D1.2

Data Management Plan v1

Related Work Package	**WP1 – Project Management and Coordination**
Related Task	Task 1.3 - Data Management Plan and Control
Lead Beneficiary	FORTH
Contributing Beneficiaries	All
Document version	1.0
Deliverable Type	Report
Distribution level	Public
Contractual Date of Delivery	30-06-2024
Actual Date of Delivery	01-07-2024

Authors	Nikolaos Tachos (FORTH)
Contributors	Eleni Tsalapati (ATC), Giuseppe Riccardo Leone (CNR), Evaggelia Anagnostopoloulou (ICCS), Daniel Diosdado López (VPF), Asbjorn Folstad (SINTEF), Konstadinos Marias (FORTH) Pilar Sala (AOA)
Reviewers	Dimitrios Fotiadis (FORTH)

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 Directorate-General for Communications Networks, Content and Technology. Neither the European Union nor the granting authority can be held responsible for them.

Version history

Version	Description	Date completed
0.0	Deliverable TOC	05.06.2024
0.3	Contents, Abbreviations, Introduction	11.06.2024
0.6	Datasets Description	21.06.2024
0.9	Version 1.0 ready for internal review	25.06.2024
1.0	Final version incorporating the replies to the review comments	01.07.2024

Statement of Originality

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

Disclaimer

This document contains material, which is the copyright of one or more FAITH consortium parties, and may not be reproduced or copied without permission.

All FAITH consortium parties have agreed to full publication of this document.

The commercial use of any information contained in this document may require a license from the proprietor of that information.

Neither the FAITH consortium as a whole, nor individual FAITH consortium parties, warrant that the information contained in this document is capable of use, nor that use of the information is free from risk, accepting no liability for loss or damage suffered by any person using this information.

Executive summary

This deliverable based on the DoW entitled “Data Management Plan v1” contains detailed information, based on our current analysis, about data that will be produced and collected within the project, whether and how it will be made accessible for re-use and further exploitation, and how it will be curated and preserved. The Data Management Plan (DMP) meticulously outlines the strategy for handling research data generated by the FAITH project, funded under the HORIZON EUROPE program and it provides an analytical overview of the DMP's key aspects, emphasizing its alignment with FAIR data principles and its contribution to the project's overall objectives.

The DMP identifies and describes the types of data expected throughout the project lifecycle, including primary data collected directly, secondary data obtained from external sources with proper attribution, and comprehensive metadata for discovery and interpretation. Data collection processes are clearly defined, outlining methods, tools, quality control procedures, and version control mechanisms to ensure data accuracy, integrity, and provenance. In Section 2, we provide a detailed description of the data gathered and processed for each work package (WP), as well as of the data collection, documentation, metadata generation and data assessment mechanisms developed by the project. Furthermore, the tools for data storage and methodologies for ensuring data security are also presented in this section.

Data sharing and open access are addressed in line with FAIR principles and HORIZON EUROPE guidelines. The DMP clarifies the extent of public availability after research publication, targeted data repositories for long-term accessibility, and anonymization strategies for sensitive data. In addition, it outlines strategies for data preservation beyond the project duration, including data format selection promoting long-term usability, metadata creation practices ensuring interpretability, and designation of a responsible party for long-term data curation.

This data management plan offers significant advantages. Rigorous data management practices enhance research quality by ensuring data integrity and facilitating research findings' reproducibility. Open access to FAIR data fosters collaboration, knowledge sharing, and the potential for new discoveries, increasing research impact. Subsequently, Section 3 describes how the data management plan is aligned with the FAIR principles. General Data Protection Regulation (GDPR) and its application to the FAITH project is described in section 4.

The FAITH Open Data project prioritizes open access for scientific publications, allowing researchers to choose between publishing models and It aligns with EU recommendations on open data access, ensuring wider dissemination of research results. These issues are described in Section 5. Finally, the project risks and proposed mitigation measures and related ethical aspects raised by FAITH research are provided in Section 6 and 7, respectively.

Furthermore, the DMP aligns with HORIZON EUROPE open access mandates and relevant data protection regulations, ensuring compliance since a clear data management strategy promotes transparency and accountability within the project and to the wider scientific community. The datasets described in this version of the DMP document represent the data collected or generated at this stage of the project. In subsequent versions of the DMP, more sensitive and detailed information

will be provided compared to this initial version. New data or changes in consortium policies may emerge, however, the core principles - as outlined in this deliverable - remain intact and are expected to remain until the end of the project.

Table of Contents

List of Abbreviations

Abbreviation	Explanation
AI	Artificial Intelligence
CSV	Comma-Separated Value
DCAT	Data Catalogue model
DICOM	Digital Imaging and Communications in Medicine
DMP	Data Management Plan
DoA	Description of Annex
DOI	Digital Object Identifier
FAIR	Findable, Accessible, Interoperable, Re-usable
GDPR	General Data Protection Regulation
HTTP	Hypertext Transfer Protocol
IPR	Intellectual Property Rights
JSON	JavaScript Object Notation
LSP	Large Scale Pilot
MRI	Magnetic Resonance Imaging
RDF	Resource Description Framework
REST API	Representational State Transfer Application Programming Interface
WP	Work Package
XML	Extensible Markup Language

List of Tables

List of Figures

1 Introduction

1.1 Purpose of the FAITH Data Management Plan

FAITH's aim is to deliver a human-centric, trustworthiness assessment framework (FAITH AI_TAF) which enables the testing/measuring/optimization of risks associated with AI trustworthiness in several critical domains. FAITH AI_TAF builds upon NIST Artificial Intelligence Risk Management Framework (AI RMF), upon the requirements imposed by the EU legislative instruments, upon ENISA guidelines on how to achieve trustworthiness by design and upon stakeholder's intelligence and users' engagement.

Seven (7) Large Scale Pilot (LSP) activities in seven (7) critical and diverse domains (robotics, education, media, transportation, healthcare, active ageing, and industry) will validate the FAITH holistic estimation of trustworthiness of selected sectoral AI systems. To this end, the proposed framework will be validated across two large scale piloting iterations/phases across focusing on assessing: (i) generic hreats of trustworthiness, and (ii) domain-specific threats and risks of trustworthiness. In addition, FAITH AI_TAF will be used to identify potential associations among the domains towards the development of a domain-independent, human-centric, risk management driven framework for AI trustworthiness evaluation.

FAITH's vision is to become a catalyst in this process by creating the first European, ethical- and GDPR compliant, quality-controlled repository, in which both large-scale data and AI algorithms will co-exist. The respective datasets produced, raw or processed by FAITH LSPs, will be carefully handled, under thorough consideration of ethical and privacy issues involved in such datasets. For all the identified FAITH datasets, specific parts that can be made publicly available have been identified in the current first version of the project's DMP. The DMP of FAITH realizes the data management regarding two types of data: on the one hand the utilization of the research data that are generated and collected within the context of the project, and on the other hand the dissemination of the scientific results generated from the project.

