any official, reviewed and Steering Committee approved update to this Data Management Plan, HELIOS will under no circumstances publish any data that is personal data, or data that may be adversely converted into personal data by e.g. merging the data set to any other known data set or data set type. The only exception is legal grounds that may force HELIOS to do so.

In subsequent updates of this Data Management Plan, making data sets openly accessible may be possible, after considering especially:

- ethical aspects,
- security measures,
- technical measures and
- organisational measures,

to guarantee privacy of pseudonymised or anonymised data and the right of vulnerable groups.

3.4.2 Non-personal data

Decisions about publishing (selected) non-personal datasets will be taken by the Project Coordinator after these datasets are reviewed by the Steering Committee of the project. A main criterion is foreseeable reusability of the data itself.

With this guideline, Project Coordinator in collaboration with project partners will act appropriately to make relevant data openly available and usable for third parties for study, teaching and research purposes, making data persistently available and findable.

The steering Committee will review the following aspects

- technical measures against de-anonymising of any data in the dataset
- protection of minority groups, especially vulnerable people

If, after the project has been closed, permission to re-use the data is required, all requests for further use of data will be considered carefully and whenever possible approved by the Project Coordinator or the person mandated with the task. Permission for data use will be granted providing there are no privacy, IPR or confidentiality issues involved or any direct overlap of research questions with the primary research. Permission will be provided by contacting the Project Coordinator, whose contact information and appropriate procedure will be provided together with other metadata.

3.5 Making data interoperable

Interoperability of the project's data relates to the impact of the project and the possibilities of re-use, migration on different platforms and extrapolation of its results. When this is possible, open file formats, open vocabularies, relevant standards and best practises will be applied to maximize interoperability.

Variables and value names will be constructed following general data processing conventions common to the research subject. List of value names and used vocabulary will be provided in a separate list. Examples of vocabulary information to be managed within the project will be e.g. number of variables / units of observation, list of variables with the name and label of each variable

as well as its values and value labels, frequency distribution of each variable, information on the classifications used and meanings of abbreviations used.

3.6 Increase data re-use

Whenever possible, data generated and used in the project will be made publicly available following FAIR principles. When this is not possible, licensing possibilities will be examined on case-by-case basis. Research publications, questionnaires, video footage and audio content are likely to be licensed under Creative Commons Attribution-NonCommercial 4.0 (strictly forbidden to use it, totally or partially, for commercial purposes) license.

Published and FAIR-compatible data will be archived in a common and open data repository. Recommended generic and certified repository services, such as CERN's Zenodo[17] (*HELIOS preference*), CSC's IDA[18] or UAB's DDD[7] will be used to enhance long-term accessibility and re-usability of the data.

Ownership of datasets will belong to involved partners after the project has completed. Creative Commons licence CC-BY-SA or CC-BY will be used for any opened datasets, unless there are compelling reasons to select more restricted type of CC-licence. Creative commons licences will by default include also a disclaimer of liability for the re-use of opened data.

No definite period, or time limit, is planned for access or re-use the data; to be defined in subsequent revisions of this document. Justification for possible case-specific embargo for published data will be decided by project consortium. Embargo will be sought primarily in connection with any potential patent application based on project results.

3.7 Allocation of resources

Costs related to research data management and opening are eligible as part of the Grant Agreement[6] of the project.

During the project, consortium partners will be responsible for managing and curating datasets at their possession. Evaluating eligibility to publish certain data sets is a part of maintaining and reporting those data sets within the project.

Overall data management is part of project management, therefore responsibility of the Project Coordinator. When closing the project, its Steering Committee will mandate the Project Coordinator or an assigned project data manager to take care of long-term preservation and sharing of datasets.

Long-term access and its costs will be discussed separately. Academic partners are expected to have procedures and best practises for related activities.

[17] Zenodo open data repository. Accessed in June 2019 at https://zenodo.org/

[18] Research data storage service IDA. accessed in June 2019 at https://www.fairdata.fi/en/ida/

4. Data security

4.1 General aspects

Access to any dataset that contains personal data must take place on a strict need-to-know basis. Consequently, these datasets must be appropriately safeguarded with commonly acceptable access policies.

