Scaling Up secure Processing, Anonymization and generation of Health Data for EU cross border collaborative research and Innovation

D1.6 — Data Management Plan

Funded by
the European Union

Grant Agreement Nr. 10109571

Project Information

Project Title	Scaling Up Secure Processing, Anonymization and Generation of Health Data for EU Cross Border Collaborative Research and Innovation		
Project Acronym	SECURED	**Project No.**	10109571
Start Date	01 January 2023	**Project Duration**	36 months
Project Website	https://secured-project.eu/		

Project Partners

Num.	Partner Name	Short Name	Country
1 (C)	Universiteit van Amsterdam	UvA	NL
2	Erasmus Universitair Medisch Centrum Rotterdam	EMC	NL
3	Budapesti Muszaki Es Gazdasagtudomanyi Egyetem	BME	HU
4	ATOS Spain SA	ATOS	ES
5	NXP Semiconductors Belgium NV	NXP	BE
6	THALES SIX GTS France SAS	THALES	FR
7	Barcelona Supercomputing Center Centro Nacional De Supercomputacion	BSC CNS	ES
8	Fundacion Para La Investigacion Biomedica Hospital Infantil Universitario Nino Jesus	HNJ	ES
9	Katholieke Universiteit Leuven	KUL	BE
10	Erevnitiko Panepistimiako Institouto Systimaton Epikoinonion Kai Ypolgiston-emp	ICCS	EL
11	Athina-Erevnitiko Kentro Kainotomias Stis Technologies Tis Pliroforias, Ton Epikoinonion Kai Tis Gnosis	ISI	EL
12	University College Cork - National University of Ireland, Cork	UCC	IE
13	Università Degli Studi di Sassari	UNISS	IT
14	Semmelweis Egyetem	SEM	HU
15	Fundacio Institut De Recerca Contra La Leucemia Josep Carreras	JCLRI	ES
16	Catalink Limited	CTL	CY
17	Circular Economy Foundation	CEF	BE

Project Coordinator: Francesco Regazzoni - University of Amsterdam - Amsterdam, The Netherlands

Copyright

Disclaimer

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the Health and Digital Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Deliverable Information

Workpackage	WP1
Workpakace Leader	Francesco Regazzoni (UvA)
Deliverable No.	D1.6
Deliverable Title	Data Management Plan
Lead Beneficiary	UvA
Type of Deliverable	Report
Dissemination Level	Public
Due Date	30/06/2023

Document Information

Delivery Date	23/03/2023
No. pages	27
Version | Status	1.0 | Final
Deliverable Leader	Francesco Regazzoni (UvA)
Internal Reviewer #1	Stephane Lorin (THALES)
Internal Reviewer #2	Nikolaos Bakalos (ICCS)

Quality Control

Approved by Internal Reviewer #1	27/07/2023
Approved by Internal Reviewer #2	31/07/2023
Approved by Workpackage Leader	31/07/2023
Approved by Quality Manager	31/07/2023
Approved by Project Coordinator	31/07/2023

List of Authors

Name(s)	Partner
Francesco Regazzoni	UvA
Paolo Palmieri	UCC
Daniela Spajic	KUL

The list of authors reflects the major contributors to the activity described in the document. The list of authors does not imply any claim of ownership on the Intellectual Properties described in this document. The authors and the publishers make no expressed or implied warranty of any kind and assume no responsibilities for errors or omissions. No liability is assumed for incidental or consequential damages in connection with or arising out of the use of the information contained in this document.

Revision History

Date	Ver.	Author(s)	Summary of main changes
10.06.2023	0.1	Francesco Regazzoni (UvA)	Created the document and the initial version of its content based on the questionnaires
15.06.2023	0.2	Francesco Regazzoni (UvA)	Added the ethical section prepared by Daniela Spajic
28.06.2023	0.3	Francesco Regazzoni (UvA)	Added section 5, 6, 7
3.07.2023	0.4	Francesco Regazzoni (UvA)	First version finalization
27.07.2023	0.5	Francesco Regazzoni (UvA)	Updated including the comment of the first reviewer
31.07.2023	1.0	Francesco Regazzoni (UvA)	Updated including the comment of the second reviewer and comments from Daniela Spajic

Table of Contents

1 Executive Summary

Data management is a fundamental part of the activities of the SECURED project. This task has to be carried out ensuring the full compliance with ethics and privacy protection regulations and best practice to ensure the security and availability of the data and results during the project and beyond. This deliverable, D1.6 Data Management Plan (DMP) contains the summary of the data as envisioned at M6 of the project, and the foreseen procedures that will be put in place to ensure the compliance with FAIR (findability, accessibility, interoperability, and reusability) data requirements, discussing also the way in which the associated metadata will be created and stored. The document also addresses the security and the ethical aspects regarding the handling of the data that are relevant to the project. The DMP is intended as a live document, that will be updated when needed and certainly as soon as new datasets will be identified and produced.