The present deliverable is developed based on the Guidelines[1] on Open Access to Scientific Publications and Research Data in Horizon Europe, as well as to the General Data Protection Regulation and is structured considering the Horizon Europe FAIR Data Management Plan[2]. According to the guidelines on Data Management in Horizon Europe, a dataset description template has been drafted to provide the main pillar for the dataset descriptions (Annex I). Relevant datasets have been identified and a detailed list has been generated, based on the datasets that have been described within the DoA to be produced during the lifetime of the project.

1.2 Background of the FAITH Data Management Plan

FAITH DMP is in accordance with the following articles of the Grant Agreement (GA):

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

[2] http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm

Article 29.2 Open access to scientific publications

Each beneficiary must ensure open access (free of charge online access for any user) to all peer-reviewed scientific publications relating to its results.

Article 29.3 Open access to research data

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) not applicable;

(iii) other data, including associated metadata, as specified and within the deadlines laid down in the '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).

Article 36 Confidentiality

During implementation of the action and for four years after the period set out in Article 3, the parties must keep confidential any data, documents or other material (in any form) that is identified as confidential at the time it is disclosed ('confidential information').

Article 39.2 Processing of personal data by the beneficiaries

The beneficiaries must process personal data under the Agreement in compliance with applicable EU and national law on data protection (including authorisations or notification requirements).

The beneficiaries may grant their personnel access only to data that is strictly necessary for implementing, managing and monitoring the Agreement.

The beneficiaries must inform the personnel whose personal data are collected and processed by the Commission. For this purpose, they must provide them with the privacy statement(s) (see above), before transmitting their data to the Commission.

2 Data description

2.1 Data Types

Different data types are collected during the course of the project, such as publications and research data, managerial and ethical documents as well as data coming from LSPs. A list of the different datasets has been established, and this list is further detailed to precise the type of data generated. Collected data includes both data already available to partners from the initiation of the project and data that are generated during the lifetime of the project.. Data formats are selected with the view to facilitate data storage and reusability. Therefore, data will be in both human-readable and machine-readable format (e.g. RDF, XLM and JSON).

2.2 Data users

The data collected and generated through the FAITH project may be exploited by a wide range of data users, including:

- Researchers
- AI model developers
- Project's partners
- Wider audience
- European Commission

2.3 FAITH Datasets

2.3.1 Datasets naming

The convention followed for naming the project datasets, is the following:

1. A prefix "DS" indicating a dataset.
2. Its unique identification number depending on the WP the dataset comes from, e.g., "DS1" for datasets coming from WP1, "DS2" for datasets coming from WP2 etc.
3. A serial number restarting at 1 for each WP indicating the sub-dataset come from the specific WP: "DS1.1", "DS1.2" etc.
4. A short name indicative of its content and purpose. e.g., "DS1.4_managerial documents".
5. If a versioning of the DS is needed then the latter in placed at the end of the naming. e.g., "DS1.4_managerial documents_v1".

2.3.2 Datasets description

In compliance with the "Dataset description template" provided in Appendix I, the following tables present in detail the information regarding the FAITH datasets, in terms of: i) generic description, ii) origin of data, iii) nature and scale of data, iv) to whom the dataset could be useful, v) related scientific publications, vi) indicative existing similar data sets, vii) partners activities and responsibilities, viii) standards and metadata, ix) data exploitation and sharing, x) archiving and preservation, xi) data security, xii) ethics

Table 1: DS1.1_Partners Contact List description.

Data identification: ***DS1.1_ Partners Contact List***	
Generic description:	
The contact details of the persons representing each partner organization in the FAITH project and participating in each WP and task. Contact details include telephone number, skype name and address.	
Origin of data:	
The data was collected at the beginning of the project and they will be updated once a change in the personnel of each organization (FAITH partner) takes place.	
Nature and scale of data:	
Spreadsheet data	
To whom the dataset could be useful:	
All partners	
Related scientific publication(s)	
N/A	
Indicative existing similar data sets (including possibilities for integration and reuse):	
N/A	
Partners activities and responsibilities	
Partner owner of the data	FAITH consortium
Partner in charge of the data analysis	FAITH consortium
Partner in charge of the data storage	FORTH
Related WP(s) and task(s)	All
Standards and metadata	
Standards, format, estimated volume of data	Excel format files
Data exploitation and sharing	
Data access policy/ Dissemination level: confidential (only for members of the Consortium and the Commission Services) or Public	Confidential (only for members of theConsortium and the Commission Services).
Data sharing, re-use, distribution, publication (How?)	Shall be limited only to be carried out between the Project Consortium members and the European Commission's Services.
Personal data protection: are they personal data? If so, have you gained (written) consent from data subjects to collect this information?	Personal data are involved. All the partners have agreed on this during the kick-off meeting of the project.
Access Procedures	None within the project consortium.
Embargo periods (if any)	None
Archiving and preservation (including storage and backup)	
Data storage (including backup): where? For how long?	On project private file repository. Shall be maintained and backed up for a period of 3years following the end of the project.

Indicative associated costs for data archiving and preservation	N/A
Indicative plan for covering the above costs	N/A
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)?	N/A
Will the data be securely stored in trusted repositories for long-term preservation and curation?	N/A
Ethics	
Describe any potential ethical issues during data collection, storage, processing and archiving, together with the ethical approval procedures related to the project.	N/A
If the research activities involve children, patients, members of vulnerable populations, use of embryonic stem cells, privacy and data protection issues or research on animals and primates, the ethical principles and relevant national, EU and international legislation must be complied with.	N/A

*Table 2: DS1.2_*Financial statements *description.*

Data identification: *DS1.2_Financial statements description*
Generic description:
The financial information of each partner of the consortium will be included in the DS1.2 dataset. They will include information about the personnel cost, the justification of travel, the justification of equipment, the justification of other goods and services, the justification of sub- contracting, the justification of linked third parties and the justification of the contributions of linked third parties.
Origin of data:
The information will be provided by each partner to the coordinator along with the interim progress report in M6, M12, M18, M24 and M36.
Nature and scale of data:
Spreadsheet data
To whom the dataset could be useful:
All partners, European Commission
Related scientific publication(s)
N/A
Indicative existing similar data sets (including possibilities for integration and reuse):

N/A	
Partners activities and responsibilities	
Partner owner of the data	FAITH consortium
Partner in charge of the data analysis	FORTH
Partner in charge of the data storage	FORTH
Related WP(s) and task(s)	WP1
Standards and metadata	
Standards, format, estimated volume of data	Excel format files
Data exploitation and sharing	
Data access policy/ Dissemination level: confidential (only for members of the Consortium and the Commission Services) or Public	Confidential (only for members of theConsortium and the Commission Services).
Data sharing, re-use, distribution, publication (How?)	Shall be limited only to be carried outbetween the Project Consortium members and the European Commission's Services.
Personal data protection: are they personal data? If so, have you gained (written) consent from data subjects to collect this information?	No personal data.
Access Procedures	None within the project consortium.
Embargo periods (if any)	None
Archiving and preservation (including storage and backup)	
Data storage (including backup): where? For how long?	On project private file repository. Shall be maintained and backed up for a period of 3years following the end of the project.
Indicative associated costs for data archiving and preservation	N/A
Indicative plan for covering the above costs	N/A
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)?	N/A
Will the data be securely stored in trusted repositories for long-term preservation and curation?	N/A
Ethics	
Describe any potential ethical issues during data collection, storage, processing and archiving, together with the ethical approval procedures related to the project.	N/A