Prior to collecting any data, the research consortium will decide and agree on the tasks, roles, responsibilities and rights relating to data collection, dataset management and data use.

During the project, research datasets will be available only to those project partners or project consortium members who have been accredited and their data usage has been approved by the Project Coordinator or an authorized project consortium member. Project partners will be responsible for curating, preserving, disseminating and deleting in appropriate manner the datasets in their possession. Retention time for curated datasets will be the same as for other project results at the project consortium partners.

Data collected or acquired within the project will be stored in a secure IT environment (e.g. behind a firewall) at a Project consortium partner's premises or in secure cloud environment provided by the Project consortium partner's IT service provider. Access to it will need registration and authentication according to generally accepted best practises. Where access is granted to research data, this will be provided through a physically and virtually secure telecommunications network.

Long-term and secure preservation of published research data will be ensured by using only certified and OpenAIRE guideline[19] compatible repositories, such as Zenodo[17].

The number of copies should be minimized. Therefore, it is advisable to access the data from a server that is protected by up-to-date and state-of-the-art endpoint security, not leaving unnecessary copies to an endpoint device itself. Communications between the server and the endpoint device should be protected by secure protocols if there is a chance that the data would pass through a publicly visible network[20].

If a copy of personal data *must* be stored locally, it is advisable to use secure storages, such as Veracrypt[21], AES encrypted zip files or secure environments, with strong passwords. If a laptop or other mobile computing unit is used, an encrypted hard drive is a bare minimum.

[19] OpenAIRE Guides. Accessed in June 2019 at https://www.openaire.eu/guides

[20] E.g. Wi-Fi networks with shared credentials may disclose transmitted data to all users who have same credentials.

[21] https://www.veracrypt.fr

In order to prevent unauthorized access, modification, replication or destruction of the project's data, a number of measures must be implemented. Those include:

- Identification security: Data is stored in online repositories, which are password protected and/or access is granted only upon successful identification. Different layers of security are implemented according to the sensitivity of the data (users' personal data, etc.)
- Location security: Access to the premises of the partners, where the project is being developed, is restricted.
- Workstation security: People working on the project are encouraged to remain protected against a possible data breach by password protecting all computers and using an up to date antivirus software. Additionally, the sharing of confidential information via email is highly discouraged.

All data use in the project will be regularly backed up and, in most cases, will be residing on cloud storage facilities, preventing this way the possibility of loss of data due to hardware failure.

To enforce data security, the Project Coordinator will remind in written respective partners about the above-mentioned measures, as well as data minimisation, whenever new user studies are launched in the project.

4.2 Platforms used for storing the data

The subsequent versions of this document will define platforms for storing and managing the data generated, OpenAIRE for linking the databases and publications and one of the following repositories for the actual data:

- CERN's Zenodo[17] (HELIOS preference)
- CSC's IDA[18]
- public repository of the UAB for the academic publications (https://ddd.uab.cat).

In addition, the project will disseminate, through its website and the social media, the public data. Internal documents and confidential deliverables can be accessed at the project's workspace.

4.3 Processing personal data

Processing personal data will always include legal obligations in a role of either a data controller or a data processor. These roles are defined in the GDPR documentation[9].

The owners of the data (data subjects) have rights to it. For instance, if they want to have their data deleted, procedures must be set to do it immediately. In addition, reports on their data must be delivered upon a request. Additional details can be found at the project's public deliverable D2.10, "Recruitment and Informed Consent Procedures".

Delivering personal data records to another entity, even within the project, must be done under a written agreement that complies with the GDPR[22].

[22] When the grounds of Art. 17.1 and 3 GDPR apply.

Description of a user consent must follow the data, defining the purpose of processing the data. GDPR compliance must be demonstrated by maintaining **documentation** of personal data, including records of given consents.

Processing personal data must be done only on computing equipment that is under control of a project partner entity (or a subcontractor), protected by up-to-date and state-of-the-art endpoint security.