1.1 Related Documents

- Deliverable D1.1 “Project Handbook Quality, Management”
- Deliverable D1.2 “GDPR and Ethics Project Guideline”
- Deliverable D1.9 “Dissemination and Exploitation Plan”

2 Introduction

This document provides the initial version of the data management plan for the SECURED project. On the one side, it describes the initial dataset and their characteristics, as currently envisioned at the beginning of the project. On the other, it presents the methods and the measures that will be used to securely handle and manage the data collected and processed during the project and the approaches that are currently planned to be used for facilitating the reuse of these data (when possible), in adherence with the FAIR principles and in compliance with the applicable local and international regulations.

The FAIR Principles [1] are a set of guiding principles aimed at making data and other research outputs Findable, Accessible, Interoperable, and Reusable. First described in 2016, these principles have rapidly come to define best practice in the management of research data, and much attention has been focused internationally on detailing what these principles mean in practice and determining how to assess the quality of their implementation. The European Commission supports the FAIR principles through the expert group on FAIR, which produced the *Turning FAIR into Reality*, the final report and action plan from the European Commission expert group on FAIR data [2], and through the European Open Science Cloud (EOSC) FAIR working group[1].

The FAIR principle of data usage requires that the data are findabile, accessibile, interoperabile, and reusabile, and their definition is reported in Appendix A

In accordance to the related guidelines, this document discusses the dataset that will be produced during the SECURED project, the approaches that will be followed to make these data FAIR, the resources that will be allocated to manage the data during the project and beyond, the measures that will be put in place to ensure the secure and reliable handling and storage of the data, and the related ethical implications.

The data management information have been collected from the consortium partners through a questionnaire. The questionnaire was directly derived from the template provided by the European Commission and it is reported the Appendix. The answers to the questionnaire are maintained in the internal project repository.

This version of the Data Management Plan provides a summary of the dataset and data management procedure at the begin of the project. Nevertheless, it has to be intended as a living document, which is constantly updated as the project proceeds. In particular, an updated version of the Data Management Plan will be released at M19.

2.1 Structure of the Document

The next sections presents the Data Management Plan of the SECURED project, following the template provided by the European Commission. Section 3 summarizes the envisioned dataset of the SECURED project. Section 4 reports the procedures that will be followed to make these data FAIR. Section 5 quickly discusses how other research outputs will be managed. Section 6 describes which resources will be allocated to managing the data. Section 7 describes the practices that will be followed to ensure security and reliability of the data. Section 8 discusses the ethical implications of the data used in the SECURED project. The Appendix reports the questionnaire send the the project partners to collect the information related to datasets and data management.

[1] https://www.eoscsecretariat.eu/working-groups/fair-working-group

3 Data Summary

In this section we will report the initial list of datasets as foreseen at the beginning of the project. Since the project is in its initial phases, the current list is, in most of the cases, a rough summary of the envisioned datasets (and their characteristics) that are very likely to be used and produced during the project and during the developing of the pilots. However, despite the early stages of the project, for some of these datasets, the ones that are already existing and that are serving as starting point, it is already possible to give a more precise picture of their sizes and characteristics.

The list of datasets and their characteristics will be constantly updated during the development of the project, providing a more detailed description of the feature of each datasets as soon as available, adding additional dataset when needed, and removing datasets that will not be used as expected in the course of the project.

Dataset Name	DS1
Main Involved Partner	EMC
Dataset Description	Raw functional-ultrasound (fUS) images (probably 85% of the EMC data)
Task and WP	T5.2 of WP5
Re-use existing data	Yes. Existing fUS and MRI data will be used for training machine-learning (ML) algorithms. Data is collected (outside the scope of SECURED) during regular care and informed consent is asked from the patients to reuse this information for research purposes.
Confidentiality level	Most likely Confidential or Restricted
Format	Raw and processed digital data generated by specific systems. The raw files will have formats specific for each system. Data will be analyzed using appropriate software, mostly customized scripts written in Matlab and Python.
Purpose	The purpose is to demonstrate the applicability of certain SECURED techniques (e.g., anonymization) for Use Case 1 on Real-time tumor classification.
Size	Currently estimated at 5 GB of prerecorded data per patient but subject to change in the coming year as the exact data streams to be used are finalized. Also, sample size of 20 patients is estimated.
Origin	The generated data can be both from pre-surgery scans and intra-surgery ones.
Data Utility Outside the project	AI researchers, neurosurgeons, ultrasound experts

Table 4 – Dataset DS1 Description.