If the research activities involve children, patients, members of vulnerable populations, use of embryonic stem cells, privacy and data protection issues or research on animals and primates, the ethical principles and relevant national, EU and international legislation must be complied with.	N/A

*Table 3: DS1.*3_RiskLog *description.*

Data identification: *DS1.3_RiskLog description*	
Generic description:	
A description of the risk, its causes, the kinds of problems that it could result in (potential effects), and risk dependencies.	
Origin of data:	
Several potential risks have been identified with direct or indirect impact on FAITH solution. Risks are grouped into four categories: a) general and administrative, b) technical and scientific, c) exploitation and dissemination, d) ethical. Each partner will estimate and evaluate the associated risks, the respective controls and will monitor the effectiveness of the controls in collaboration with RM.	
Nature and scale of data:	
Spreadsheet data. The Risk Log for the project is using PM2 Risk Log template and no changes have been done to the structure, fields or values.	
To whom the dataset could be useful:	
Executive Board and Project Core	
Related scientific publication(s)	
N/A	
Indicative existing similar data sets (including possibilities for integration and reuse):	
N/A	
Partners activities and responsibilities	
Partner owner of the data	FAITH consortium
Partner in charge of the data analysis	FORTH
Partner in charge of the data storage	FORTH
Related WP(s) and task(s)	WP1
Standards and metadata	
Standards, format, estimated volume of data	PM2 Risk Log template
Data exploitation and sharing	
Data access policy/ Dissemination level: confidential (only for members of the Consortium and the Commission Services) or Public	Confidential (only for members of theConsortium and the Commission Services).
Data sharing, re-use, distribution, publication (How?)	Shall be limited only to be carried outbetween the Project Consortium members and the European

	Commission's Services.
Personal data protection: are they personal data? If so, have you gained (written) consent from data subjects to collect this information?	No personal data.
Access Procedures	None within the project consortium.
Embargo periods (if any)	None
Archiving and preservation (including storage and backup)	
Data storage (including backup): where? For how long?	On project private file repository. Shall be maintained and backed up for a period of 3years following the end of the project.
Indicative associated costs for data archiving and preservation	N/A
Indicative plan for covering the above costs	N/A
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)?	N/A
Will the data be securely stored in trusted repositories for long-term preservation and curation?	N/A
Ethics	
Describe any potential ethical issues during data collection, storage, processing and archiving, together with the ethical approval procedures related to the project.	N/A
If the research activities involve children, patients, members of vulnerable populations, use of embryonic stem cells, privacy and data protection issues or research on animals and primates, the ethical principles and relevant national, EU and international legislation must be complied with.	N/A

*Table 4: DS1.*4_Managerial documents *description.*

Data identification: ***DS1.4_Managerial documents description***
Generic description:
Information related to the project management and coordination.
Origin of data:
The information will be collected through the whole lifecycle of the project.
Nature and scale of data:

Documents in excel, word and pdf format.	
To whom the dataset could be useful:	
All partners, European Commission.	
Related scientific publication(s)	
N/A	
Indicative existing similar data sets (including possibilities for integration and reuse):	
N/A	
Partners activities and responsibilities	
Partner owner of the data	FAITH consortium
Partner in charge of the data analysis	FORTH
Partner in charge of the data storage	FORTH
Related WP(s) and task(s)	WP1
Standards and metadata	
Standards, format, estimated volume of data	Excel format files. Word format files. PDF format files
Data exploitation and sharing	
Data access policy/ Dissemination level: confidential (only for members of the Consortium and the Commission Services) or Public	Confidential (only for members of theConsortium and the Commission Services).
Data sharing, re-use, distribution, publication (How?)	Shall be limited only to be carried outbetween the Project Consortium members and the European Commission's Services.
Personal data protection: are they personal data? If so, have you gained (written) consent from data subjects to collect this information?	No personal data.
Access Procedures	None within the project consortium.
Embargo periods (if any)	None
Archiving and preservation (including storage and backup)	
Data storage (including backup): where? For how long?	On project private file repository. Shall be maintained and backed up for a period of 3years following the end of the project.
Indicative associated costs for data archiving and preservation	N/A
Indicative plan for covering the above costs	N/A
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)?	N/A

Will the data be securely stored in trusted repositories for long-term preservation and curation?	N/A
Ethics	
Describe any potential ethical issues during data collection, storage, processing and archiving, together with the ethical approval procedures related to the project.	N/A
If the research activities involve children, patients, members of vulnerable populations, use of embryonic stem cells, privacy and data protection issues or research on animals and primates, the ethical principles and relevant national, EU and international legislation must be complied with.	N/A

*Table 5: DS1.*5 & DS1.6 datasets *description.*

Data identification: *DS1.5 _ Ethical approvals* *DS1.6_ GDPR documents*	
Generic description:	
The datasets DS1.5 and DS1.6, are related to the legal and ethical issues related to theFAITH project (ethical approvals, informed consents, GDPR documents etc.).	
Origin of data:	
The datasets will be updated during the lifecycle of the project.	
Nature and scale of data:	
The nature of these dataset can be Word and or pdf documents.	
To whom the dataset could be useful:	
All partners related to the data collection and data processing.	
Related scientific publication(s)	
N/A	
Indicative existing similar data sets (including possibilities for integration and reuse):	
N/A	
Partners activities and responsibilities	
Partner owner of the data	All partners
Partner in charge of the data analysis	Legal and Ethical Committee
Partner in charge of the data storage	All partners
Related WP(s) and task(s)	All WPs
Standards and metadata	
Standards, format, estimated volume of data	Word format files. PDF format files

Data exploitation and sharing	
Data access policy/ Dissemination level: confidential (only for members of the Consortium and the Commission Services) or Public	Confidential (only for members of theConsortium and the Commission Services).
Data sharing, re-use, distribution, publication (How?)	Shall be limited only to be carried outbetween the Project Consortium members and the European Commission's Services.
Personal data protection: are they personal data? If so, have you gained (written) consent from data subjects to collect this information?	Personal data are involved. Informed consent will be collected before the proof-of-concept study starting date.
Access Procedures	None within the project consortium.
Embargo periods (if any)	None
Archiving and preservation (including storage and backup)	
Data storage (including backup): where? For how long?	The informed consent document will be prepared during work with the clinical project protocol. The consent form will be signed and kept either in a physical format and/or in an electronic format using safe storage platforms on the premises of Clinical centers. Other documents will be stored in the private file repository of the project.
Indicative associated costs for data archiving and preservation	N/A
Indicative plan for covering the above costs	N/A
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)?	N/A
Will the data be securely stored in trusted repositories for long-term preservation and curation?	N/A
Ethics	
Describe any potential ethical issues during data collection, storage, processing and archiving, together with the ethical approval procedures related to the project.	N/A
If the research activities involve children, patients, members of vulnerable populations, use of embryonic stem cells, privacy and data protection issues or research on animals and primates, the	N/A

ethical principles and relevant national, EU and international legislation must be complied with.	