4.4 User profiling

Profiling is defined in Article 4.4 GDPR as ”any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person’s performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements.” It consists of:[23]

- an *automated* form of processing;
- carried out *on personal data*, data collection including *sensitive* characteristics;
- the objective of the profiling must be *to evaluate personal aspects* about a natural person, which means identifying patterns for present and future behaviours.

Evaluating personal aspects involves making statistical deductions to predict behaviours of a natural person. When the purpose includes assessing individual characteristics, then profiling is going on.[24] Depending on how the collected data is used, the patterns can lead to automated decision-making, which not necessarily leads to profiling unless monitored habits are linked to an individual. Both can be linked and may overlap. According to Article 22 of the GDPR, data subjects have the right not to be subjected only to automated-processing, producing legal effects or significantly affecting the person. Exceptions of Article 22.2 shall be considered.

In the studies of HELIOS, there may be elements of user profiling: In the development of a content-aware social graph, in WP4 task 4.8, the aim is to aggregate multimedia content consumed and exchanged by end users, followed by semantic spatio-temporal analysis with the goal of building content-aware social graphs. Content consumption preferences and patterns related to multilevel contexts will be exploited to reveal implicit links between users and their connections, as well as between users that are not explicitly connected (but could potentially be highly similar in terms of preferences, tastes and goals). The outcome of this task will enable a more nuanced semantic profiling of users and will make it possible to map of their diverse “identities” based on contextual attributes, enabling the construction and mining of multiple “local” social graphs (ego-networks). Local social graphs and profiling will be leveraged to help to dynamically filter content and avoid overload.

In order for HELIOS developers to comply with profiling and automated decision making, they must comply with the principles of processing: lawful, fair and transparent processing, complying with data

[23] WP29 Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679, 3 October 2017, p. 6.

[24] WP29 Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679, 3 October 2017, p. 7.

minimisation principles, The legal basis for profiling and automated decision-making rely in the consent of the individual,[25] who needs to be informed[26] and be aware of the consequences of profiling, and/or on the legitimate interest pursued by the controller, authorised by the EU safeguarding data subject´s rights and freedoms.[27] Taking into account the use of public data obtained from data sets and scientific research purposes, Article 89 GDPR will be applicable. HELIOS design fulfils a legitimate purpose.

The safeguards to be applied, if profiling were to be done not utilising public data sets where individuals are identified, shall include a way for data subjects to acquire human intervention, express their point of view, challenge the decision and have simple access to exercise these rights.[28] Such specifications are detailed in the consent forms[29] provided to participants in the development of validation activities in WP7. This WP does not entail profiling as such. Solely automated decision-making is not recommended and must be specifically brought to the attention of the Ethical Manager of HELIOS. Minors will not be involved in the project.

Those HELIOS partners that process personal data shall be aware of the risks of profiling. Profiling presents risks to fundamental or human rights and freedoms because profiling may lead to inaccuracy, discrimination, be unfair and produce legal effects.[30]

Examples of rights affected are “the freedom to associate with others, vote in an election, or take legal action. A legal effect may also be something that affects a person’s legal status or their rights under a contract.”[31] Such effects may have a (significant) impact in the life of a person e.g. the refusal of a credit application, the termination of a contract, denial or entitlement of social benefits, or denial or granting of a residence permit. ”For data processing to significantly affect someone the effects of the processing must be sufficiently great or important to be worthy of attention,” potentially[31]

- affecting the circumstances, behaviour or choices of the individuals concerned;
- having a prolonged or permanent impact on the data subject; or
- leading to the exclusion or discrimination of individuals.

However, the profiling measures utilised in the development of the HELIOS platform under scientific research purposes will not reach a level of dissemination where data of any participant could face such a risk.

[25] Article 6.1.a) GDPR.

[26] Article 13 and 14 GDPR.

[27] Article 6.1.f), 22.2.b) and Recital 71 GDPR.

[28] WP29 Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679, 3 October 2017, p. 27.