Dataset Name	DS2
Main Involved Partner	EMC
Dataset Description	Processed images (probably 10% of the EMC data)
Task and WP	Same of Dataset D1
Re-use existing data	Same of Dataset D1
Confidentiality level	Same of Dataset D1
Format	Processed files will be timestamped and linked to the raw data stored.
Purpose	Same of Dataset D1
Size	Same of Dataset D1

Dataset Name	DS2
Origin	Same of Dataset D1
Data Utility Outside the project	Same of Dataset D1

Table 5 – Dataset DS2 Description.

Dataset Name	DS3
Main Involved Partner	EMC
Dataset Description	Analyzed imagery data (probably 5% of the EMC data)
Task and WP	Same of Dataset D1
Re-use existing data	Same of Dataset D1
Confidentiality level	Same of Dataset D1
Format	Statistical analysis will be performed using tools like SPSS and R.
Purpose	Same of Dataset D1
Size	Same of Dataset D1
Origin	Same of Dataset D1
Data Utility Outside the project	Same of Dataset D1

Table 6 – Dataset DS3 Description.

Dataset Name	DS4
Main Involved Partner	BME
Dataset Description	Dataset derived from publicly available medical data (only if necessary, and in full compliance with its terms of use and ethical guidelines and data protection standards and regulations).
Task and WP	T3.2 of WP3 and T2.3 of WP2
Re-use existing data	Time-series and image data, such as medical images, CTG, etc. (used only if necessary to build synthetic generative models for the purpose of anonymization.
Confidentiality level	Same as the starting data (most likely openly available)
Format	Time-series and images, perhaps tabular data.
Purpose	To generate synthetic data which preserves the general statistics of the original data for the purpose of anonymization and data augmentation
Size	Few hundreds of MBs at most.
Origin	Public data.
Data Utility Outside the project	Realistic synthetic data can further be used for the same purpose as its original counterpart.

Table 7 – Dataset DS4 Description.

Dataset Name	DS5
Main Involved Partner	ATOS
Dataset Description	Artificial or synthetic health data for testing the assets provided
Task and WP	T2.1 of WP2
Re-use existing data	When possible, existing artificial or synthetic health data will be used.
Confidentiality level	Most likely Restricted
Format	Mostly CSV / XLSX
Purpose	Generate anonymised data based on synthetic data for fulfilling GDPR regulation. Related to objectives in WP2: O2.1: Research and develop data anonymization techniques for cross-border data sharing; O2.2: Research and develop synthetic-data generation techniques for data anonymization and data augmentation
Size	> 5 MB
Origin	Artificial data for testing asset. Re-used synthetic data from T2.2
Data Utility Outside the project	Researchers, health governmental bodies, pharma or insurance private companies.

Table 8 – Dataset DS5 Description.

Dataset Name	DS6
Main Involved Partner	HNJ
Dataset Description	Records of clinical parameters of patients generated by different sensors at the patient's bed (home and hospital), collected by a hub and sent to be stored in a server at the hospital's premises.
Task and WP	T5.3 of WP5
Re-use existing data	We are starting off by using previously generated data that have only been used from clinicians to assess the evolution of the patient.
Confidentiality level	Most likely Confidential or Restricted
Format	The data are collected and stored in the form of records of pairs clinical parameter/value, together with an ad hoc id and a time stamp
Purpose	Data will be used in the context of WP5 to test the tools produced by the project
Size	Hundreds of millions of records
Origin	We have currently a set of 40 million of records. We expect to generate an order of magnitude more during the time of the project
Data Utility Outside the project	The ultimate purpose is to increase patients' safety. This high-density clinical data will help create models to make predictions about the future evolution of the health conditions of patients particularly in hospitals ICUs and when they are monitored at their homes.

Table 9 – Dataset DS6 Description.

Dataset Name	DS7
Main Involved Partner	SEM
Dataset Description	Records and scanned images at different zoom levels
Task and WP	Task 5.4 of WP5
Re-use existing data	We plan reusing data and we plan to reuse them for education and research.
Confidentiality level	Most likely Confidential or Restricted
Format	Images, time series, tables, natural language texts
Purpose	Education: synthetic data generation for education; Research: provide aggregated/anonymous data for research in project based education
Size	Images: 100Gb-10Tb; time series: 1-10Gb; tables: 10-100Gb; text: 1Gb
Origin	Origin: real cases from electronic health records Provenance:patient ID, date of data access/generation, ID of the algorithm/docker image that was used
Data Utility Outside the project	Teachers, researchers and students in health care

Table 10 – Dataset DS7 Description.