*Table 6: DS*3.1, DS3.2 & DS3.3 datasets *description.*

Data identification: DS3.*1_Communication KPIs* *DS3.2 _ Dissemination materials* *DS3.3_ Exploitation plan*	
Generic description:	
The datasets DS3.1 and DS3.2 include information regarding the dissemination, exploitation and Innovation management activities of the FAITH consortium.	
Origin of data:	
The datasets will be updated during the lifecycle of the project, while a final consolidate version of them will be included in the deliverables of WP3.	
Nature and scale of data:	
The nature of these dataset can be Excel, Word, pdf documents, while the content of the dissemination materials can be web pages, brochures, flyers, PowerPoint presentations, papers in journal and conferences, videos, images etc.	
To whom the dataset could be useful:	
All partners	
Related scientific publication(s)	
Journal and conferences publications that are produced during the lifecycle of the project.	
Indicative existing similar data sets (including possibilities for integration and reuse):	
N/A	
Partners activities and responsibilities	
Partner owner of the data	FAITH consortium
Partner in charge of the data analysis	Dissemination and Exploitation Manager
Partner in charge of the data storage	FAITH consortium
Related WP(s) and task(s)	All Tasks
Standards and metadata	
Standards, format, estimated volume of data	Excel, Word, PowerPoint, PDF Image formats (*.tiff, *.png, *.jpeg etc.)Video formats
Data exploitation and sharing	
Data access policy/ Dissemination level: confidential (only for members of the Consortium and the Commission Services) or Public	Public for the dissemination materials. Exploitation plan and IPR that will be confidential to the consortium partners and the Commission's Services
Data sharing, re-use, distribution, publication (How?)	The dissemination materials can be shared, re-used and distributed following copyright agreements. Exploitation plan and IPR that shall be limited only to

	be carried out between the Project Consortium members and the Commission's Services.
Personal data protection: are they personal data? If so, have you gained (written) consent from data subjects to collect this information?	No Personal data are involved.
Access Procedures	For all the public datasets non access procedures are applied
Embargo periods (if any)	None
Archiving and preservation (including storage and backup)	
Data storage (including backup): where? For how long?	Shall be maintained and backed up for a period of 3 years following the end of theproject.
Indicative associated costs for data archiving and preservation	N/A
Indicative plan for covering the above costs	N/A
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)?	N/A
Will the data be securely stored in trusted repositories for long-term preservation and curation?	N/A
Ethics	
Describe any potential ethical issues during data collection, storage, processing and archiving, together with the ethical approval procedures related to the project.	N/A
If the research activities involve children, patients, members of vulnerable populations, use of embryonic stem cells, privacy and data protection issues or research on animals and primates, the ethical principles and relevant national, EU and international legislation must be complied with.	N/A

*Table 7: DS*4.1 LSP1 *description.*

Data identification: ***DS4.1_LSP1 description***
Generic description:
The data that will be used for the media LSP include learning materials, questionnaires, surveys and mockups created by ATC in collaboration to FH. Additionally, the AI-tool will store interaction data of LSP participants (media experts), such as log data, resources accessed and feedback on the functionality of the AI-tool. Based

on these data, the AI-tool will be optimized and its trustworthiness enhancement will be monitored throughout the project.	
Origin of data:	
Learning materials, questionnaires, surveys and mock-ups are provided by ATC in collaboration with partner FH. Interaction data generated from the media experts' use of the AI-tool are collected automatically while media experts interact with the platform.	
Nature and scale of data:	
Data will be stored in a database (e.g., mongoDB, PostgresSQL)	
To whom the dataset could be useful:	
The data will be useful to FAITH to monitor the trustworthiness enhancement of the AI platform and evaluate the FAITH trustworthiness optimization framework. The data will also be useful to ATC to optimize the AI-tool piloted in FAITH project.	
Related scientific publication(s)	
N/A	
Indicative existing similar data sets (including possibilities for integration and reuse):	
No this AI-tool has not been tested before	
Partners activities and responsibilities	
Partner owner of the data	FH
Partner in charge of the data analysis	ATC
Partner in charge of the data storage	ATC
Related WP(s) and task(s)	WP4, T4.1, T4.2, T4.3
Standards and metadata	
Standards, format, estimated volume of data	The dataset complies with GDPR rules for the protection of personal data. All the data will be stored in PostgreSQL. The dataset format will be relational with capability to export in other formats, such as csv or json. The estimated size of dataset is few KB.
Data exploitation and sharing	
Data access policy/ Dissemination level: confidential (only for members of the Consortium and the Commission Services) or Public	Confidential
Data sharing, re-use, distribution, publication (How?)	Summaries/statistics of the results from the pilots may be used for research publication.
Personal data protection: are they personal data? If so, have you gained (written) consent from data subjects to collect this information?	The dataset does not include any sensitive personal data. The users will log in anonymously to the platform hence it will no be possible to be traced.
Access Procedures	N/A
Embargo periods (if any)	N/A
Archiving and preservation (including storage and backup)	

Data storage (including backup): where? For how long?	The data will be stored in ATC services until the end of the project.
Indicative associated costs for data archiving and preservation	The indicative associated cost is 300€/annum.
Indicative plan for covering the above costs	These costs have been provisioned in the GA and will be covered from C.3 (Other goods, works and services).
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)?	We will investigate whether we will use the ATC infrastructure or Azure cloud services. If ATC infrastructure is going to be used, then the data will be stored in secure, access-controlled environments provided by ATC services, compliant with all relevant standards. Similarly in the case where Azure cloud services are employed.
Will the data be securely stored in trusted repositories for long-term preservation and curation?	The data will be securely stored in trusted repositories for long-term preservation and curation provided by ATC/Azure.
Ethics	
Describe any potential ethical issues during data collection, storage, processing and archiving, together with the ethical approval procedures related to the project.	No ethical issues since the users log in anonymously and cannot be traced.
If the research activities involve children, patients, members of vulnerable populations, use of embryonic stem cells, privacy and data protection issues or research on animals and primates, the ethical principles and relevant national, EU and international legislation must be complied with.	

*Table 8: DS*5.1 LSP2 *description.*

Data identification: ***DS5.1_LSP2 description***
Generic description:
The data that will be used for the transportation LSP includes images captured by surveillance cameras inside the trains. The information collected includes: - Passengers getting on and off the train. To evaluate the turnout in the different time slots and understand which stops the greatest number of passengers pass through.