[29] HELIOS Deliverable D2.11 “Informed consent forms and information sheets” (M4)

[30] WP29 Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679, 3 October 2017, p. 10.

[31] WP29 Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679, 3 October 2017, p. 21.

4.5 Instructions when suspecting any unauthorized use of data

If any personal data or a copy of it is breached, e.g. by losing a USB memory stick, or detecting an unauthorized access to a server, the incident MUST BE REPORTED immediately, WITHIN 72 HOURS at latest, to the supervisory authority of the respective EU Member State.

All individuals, whose rights or freedoms are at a *high risk* due to the breach or due to profiling against their consent, must be contacted; contact immediately everyone whose data was there. Therefore, it is highly advisable to minimise any personal data in the data sets.

Furthermore, if there are any reasons to suspect that the data has been used against users' consent, the users, the Project Coordinator and the Steering Committee must be informed promptly, preferably in this order.

5. Ethical aspects

Privacy of the project participants and persons involved will be secured by following closely all the relevant EU General Data Protection Regulations. No person or organisation involved will be unintentionally identifiable directly or indirectly in the datasets. Besides storing separately from the data, all direct identifiers of any respondents or subjects (e.g. names and contact information of persons and organisations) - also indirect references to e.g. lines of businesses, branches or industries - will be removed and destroyed after the anonymised dataset has been checked and validated.

Research integrity and ethical principles related to data collection and use are covered in detail in the ethics self-assessment section of the grant application. According to guidelines set by the Ethical Manager of the project, ethics review is not required for the project as sensitive personal data will not be collected or handled within the project.

Ethical requirements have also been covered in other deliverables of the project, such as a public deliverable D2.10, "Recruitment and Informed Consent Procedures".

6. Updating this document

This is the update of the Data Management Plan (M18), originating from the initial plan (M6) and a final version (M28). Up-to-date version with possible smaller updates will be available for project members in project document repository.

Annex: Dataset descriptions (June 24th, 2020)

The table below contains a list of HELIOS Datasets published of planned by June 24th, 2020:

Responsible partner	Dataset name	URL (if known)	Availability date (est.)	How the data is generated [equipment, software, research infra…]	Expected size	Data format (JSON, CSV, etc)	Description (more details in the dataset location)
TCD	TBD	TBD	2020	TBD	TBD	TBD	NLP for user interest prediction (training set)
UPV	Social Audio Messages in-the-Wild (SAMEW)	TBD	July 2021	Previously recorded social media audios	10 MB	CSV	Social Audio Messages in-the-Wild (SAMEW) dataset is a spontaneous speech dataset which contains 1000 audio messages up to one minute collected from real WhatsApp conversations of 100 Spanish speakers, gender balanced. Voice messages were produced in-the-wild conditions before participants were recruited, avoiding any conscious bias due to laboratory environment. Samples were labelled by three evaluators in terms of valence and arousal using the Self-Assessment Manikin (SAM) procedure. SAMEW dataset includes acoustic properties for each file including temporal, spectral and cepstral domain features, as well as text transcription and emotional labelling
VTT	Message reaction time	TBD	2020	Lab test equipment		TBD	IF lab tests are possible with actual users, the test data for message reaction times *might* be usable if a) it can be totally anonymous and b) there is enough data.
CERTH	PIPD2020 - Pinterest Interests Profiling Dataset 2020	https://zenodo.org/record/3895162#.XuhqJUUzaUk	15.06.2020	EfficientNet-B3	1.7 GB	HDF5 & JSON	This dataset contains features extracted with EfficientNet-B3 from over 60000 images classified in 15 interests categories 4 of which are further analyzed in subcategories, features from the images of 12 users with hand-labeled Pinboards, the text extracted from the previous images and over 200000 features from the images of 422 Pinterest users along with their follower-followee relationships.
CERTH	Feature Dataset of Centralized & Decentralized Communication Apps	https://zenodo.org/record/3903184#.XvCULmgzaUk	22.6.2020	manually	17 kB	CSV	This dataset contains features for 38 centralized and decentralized communication applications. In total, 78 different features identified in these 38 applications.