Dataset Name	DS8
Main Involved Partner	JCLRI
Dataset Description	We will collect and process publicly available genetic, clinical and lifestyle data from individuals from different Genome Wide Association Studies (GWAS) or that are part of the UK BioBank.
Task and WP	Task 5.5 in WP5
Re-use existing data	All the data will be re-used, as it has been collected by third parties who made it available via dbGaP or the UK BioBank
Confidentiality level	Most likely Confidential or Restricted
Format	The types of data include: Genotyping arrays (in plink format); Clinical data (in tabular or plink format); Other meta-data regarding the lifestyles of the participants (in tabular format)
Purpose	The purpose is to test the use of the SECURED framework to analyze protected genetic data in the context of quantifying genetic risk as well as identifying germline variants associated with different phenotypes
Size	The data is approximately 10 Tb
Origin	The origin are public databases such as dbGaP or the UK BioBank
Data Utility Outside the project	The data will be useful for scientists studying cancer genetics

Table 11 – Dataset DS8 Description.

The following dataset includes all the administrative data needed for managing the project. The will be stored and maintained according to the relevant regulation on a secure storage server offered by the project coordinator's premises and will be deleted at the end of the project.

Dataset Name	DA
Main Involved Partner	UvA
Dataset Description	This is the Consortium internal directory. It identifies the Partner Name, PI / Task Leader name, Email, role in the project.
Task and WP	All Tasks and and WP
Re-use existing data	Not Applicable
Confidentiality level	Confidential
Format	The data are collected and stored in the form of .xlsx in the UvA research drive.
Purpose	The purpose is to establish segmented and organized mailing lists for each task, technical achievements, leadership groups, and other emerging necessities.
Size	<200 KB
Origin	Administrative data from partners/collaborators.
Data Utility Outside the project	These data will not be used outside of the project.

Table 12 – Dataset DA Description.

We envision that other datasets could be generated in the course of the project, also in relatetion with the open call foreseen in the SECURED project [2]. Once identified, these datasets will be promptly included in the Data Management Plan.

[2] At the end of the second year of the SECURED project, it is planned an open call where SMEs, innovators and researchers will be invited to experiment with the technologies under development

4 FAIR data

This section describes the measures that will be put in place to make the data produced and used in the SECURED project adhere to the FAIR data principles of the European Commission. The FAIR data principles require that data are Findable, Accessible, Interoperable, and Reusable. The SECURED project will deal with at least three different families of data characterized by their level of accessibility. In particular, we will have:

- **data openly accessible** these family of data include data that are generated within the SECURED project and that will be made available, but also data that are already openly available and that will be reused during the project
- **confidential data** that can be accessed only by consortium partners
- **restricted data** that are accessible only by a restricted amount of people within the consortium.

A decision on the access level of each data set will be finalized in the course of the project. SECURED will apply the FAIR data principles to the dataset generated and used in the project, regardless of the level of the accessibility described above. In the rest of this section, we will describe how each of the requirements of the FAIR principle will be addressed during SECURED.

4.1 Making data findable

Each data used in the project will be named using a consistent naming convention and version tracking allowing to quickly identify the data and the version. The naming convention will include the following elements:

- The project name, **SECURED**
- A unique dataset name, following the format **DSn**, where N is the unique dataset number assigned in the project
- The date of the creation of each dataset, using the format YYYYMMDD
- The version number

When compliant with the access policies of the specific dataset, public repositories with digital object identifiers (such as Zenodo [3]) will be selected to store the data. Datasets will be accompanied by metadata, stored in text files, that describe the key aspects of the dataset and would simplify the access and the location of the data. The secured project is currently considering to use the DataCite [4] metadata or the Dublin Core [5] as format for metadata. To optimize the possibility of discover and indexing the data and to maximize the potential reuse, search keywords will be provided together the metadata (and, when applicable, when data will be uploaded in public repositories).

4.2 Making data accessible

SECURED is committed to make results, finding and data publicly available as much as possible. After a careful evaluation of possible conflicts with confidentiality, intellectual property protection, scientific results and findings will be disseminated to the relevant scientific communities and to stakeholders by means of public presentations, tutorial, and webinars and the related publications will be made available on open access repositories.