- Passengers moving along the train, using images of individual people. - Passenger counting. Collecting the number of passengers in each carriage, counting how many people are sitting, how many are standing, and monitoring where people sit most frequently. - Ticket validation processes. - Baggage handling. - Security checks. - Interactions between passengers and staff. - Accidents or emergencies that occur on the train. - Train maintenance and cleaning activities. - Safety measures implemented on trains.	
Origin of data:	
The data is collected from CCTV cameras installed on board the trains. The images are annotated by company personnel.	
Nature and scale of data:	
All data is stored in local database protected by access limitation (SW access limitation and physical access limitation)	
To whom the dataset could be useful:	
The data is useful for train management: for analyzing passenger flow, security monitoring, safety assessments, and improving operational efficiency within the train transportation system.	
Related scientific publication(s)	
Not yet, however, the publication of datasets and associated scientific papers will be considered. Open access, either gold or green, will also be taken into account.	
Indicative existing similar data sets (including possibilities for integration and reuse):	
There exist general image and video datasets that might be used to pre-train some models before performing fine-tuning, testing, and validation on the datasets that will be acquired in the project. However, to the best of our knowledge, no datasets on the specific theme that could be reused or integrated currently exist.	
Partners activities and responsibilities	
Partner owner of the data	MERMEC
Partner in charge of the data analysis	CNR, MERMEC
Partner in charge of the data storage	CNR, MERMEC
Related WP(s) and task(s)	WP5, T5.1, T5.2, T5.3
Standards and metadata	
Standards, format, estimated volume of data	Encoded video files (eg H264/H265) with accompanying metadata and annotation files in JSON format. The estimated size if 1MB/s/camera. One our session for a multicamera system with 8 cameras will have size approx 30GB size.

	The expected total size will be about 10 TB. The size of the metadata and annotation is negligible with respect to the primary data.
Data exploitation and sharing	
Data access policy/ Dissemination level: confidential (only for members of the Consortium and the Commission Services) or Public	Raw video files will be kept confidential and disregarded. Curated datasets will be made available according to the open data principles; offuscated data will be provided in some circumstances.
Data sharing, re-use, distribution, publication (How?)	To be agreed - CC BY-NC-ND 4.0 might be considered.
Personal data protection: are they personal data? If so, have you gained (written) consent from data subjects to collect this information?	In case of the presence of persons in videos, prior explicit written consent will be acquired and in any case signage of video recording and advertisement will be ensured (eg in the case of the use on carriages in public service, adequate advertising of experimentation will be issued according to the recommendations by the ethical committee that will be set up for this specific LSP).
Access Procedures	Public Dataset will be made available; other enquiries will have to be directed to the data controller (according to GDPR).
Embargo periods (if any)	N/A
Archiving and preservation (including storage and backup)	
Data storage (including backup): where? For how long?	Raw videos will be acquired on portable hard disk in two copies and will be password encrypted. Then data transfer on a central storage unit using an hardware RAID controlled with redundancy will be used. Procedure about security of this unit will be detailed in the (planned) DPIA (Data protection Impact Assessment)
Indicative associated costs for data archiving and preservation	1k€ per year
Indicative plan for covering the above costs	Indirect costs of the project (overheads)
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)?	Linux storage server administered and regularly updated to address critical issues. It features hardware redundancy and RAID for enhanced

	reliability and data protection. Details will be provided in the DPIA. Suitability will be assessed by the Ethical Committee to which we will submit the plan of the study involved in the pilot.
Will the data be securely stored in trusted repositories for long-term preservation and curation?	YES
Ethics	
Describe any potential ethical issues during data collection, storage, processing and archiving, together with the ethical approval procedures related to the project.	The LSP uses cameras to capture video streams that include human figures on board public transit and possibly in related premises (stations, platforms, shelters). These streams are processed to gather information related to the safety and security of passengers and to understand mobility patterns. Extracted data from these human figures may include anthropometric (but not truly biometric) characteristics. For the sole purpose of detecting events of interest, the original video streams do not need to be transferred from the local processing unit and can be deleted after processing. However, relevant aggregated information from the observed events can be transmitted from the local processing unit to a remote location for storage and further processing and analysis. In some cases, it might be necessary to keep "video summaries" of observed events or other evidence for accountability or explanation purposes of the AI models employees, also on the basis of the results of the currently ongoing activities in WP2. It is possible to consider blurring or other methods to obfuscate such videos or, even, the actual real figures might be replaced by avatars or skeleton models.
If the research activities involve children, patients, members of vulnerable populations, use of embryonic stem cells, privacy and data protection issues or research on animals and primates, the	*In the case of data acquisition on carriages in normal services, datasets will not be released in their original form but suitable obfuscation will be applied first. Areas of recording will be adequately advertised. However, it is possible that individuals unable to give*

ethical principles and relevant national, EU and international legislation must be complied with.	*consent (such as children or other vulnerable categories) and traveling alone may enter those areas.*

*Table 9: DS*6.1 LSP3 *description.*

Data identification: ***DS6.1_LSP3 description***	
Generic description:	
The data that will be used for the education LSP include learning materials, questionnaires, and inquiry-based scenarios tailored to specific STEM concepts, which will be created by the teachers and imported through the authoring tool. Additionally, the learning environment will store interaction data of students, such as experimental data and student responses to questions. Interaction data includes logs of student activities, such as time spent on different tasks, resources accessed, and learning paths of students. Based on the interaction data of students, the platform provides analytics of student performance, identifies student learning patterns, and predicts student performance.	
Origin of data:	
Learning materials, questionnaires, and scenarios are provided by teachers and imported into the system using an authoring tool. Interaction data generated from the students' use of the learning environment are collected automatically as students interact with the platform.	
Nature and scale of data:	
All the data are stored in PostgreSQL.	
To whom the dataset could be useful:	
The data could be useful to educators for assessing the performance of their class, identifying areas of improvement, and tailoring instructional approaches to meet individual learning needs.	
Related scientific publication(s)	
N/A	
Indicative existing similar data sets (including possibilities for integration and reuse):	
N/A	
Partners activities and responsibilities	
Partner owner of the data	EA
Partner in charge of the data analysis	ICCS
Partner in charge of the data storage	ICCS
Related WP(s) and task(s)	WP6, T6.1, T6.2, T6.3
Standards and metadata	
Standards, format, estimated volume of data	The dataset complies with GDPR rules for the protection of personal data. All the data will be stored in PostgreSQL. The dataset format will be relational with capability to export in other formats, such as csv or json. The estimated size of dataset is 400GB.

Data exploitation and sharing	
Data access policy/ Dissemination level: confidential (only for members of the Consortium and the Commission Services) or Public	Confidential
Data sharing, re-use, distribution, publication (How?)	The results from the pilots will be used for research publication.
Personal data protection: are they personal data? If so, have you gained (written) consent from data subjects to collect this information?	The dataset does not include any sensitive personal data.
Access Procedures	Authorized employees or researchers
Embargo periods (if any)	-
Archiving and preservation (including storage and backup)	
Data storage (including backup): where? For how long?	The data will be stored on AWS Amazon services for the duration of the project.
Indicative associated costs for data archiving and preservation	Indicative associated cost is 500€/annum.
Indicative plan for covering the above costs	Budget Sponsored by cloud provider for academic use.
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)?	The data will be stored in secure, access-controlled environments provided by AWS services.
Will the data be securely stored in trusted repositories for long-term preservation and curation?	The data will be securely stored in trusted repositories for long-term preservation and curation provided by AWS. The platform has different user roles. Access to data related to student interactions is granted only to the students' teachers and the system administrators.
Ethics	
Describe any potential ethical issues during data collection, storage, processing and archiving, together with the ethical approval procedures related to the project.	Potential ethical issues during data collection, storage, processing, and archiving include privacy breaches, unauthorized access to sensitive data, and misuse of data. The platform does not store any sensitive personal data. All the interaction data of students are anonymized.
If the research activities involve children, patients, members of vulnerable populations, use of embryonic stem cells, privacy and data protection	Our research activities involve children and are compliant with EU policies.

issues or research on animals and primates, the ethical principles and relevant national, EU and international legislation must be complied with.	