Datasets (and related metadata) that, according to the restrictions and the specific access policies, can be made openly accessible will be made either openly available to the community using public repositories such

[3] https://zenodo.org
[4] https://schema.datacite.org
[5] https://dublincore.org/

as Zenodo or provided only upon request (this could be the case where metadata could be made available to the public, but data could be provided only after specific agreements). Data that are confidential and that can thus be accessed only by consortium members will be stored securely and made available to the consortium partners using the research drive platform offered by Surf [6] and provided to the consortium by the coordinating entity. The research drive platform for the SECURED project is managed by the coordinator and can be accessed only by members of the consortium identified by credentials. The list of account is maintained and regularly updated to ensure that access to data is granted only to authorized consortium members. Release on public repositories or catalogues (such as Zenodo or GWAS [7]) of the metadata associated with these confidential data or the release of part of these data will be considered and carefully examined during the development of the project, and a final decision will be taken on a case-by-case basis, ensuring that the decision do not negatively affect the intellectual properly protection and is full compliant with applicable regulations. Finally, data that are accessible only by a restricted number of people will be shared following the procedure indicated by the partner that owns them (typically, they will stay at the partner location and the access will be protected using state of the art authentication).

Information related to data used in scientific publication will be timely made available at the moment of the publication. Similarly, where access policies allow to do so, also data will be made available at the moment of the publication. Where an embargo period is needed, data will be prepared for publication and temporary stored on dedicated folders of the SECURED Research Drive platform till the end of the embargo.

To ensure the widest possible access to the data, the SECURED project plans make data available in standard file formats (csv, txt, pdf, ...). These formats of files can be accessed using regular software commodities commonly available (also as opensource). Where needed, a detailed list of software needed to access the data will be provided.

4.3 Making data interoperable

The SECURED project is committed to make the data created in the project interoperable and to allow data exchange and reuse across disciplines. To achieve this goal, data and metadata vocabularies, formats and standards that will be used will be carefully selected following interoperability best practice. The specific format of the used data depends from the application domain of the pilot. For some of the pilots, the data format has been already defined. For instance, for the neuro-imaging data, it has been adopted the NIfTI data format [8] which is a very popular format for the relevant community and recognized to be interoperable. Also, when existing data will be reused, we will maintain the same standard as of the original data. The exact format of the data, metadata and ontologies has however not yet been defined for all the data. Nevertheless, possible candidates have been identified. Among them, there are health interoperability standards such as HL7/FHIR [9] or SNOMED [10] concerning the data, and Phenotype Ontology [11] or the Disease Ontology [12] concerning ontologies. Since it is possible that several data format needs to be supported, we are also considering the possibility to create a common semantic model for project data, that might include project specific ontologies, but that will also provide the appropriate documentation for accessing it and traversing it, and a mapping to commonly used ontologies more specific of each domain.

4.4 Increase data re-use

Reuse of the data used and produced in the project will be facilitate by the production and maintenance of the appropriated documentation. Data will be accompanied by readme files containing detailed information and explanation about them, such as unit of measure used, process followed for the creation, provenience, tools

[6] https://wiki.surfnet.nl/display/RDRIVE/Research+Drive
[7] https://www.ebi.ac.uk/gwas/
[8] https://nifti.nimh.nih.gov
[9] https://www.hl7.org/fhir/
[10] https://www.nlm.nih.gov/healthit/snomedct/index.html
[11] https://hpo.jax.org/app/
[12] https://disease-ontology.org/

and procedures used for the analysis, etc. The documentation material will be stored in repositories supporting versioning together with the data that it describes. A quality assurance procedure to ensure that the data are complete and accompanied by the proper documentation will be put in place and supervised by the SECURED quality manager Dr. Paolo Palmieri.

In this phase of the project, the decision about making pilots data freely available to third parties has not been finalized. Certain data are confidential or restricted (for instance, certain patient data), and their confidentiality level will be maintained. However, the possibility of making the data available and to do so with standard reuse licenses will be certainly carefully considered and evaluated as soon as the data will be created. This could be, for instance, the case of generated (e.g., derivative) data during the project, provided the sharing of such data does not pose any privacy risk (e.g. for the owners of the original data), the case of final summary data, where allowed.

5 Other research outputs

Any other research output produced and generated within the SECURED project will be managed according to the FAIR principles and, if possible, openly released. Software, workflows and models developed in the project will be accompanied by the necessary documentation and will be maintained in repositories supporting versioning and suitable to enforce the needed access permission. Adherence to the FAIR principles will be also requested to participants to the SECURED open call.

Interoperability will be also considered during the design phase of the components of the SECURED project and the SECURED Innohub. Software libraries, in particular, will be designed to enable interoperability with existing libraries and software and re-use to the maximum extent possible. Similarly, where possible, hardware components will be designed using standard interfaces to facilitate their reuse.

In the current phase of the project, which is mostly focused on requirements collection, use case formalization, and architecture definition, the implementation phase has not been started yet. Thus, at this stage of the project, it has not been yet clearly decided which elements of the SECURED project will be released openly. An in depth discussion about releasing and licensing of the output of the SECURED project, including the exploitation plan of each partner, will be subject of the deliverables D1.4 “Market Analysis, IPR Management Exploitation and Standardization” (due at M18) and D1.5 “Updated Market Analysis, IPR Management Exploitation and Standardization” (due at M36) and will be also addressed in updated version of the Data Management Plan D1.8 (due at M19).

6 Allocation of resources

Datasets will be stored at different locations, including servers located at the respective partner's premises, the internal project repositories maintained by the project coordinator entity and openly available repositories such as Zenodo. Costs for setting up and maintaining the infrastructure needed for securely storing the data include disk-storage (and backup), hardware firewall and VPN service, computer upkeep, network and compute-cluster maintenance. If compliant with the Grant Agreement conditions, costs for making data or other research outputs FAIR in the SECURED project are eligible as part of the Horizon Europe grant. Where applicable, each partner is responsible for claiming these costs.

Each partner is responsible for preparing the datasets produced during its own activities and to ensure their compliance with all the applicable local and international guidelines, regulations and laws. The Project Coordinator Dr. Francesco Regazzoni (UvA) with the support of the Quality Manager Dr. Paolo Palmieri and the Dissemination Manager Dr. Apostolos Fournaris (ISI) will timely update the data management plan in the SECURED project accordingly.

Where applicable, partners that manage data at their premises will identify an internal person responsible for managing the data in adherence to the SECURED Data Management Plan. A list including an overview of the responsible contact persons for each institution will be maintained in the project repository, together with the other administrative data of the SECURED project.

7 Data security

Security of the data handled and stored during the SECURED project is of utmost importance. SECURED is fully committed that the data used in the project will remain confidential and accessible. This section highlights the measures and the guidelines that will be used to ensure this.

In addition to complying with the relevant local, EU, and international applicable legislations (such as the GDPR), all partners will adhere to the state of the art best practices for the security and back up of the data. As mentioned before, the internal data of the consortium will be safely and securely stored on the Research Drive platform of Surf and provided to the consortium by UvA. The platform can be accessed only by authenticated members of the consortium, ensuring proper access control and security. All members of the consortium who need access must submit a request for the creation of an account to the project coordinator (who is also in charge of managing the correct access rights and maintaining the list updated). Datasets that will be maintained at the partners' premises will be securely stored and maintained in internal repositories under the responsibility of the owning partner. In case a transfer of data will be required, VPNs or encrypted communication will be used.

Strict adherence with institutional IT regulations, best practices and common security measures are recommended to all the partners to ensure the security of the data and increase consistent standards within the consortium. Common security measures recommended by SECURED include the following guidelines:

- To avoid loss of data, use appropriate redundancy (e.g. storing data in at least two separate locations, use of NAS and RAID solutions).
- To secure data at rest, use data encryption solutions wherever possible. Device-level encryption should be the preferred choice.
- To prevent data loss and potential breaches, limit to the maximum feasible extent the use of portable USB devices, and always use encryption on such devices when their use is unavoidable.
- To prevent unathorized access, apply appropriate partner- and consortium-based authentication and access control policies and mechanisms, and consider the adoption of two-factor authentication.
- To prevent data misplacement, follow rigorously the naming convention detailed in Section 4.1.

8 Ethics

Several potential legal and ethical concerns occur concerning the collection, processing and sharing of (personal) data within the SECURED project.

Ethical issues may arise with regard to some of the objectives of the SECURED project. Firstly, the project methodology includes the processing of large amounts of sensitive data, particularly data concerning health and genetic data. Secondly, these data will be collected from vulnerable individuals, including adult patients and possibly paediatric patients, and are envisioned to be shared among consortium partners to develop the SECURED tools and Innohub platform. Thirdly, the creation and application of new data analytics techniques, such as medical synthetic data generation or federated learning, raise questions regarding the risks for re-identification of the data subject concerned.

In addition to the potential ethics issues, multiple legal requirements apply to the processing of personal data and special categories of personal data in the SECURED project and to the creation of the SECURED solutions. Personal data is defined in the General Data Protection Regulation (GDPR) as “any information relating to an identified or identifiable natural person (‘data subject’)" [13]. An identifiable natural person is someone “who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person" [14]. Special categories of personal data are data concerning health or genetic data, which are more sensitive, and, thus, receive more protection under the GDPR [15].