Table 10: DS7.1 LSP4 *description.*

Data identification: *DS7.1_LSP4 description*	
Generic description:	
The data set that will be used will be collected on-site by UNIFI's robots. It will consist of: - Acoustic data obtained from the Forward Looking Sonar (FLS) - Optical images from cameras - Bathymetric data from Multibeam Echosounder (MBES) or similar 3D Sonar - Navigation data from the AUV localization sensors and algorithm	
Origin of data:	
From the port visits during testing prior to the LSP executions and, data from the actual day of the LSP.	
Nature and scale of data:	
TBC	
To whom the dataset could be useful:	
UNIFI as providers of the AI models.	
Related scientific publication(s)	
N/A	
Indicative existing similar data sets (including possibilities for integration and reuse):	
Collections of public datasets to perform object detection underwater (either with optical or acoustic imagery) are available at the following repositories: https://github.com/mousecpn/Collection-of-Underwater-Object-Detection-Dataset https://github.com/xahidbuffon/Awesome_Underwater_Datasets?tab=readme-ov-file https://github.com/xinzhichao/Underwater_Datasets However, due to the high variability in optical frames (depending on the location of acquisition) and the specificity of acoustic images (related to the particular sensor used), the utility of utilizing such public datasets remains uncertain.	
Partners activities and responsibilities	
Partner owner of the data	- UNIFI & APRA (for the data related to the port of Ravenna) - UNIFI & VPF for the data related to the port of Valencia.
Partner in charge of the data analysis	- UNIFI

Partner in charge of the data storage	- UNIFI, APRA, VPF
Related WP(s) and task(s)	- WP7
Standards and metadata	
Standards, format, estimated volume of data	- Acoustic data format: either raw or compressed image -estimated size proportional to order of magnitude of 1 GB - Optical images format: either raw or compressed image - estimated size proportional to order of magnitude of 1 GB - Bathymetric data format: sensor proprietary format (depending on the specific sensor to be acquired), georeferenced 3D/2D data, point cloud format - estimated size proportional to order of magnitude of 1 or 10 GB - Navigation data format: tabular data - estimated size proportional to order of magnitude of 0.1 GB
Data exploitation and sharing	
Data access policy/ Dissemination level: confidential (only for members of the Consortium and the Commission Services) or Public	Confidential
Data sharing, re-use, distribution, publication (How?)	-within WP7 partners
Personal data protection: are they personal data? If so, have you gained (written) consent from data subjects to collect this information?	N/A
Access Procedures	-Formal request
Embargo periods (if any)	-None
Archiving and preservation (including storage and backup)	
Data storage (including backup): where? For how long?	The data will be stored in UNIFI’s services until the end of the project.
Indicative associated costs for data archiving and preservation	-None
Indicative plan for covering the above costs	-Not necessary
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)?	-Strict access controls: only authorized personnel will have access to data. -Redundant storage systems and data replication (in repositories and physical storage disks owned by the

	partners) to ensure that data is not lost if one storage system fails.
Will the data be securely stored in trusted repositories for long-term preservation and curation?	- The data will be securely stored in trusted repositories with strict access control. The data will be available to authorized individuals only.
Ethics	
Describe any potential ethical issues during data collection, storage, processing and archiving, together with the ethical approval procedures related to the project.	N/A
If the research activities involve children, patients, members of vulnerable populations, use of embryonic stem cells, privacy and data protection issues or research on animals and primates, the ethical principles and relevant national, EU and international legislation must be complied with.	N/A

*Table 11: DS*7.2 LSP4 bathymetry *description*.

Data identification: *DS7.1_LSP4 bathymetry description*	
Generic description:	
This type of data is the Bathymetry data of the port of Valencia, which is the sea floor topography. It provides accurate readings of water depths at various locations within the port. This data will be used to feed the AI models prepared by UNIFI. It will also be complemented by the data gathered on site from the actual robots.	
Origin of data:	
Obtained from the last bathymetry campaign of the port authority of Valencia. It is retrieved by the port's topography technicians with their specific equipment.	
Nature and scale of data:	
.xyz file with the bathymetry levels	
To whom the dataset could be useful:	
UNIFI, for training the AI model to be used in LSP4 – Robotics/Drones (WP7)	
Related scientific publication(s)	
N/A	
Indicative existing similar data sets (including possibilities for integration and reuse):	
There are no public datasets available of bathymetries inside ports.	
Partners activities and responsibilities	
Partner owner of the data	FVP
Partner in charge of the data analysis	UNIFI

Partner in charge of the data storage	FVP
Related WP(s) and task(s)	WP7, all tasks
Standards and metadata	
Standards, format, estimated volume of data	ETRS89 Huso30; The bathymetry levels are positive, measured every 2 meters.
Data exploitation and sharing	
Data access policy/ Dissemination level: confidential (only for members of the Consortium and the Commission Services) or Public	Confidential
Data sharing, re-use, distribution, publication (How?)	Data will be shared and distributed to the technical partners that need it for the project
Personal data protection: are they personal data? If so, have you gained (written) consent from data subjects to collect this information?	There is no personal data.
Access Procedures	Formal request to the Valenciaport Foundation
Embargo periods (if any)	None
Archiving and preservation (including storage and backup)	
Data storage (including backup): where? For how long?	On Sharepoint and locally at VPF
Indicative associated costs for data archiving and preservation	Negligible (3MB)
Indicative plan for covering the above costs	None
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)?	Cloud storage plus local copies of the dataset
Will the data be securely stored in trusted repositories for long-term preservation and curation?	We use Sharepoint, managed by Microsoft, to store our files in the cloud
Ethics	
Describe any potential ethical issues during data collection, storage, processing and archiving, together with the ethical approval procedures related to the project.	N/A
If the research activities involve children, patients, members of vulnerable populations, use of embryonic stem cells, privacy and data protection issues or research on animals and primates, the ethical principles and relevant national, EU and international legislation must be complied with.	N/A

*Table 12: DS*8.1 LSP5 *description.*

Data identification: ***DS8.1_LSP5 description***	
Generic description:	
The dataset for LSP5 encompasses data from the wastewater treatment process at project partner Veas, who has Norway's largest processing plant for wastewater cleaning.	
Origin of data:	
The data originates from the fully automated process plant, which includes 15K objects and 60K signals. Data from the process are logged and stored, with a potential history of several years. Since 2021, the data has been made accessible through the currently used environment, ABB Edge Insight, into one cloud platform.	
Nature and scale of data:	
The nature and scale of data used for the research tasks in FAITH, depends on the specific tasks. Potentially, data from up to 10 years history can be used. Data from 2021 will be immediately available through the current environment.	
To whom the dataset could be useful:	
The dataset is useful to Veas and to researchers working on the AI-models to be applied by Veas.	
Related scientific publication(s)	
The work in LSP5 will result in scientific publications.	
Indicative existing similar data sets (including possibilities for integration and reuse):	
We are not aware of public datasets similar to the dataset in LSP5.	
Partners activities and responsibilities	
Partner owner of the data	Veas
Partner in charge of the data analysis	SINTEF
Partner in charge of the data storage	Veas (SINTEF stores and processes data used for specific analyses)
Related WP(s) and task(s)	WP8
Standards and metadata	
Standards, format, estimated volume of data	For extracted datasets for specific analyses: Format: .csv files for measurements, .xlsx for metadata Estimated size: Less than 1GB
Data exploitation and sharing	
Data access policy/ Dissemination level: confidential (only for members of the Consortium and the Commission Services) or Public	Due to security aspects of Veas data, these will only be accessible to researchers with confidentiality agreements.