Addressing the legal and ethical challenges is a pivotal part of the SECURED project’s work plan. Special attention will be paid to the issues since the beginning of the project. The legal and ethical inventory, which will be provided throughout two deliverables by KU Leuven as the law and ethics partner, delivers an overview of the relevant legal and ethical framework that relate to the ethical and legal issues to offer consortium partners the necessary tools to address them. More specifically, D1.2 “GDPR and Ethics Project Guidelines" (M6) provides an overview of the essential ethics principles and legal provisions of the GDPR as the primary data protection framework applicable to data processing, including the collection and sharing of data. D2.5 “Legal and ethical framework and analysis" (M9) builds upon the first deliverable and, beyond that, also discusses other frameworks relevant to the SECURED project, including data (governance) laws, artificial intelligence and health technology frameworks. Furthermore, deliverables D5.4 “Evaluation and implementation of legal and ethical requirements” (M24) and D4.6 “Legal Validation and Recommendation report” will address the legal and ethical requirements.

For personal data processing in the context of SECURED solutions, relevant controller-processor agreements for data sharing will be concluded between the partners involved where necessary.

Anonymised data are not subject to such legal requirements [16]. Once the data has been successfully anonymised, it is no longer possible to (re-)identify the data subject. Consequently, EU law does not hinder the dissemination or further use of anonymised data provided that it is anonymised in a way that absolutely prevents the data subject from being identified. However, anonymisation is a data processing operation involving personal data, so the GDPR requirements apply before and during the processing operation. To this end, particular attention must be paid to the data minimisation and purpose limitation principle.

Open research data (which can be any data resulting from research regardless of whether it is anonymous, pseudonymous, or personal data) from the SECURED project will be anonymised before it is made publicly available. If anonymisation is not possible, they should be pseudonymised [17]. The data provider is responsible for the anonymisation or pseudonymization process and for ensuring that identifiable variables (e.g., name, email address, ID number) are not transferred.

[13] Article 4(1) GDPR.
[14] Article 4(1) GDPR.
[15] See Article 9 GDPR.
[16] See recital 26 GDPR.
[17] Article 4(5) GDPR.

(Explicit) informed consent authorizing the use of personal data will be collected from every possible patient when required. A patient information sheet will be distributed to provide the relevant information as per Articles 13 and 14 General Data Protection Regulation regarding collecting, using and sharing personal data. The participants have the right to withdraw from the research at any time without any adverse consequences.

9 Conclusions

The current version of the Data Management Plan includes a preliminary description of the envisioned datasets that will be produced in the project and their characteristics as they can be foreseen at the M6 of the project. It also provides a description of the procedure that will be put in place for ensuring adherence to the FAIR guidelines and addresses security and ethical aspects of the data handling in the project.

It must be emphasized that the current document represents a picture of the situation and the plans at the beginning of the project, where the definition of the overall requirements is still an ongoing process. Because of this, some decisions on the exact access policies to data are not finalized while it is also very likely that currently reported datasets could be subject to changes and that new datasets will be added (certainly, we expect so as outcome of the SECURED Open Call). As a result, the Data Managment Plan will be constantly and timely updated, and promptly made available to partners to ensure compliance.

References

[1] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, and B. Mons, "The fair guiding principles for scientific data management and stewardship," *Scientific Data*, vol. 3, no. 1, p. 160018, Mar 2016. [Online]. Available: https://doi.org/10.1038/sdata.2016.18

[2] European Commission and Directorate-General for Research and Innovation, *Turning FAIR data into reality: final report and action plan from the European Commission expert group on FAIR data.*, 2018, oCLC: 1111111608. [Online]. Available: http://publications.europa.eu/publication/manifestation_identifier/PUB_KI0618206ENN

A FAIR Data Principles

The FAIR principles, defined in [1] and as quoted in Turning FAIR data into reality, the final report and action plan from the European Commission expert group on FAIR data [2].

Findable

- **F1.** (Meta)data are assigned a globally unique and persistent identifier
- **F2.** Data are described with rich metadata
- **F3.** Metadata clearly and explicitly include the identifier of the data they describe
- **F4.** (Meta)data are registered or indexed in a searchable resource

Accessible

- **A1.** (Meta)data are retrievable by their identifier using a standardised communications protocol
- **A1.1.** The protocol is open, free, and universally implementable
- **A1.2.** The protocol allows for an authentication and authorisation procedure, where necessary
- **A2.** Metadata are accessible, even when the data are no longer available

Interoperable

- **I1.** (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- **I2.** (Meta)data use vocabularies that follow FAIR principles
- **I3.** (Meta)data include qualified references to other (meta)data

Reusable

- **R1.** Meta(data) are richly described with a plurality of accurate and relevant attributes
- **R1.1.** (Meta)data are released with a clear and accessible data usage license
- **R1.2.** (Meta)data are associated with detailed provenance
- **R1.3.** (Meta)data meet domain-relevant community standards

B Data Management Plan Questionnaire

B.1 Data Summary

- Will you re-use any existing data and what will you re-use it for? State the reasons if re-use of any existing data has been considered but discarded.
- What types and formats of data will the project generate or re-use?
- What is the purpose of the data generation or re-use and its relation to the objectives of the project?
- What is the expected size of the data that you intend to generate or re-use?
- What is the origin/provenance of the data, either generated or re-used?
- To whom might your data be useful ('data utility'), outside your project?