Data sharing, re-use, distribution, publication (How?)	Limited extracted datasets can be made available to others, e.g. related to academic publications, on a case to case basis.
Personal data protection: are they personal data? If so, have you gained (written) consent from data subjects to collect this information?	The dataset does not contain personal data
Access Procedures	N/A
Embargo periods (if any)	N/A
Archiving and preservation (including storage and backup)	
Data storage (including backup): where? For how long?	The data is stored through the Veas installation of ABB Edge Insight
Indicative associated costs for data archiving and preservation	N/A
Indicative plan for covering the above costs	N/A
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)?	Data security in line with Veas processes for security and quality control
Will the data be securely stored in trusted repositories for long-term preservation and curation?	Yes, stored in the Veas installation of ABB Edge insight
Ethics	
Describe any potential ethical issues during data collection, storage, processing and archiving, together with the ethical approval procedures related to the project.	N/A
If the research activities involve children, patients, members of vulnerable populations, use of embryonic stem cells, privacy and data protection issues or research on animals and primates, the ethical principles and relevant national, EU and international legislation must be complied with.	N/A

*Table 13: DS*9.1 LSP6 *description.*

Data identification: ***DS9.1_LSP6 description***
Generic description:
The collected data includes prostate MR images from various vendors and will be used in the healthcare LSP of FAITH project. The data will be exploited to assess the trustworthiness of the AI-based system which provides automatic segmentations of Prostate and its anatomical regions (Peripheral zone, Transition zone).

In this perspective, the data will also be used to enable the fine tuning of the already developed deep learning model. The vision: such a pipeline is the AI based system, designed and developed to replace the second reader (radiologist) in order to minimize the burden in the healthcare sector. Further, model vulnerabilities will be identified and assessed as described in the FAITH AI_TAF.	
Origin of data:	
Prostate T2-weighted patient scans are provided by the clinical experts for evaluation and model refinement. • PAGNI Hospital Heraklion (1st Pilot) • UNIPI (replication Pilot)	
Nature and scale of data:	
The data are not transferred outside the premises of clinical experts but stored anonymized on local dicom server (e.g. ORTHANC). The data are MRI T2-weighted, DWI/ADC and when available DCE sequences in DICOM format.	
To whom the dataset could be useful:	
Clinical Experts may evaluate the performance of the AI segmentation tool and refine the system output (segmentation masks) when they deem those are inappropriate. AI developers will refine the model based on the corrections/suggestions from the clinical experts. The vision is to build a trustworthy AI-system for all involved stakeholders.	
Related scientific publication(s)	
<Is the dataset related to a scientific publication? Is the latter Gold or Green Open Access?> No	
Indicative existing similar data sets (including possibilities for integration and reuse):	
- PICAI Dataset - Prostate-158 - Prostate X2 - Prostate 3T	
Partners activities and responsibilities	
Partner owner of the data	Hospital of PAGNI, UNIPI
Partner in charge of the data analysis	FORTH
Partner in charge of the data storage	Hospital of PAGNI, UNIPI (premises)
Related WP(s) and task(s)	WP9, T9.1, T9.2, T9.3
Standards and metadata	
Standards, format, estimated volume of data	Format: DICOM Estimated Volume: 2GB The dataset complies with GDPR rules for the protection of personal data.
Data exploitation and sharing	

Data access policy/ Dissemination level: confidential (only for members of the Consortium and the Commission Services) or Public	Confidential
Data sharing, re-use, distribution, publication (How?)	The results from the pilots will be used for research publication and to validate the FAITH AI_TAF
Personal data protection: are they personal data? If so, have you gained (written) consent from data subjects to collect this information?	The dataset will be strictly anonymized since any health-related dataset include sensitive data. Ethical approvals will be provided and explicit inform consent will be in place.
Access Procedures	Authorized employees or researchers
Embargo periods (if any)	N/A
Archiving and preservation (including storage and backup)	
Data storage (including backup): where? For how long?	On the premises of the clinical experts involved in the LSP6 for the duration of the project.
Indicative associated costs for data archiving and preservation	
Indicative plan for covering the above costs	
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)?	The data will be stored on the premises of the clinical partners, Data in Rest will be encrypted exploiting DiskEncryption of the filesystem. DICOM index will be stored in a database index of an encrypted PostgreSQL
Will the data be securely stored in trusted repositories for long-term preservation and curation?	The data will be securely stored the premises of clinical partners during FAITH lifetime.
Ethics	
Describe any potential ethical issues during data collection, storage, processing and archiving, together with the ethical approval procedures related to the project.	Potential ethical issues during data collection, storage, processing, and archiving include privacy breaches, unauthorized access to sensitive data, and misuse of data. Since it is not envisaged to have data transfer any risk is minimized. Only data transfer will be on metadata/responses from the experts to the performance and other trustworthiness characteristics of the proposed AI-based system.
If the research activities involve children, patients, members of vulnerable populations, use of embryonic stem cells, privacy and data protection issues or research on animals and primates, the	Our research activities involve patient and are compliant with EU policies.

ethical principles and relevant national, EU and international legislation must be complied with.	

*Table 14: DS*10.1 LSP7 *description.*

Data identification: *DS10.1_LSP7 description*
Generic description:
The data that will be used for the active ageing LSP includes the data obtained from the ACTIVAGE application for the real time monitoring of the users participating in the study. The data is obtained from sensors installed at the houses of the participants. The information collected includes: - Presence detection in the different home rooms - Open and close of main entrance door. - Time user is away from home - Time spent in every room - Number of visits to the bathroom - Time spent in the bathroom - Daily active time - Daily rest time - Alerts triggered
Origin of data:
The dataset that is being used to develop the AI models comes from the results of the ACTIVAGE project. It consists of the data collected during the pilot execution in Valencia where presence sensors and door sensor where deployed in 500 homes during one year. The dataset to train and improve the AI models according to FAITH framework will be collected from the deployment of ACTIVAGE application in the homes of the FAITH LSP 7 participants.
Nature and scale of data:
The initial dataset is in JSON format and it is stored in a local database protected by access limitation. The FAITH dataset will be stored in a Mongo DB in JSON format in a private cloud instance.
To whom the dataset could be useful:
The dataset could be useful for providers of social and health services that want to improve the efficiency of their services with a more proactive and personalized approach
Related scientific publication(s)
Not yet; however, the publication of datasets and associated scientific papers will be considered. Open access, either gold or green, will also be taken into account.
Indicative existing similar data sets (including possibilities for integration and reuse):
To the best of our knowledge, no datasets on the specific theme that could be reused or integrated currently exist.
Partners activities and responsibilities