B.2 FAIR Data

B.2.1 Making data findable

- Will data be identified by a persistent identifier?
- Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.
- Will search keywords be provided in the metadata to optimize the possibility for discovery and then potential re-use?
- Will metadata be offered in such a way that it can be harvested and indexed?

B.3 Making data accessible

Repository

- Will the data be deposited in a trusted repository?
- Have you explored appropriate arrangements with the identified repository where your data will be deposited?
- Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object?

Data

- Will all data be made openly available? If certain datasets cannot be shared (or need to be shared under restricted access conditions), explain why, clearly separating legal and contractual reasons from intentional restrictions. Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if opening their data goes against their legitimate interests or other constraints as per the Grant Agreement.
- If an embargo is applied to give time to publish or seek protection of the intellectual property (e.g. patents), specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

- Will the data be accessible through a free and standardized access protocol?If there are restrictions on use, how will access be provided to the data, both during and after the end of the project?
- How will the identity of the person accessing the data be ascertained?
- Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)?

Metadata

- Will metadata be made openly available and licenced under a public domain dedication CC0, as per the Grant Agreement? If not, please clarify why. Will metadata contain information to enable the user to access the data?
- How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available?
- Will documentation or reference about any software be needed to access or read the data be included? Will it be possible to include the relevant software (e.g. in open source code)?

B.3.1 Making data interoperable

- What data and metadata vocabularies, standards, formats or methodologies will you follow to make your data interoperable to allow data exchange and re-use within and across disciplines? Will you follow community-endorsed interoperability best practices? Which ones?
- In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining or extending them?
- Will your data include qualified references [18] to other data (e.g. other data from your project, or datasets from previous research)?

B.3.2 Increase data re-use

- How will you provide documentation needed to validate data analysis and facilitate data re-use (e.g. readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, units of measurement, etc.)?
- Will your data be made freely available in the public domain to permit the widest re-use possible?
- Will your data be licensed using standard reuse licenses, in line with the obligations set out in the Grant Agreement?
- Will the data produced in the project be useable by third parties, in particular after the end of the project?
- Will the provenance of the data be thoroughly documented using the appropriate standards? Describe all relevant data quality assurance processes.
- Further to the FAIR principles, DMPs should also address research outputs other than data, and should carefully consider aspects related to the allocation of resources, data security and ethical aspects.

[18] A qualified reference is a cross-reference that explains its intent. For example, X is regulator of Y is a much more qualified reference than X is associated with Y, or X see also Y. The goal therefore is to create as many meaningful links as possible between (meta)data resources to enrich the contextual knowledge about the data. (Source: https://www.go-fair.org/fair-principles/i3-metadata-include-qualified-references-metadata/)

B.4 Other research outputs

- In addition to the management of data, beneficiaries should also consider and plan for the management of other research outputs that may be generated or re-used throughout their projects. Such outputs can be either digital (e.g. software, workflows, protocols, models, etc.) or physical (e.g. new materials, antibodies, reagents, samples, etc.).
- Beneficiaries should consider which of the questions pertaining to FAIR data above, can apply to the management of other research outputs, and should strive to provide sufficient detail on how their research outputs will be managed and shared, or made available for re-use, in line with the FAIR principles.

B.5 Allocation of resources

- What will the costs be for making data or other research outputs FAIR in your project (e.g. direct and indirect costs related to storage, archiving, re-use, security, etc.) ?
- How will these be covered? Note that costs related to research data/output management are eligible as part of the Horizon Europe grant (if compliant with the Grant Agreement conditions)
- Who will be responsible for data management in your project?
- How will long term preservation be ensured? Discuss the necessary resources to accomplish this (costs and potential value, who decides and how, what data will be kept and for how long)?

B.6 Data security

- *What provisions are or will be in place for data security (including data recovery as well as secure storage/archiving and transfer of sensitive data)?*
- *Will the data be safely stored in trusted repositories for long term preservation and curation?*

B.7 Ethical Aspects

- Are there, or could there be, any ethics or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).
- Will informed consent for data sharing and long term preservation be included in questionnaires dealing with personal data?

B.8 Other Issues

- Do you, or will you, make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones (please list and briefly describe them)?