Partner owner of the data	ACTIVAGE
Partner in charge of the data analysis	BRDG
Partner in charge of the data storage	ACTIVAGE
Related WP(s) and task(s)	WP10
Standards and metadata	
Standards, format, estimated volume of data	The initial dataset is in JSON format and its size is 300 Gb The dataset that will be generated during the pilot in FAITH project will be also in JSON format and the expected size is similar to the previous one.
Data exploitation and sharing	
Data access policy/ Dissemination level: confidential (only for members of the Consortium and the Commission Services) or Public	Raw sensor data will be kept confidential. Curated datasets will be made available following the principles of open data and FAITH strategy of dissemination
Data sharing, re-use, distribution, publication (How?)	The data sharing policy will be based on the specific license established in FAITH. It will be based on a permissive base licence that allows the negotiation of sub-licences in the style of MIT-type licences, allowing the exploitation and level of access and specific use for each request for access to the data. In addition, public access will be given to the metadata describing the dataset to encourage its search, interoperability, preservation and dissemination, as well as the assignment of a DOI.
Personal data protection: are they personal data? If so, have you gained (written) consent from data subjects to collect this information?	The dataset from ACTIVAGE project is anonymized and doesn't contain sensitive personal data. The dataset to be collected during FAITH pilot will follow the same approach and it will not contain personal information or any identification information
Access Procedures	Public Dataset will be made available in the project designated repository, e.j. Zenodo; other enquiries will have to be directed to the data controller (according to GDPR).-
Embargo periods (if any)	NA
Archiving and preservation (including storage and backup)	
Data storage (including backup): where? For how long?	The dataset from ACTIVAGE project is stored locally in an access restricted server. The dataset created from FAITH pilot will be also stored locally in the same server until the moment to open it

	in a public repository. The server has implemented a policy of daily incremental backups and monthly complete backup.
Indicative associated costs for data archiving and preservation	1k€ per year
Indicative plan for covering the above costs	They are included in the indirect costs of the project
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)?	Storage server is managed and regularly updated to address critical issues. It has hardware redundancy and RAID for enhanced reliability and data protection. Security measures are implemented for access control to the data and fully details will be provided in the DPIA.
Will the data be securely stored in trusted repositories for long-term preservation and curation?	Yes
Ethics	
Describe any potential ethical issues during data collection, storage, processing and archiving, together with the ethical approval procedures related to the project.	Data collected in the datasets can't be considered per se sensitive data, however, when processed by AI technologies it could produce artifacts that could infer user behaviour, thus potentially raising ethics concerns. Users participating in the pilot will be presented with an informed consent where data collection, data protection and data processing procedures will be described and consent will be asked from them.
If the research activities involve children, patients, members of vulnerable populations, use of embryonic stem cells, privacy and data protection issues or research on animals and primates, the ethical principles and relevant national, EU and international legislation must be complied with.	EU ethical principles and EU regulations will be followed and complied with in regards of the collection, processing and storage of data.

2.4 Data security

Security is of high importance for the FAITH consortium because it is a multifaceted quality attribute that affects the functionality, user experience, availability, and data protection. Thus, FAITH is committed to protecting information by mitigating information risks and implementing IT security measures. The primary focus of information security is the balanced protection of the confidentiality, integrity and availability of data while maintaining an efficient execution of the FAITH goals, all without hampering the productivity within the project. Accordingly, FAITH will implement measures towards achieving the three tenets of information security (also known as the CIA-triad model), including the following:

- **Confidentiality:**
  - Mechanisms for protecting information from unwanted exposure, tampering, or destruction (e.g., access controls such that only authorised users and processes will be able to access or modify data processed within FAITH);
- **Integrity:**
  - Measures to ensure that information and software remain intact and correct (e.g., back-ups of data as well as system audits to check whether data related to FAITH is maintained in a correct state and no improper modifications (either accidental or malicious) are possible);
- **Availability:**
  - Mechanisms ensuring that only authorised users can access data on need to know basis whenever necessary for FAITH (e.g., through procedures for handling IT system users, i.e., identification, authentication and authorisation).

Thus, the CIA-triad will guide the FAITH consortium's efforts and policies aimed at keeping all the data secure. While not explicitly foreseen, the CIA-triad can largely be achieved by implementing the IT controls and security measures mandated by the applicable data protection legislation. Accordingly, all partners of the FAITH consortium will adopt good practice data security procedures in the project, in accordance with the relevant data protection rules (including the GDPR). In this manner, FAITH will be able to avoid unforeseen usage or disclosure of data, including the mosaic effect, i.e., obtaining identification by merging multiple sources). Relevant measures include restrictive access controls via secure log-ins, installation of up-to-date security software on devices, regular data backups, etc.

Other IT security processes to be considered and adopted during FAITH include the following:

- Addressing common security problems that affect IT systems by implementing standard best practices (for instance, the use of software tools, such as virus checkers, antimalware protection, firewalls and updating software and patches);
- Protecting the hardware and software used for project activities from a wide variety of IT security threats (such as environmental factors, e.g., vandalism, sabotage, and theft);
- Appointing an internal team of IT security experts at each partner organisation, which will be tasked to manage security issues to ensure effective incident responses;
- Having security alerts in place such that computer or system users are able to understand their role in IT security and take appropriate actions in response to potential security threats (e.g., the quickly reporting security incidents to response teams such that security threats can be effectively addressed).

For both all the project’s repositories the following aspects are addressed:

- **Authentication**: provide and validate information about the person or the system interacting with the platform

- **Authorisation**: restrict access control based on the users' identity and their access rights on the data and compute resources
- **Audit**: logging and monitor of users' and systems' behaviour throughout the platform
- **Confidentiality**: all interactions are encrypted and protected from unauthorized access and eavesdropping

Specifically, for the FAITH internal file repository, all the information in transit is performed over secure channels. The use of Transport Layer Security (TLS) with strong ciphers is the established best practice for securing network communication. Additionally, TLS provides integrity and authenticity of the interacting peers. In addition, the TLS-secured network communication channels and HTTPS is used both internally (among the platform's components) and externally (when the system is accessed by its users or other systems). The identification and authenticity of the interacting FAITH components is verified with digital certificates signed by either using well known and trusted Certificate Authorities (CA) or an internal CA of the platform in order to simplify deployment. The latter can facilitate the testing of the components and itis certainly easier, at the cost of supporting only the internal communication. Of use of strong private keys (2018-bit RSA or 256-bit ECDSA), recent versions of TLS (TLS 1.2 and 1.3), and a short list of strong ciphers that offer at least 128-bit encryption will be utilized.

Regarding the private repository based on the Nextcloud technology integrates logging and intrusion detection and works with existing authentication mechanisms like SAML, Kerberos and LDAP. Administrators can set permissions on sharing and access to files using groups. The CBMLBox employs standard TLS to encrypt data in transfer and offers Server Side Encryption on the local storage.

3 Alignment to the Findable, Accessible, Interoperable, Re-usable (FAIR) data principles

The FAIR data principles were first published in 2016 aiming to enhance the findability, accessibility, interoperability, and reusability of digital resources for both human and machines. These principles consider applications and computational agents as stakeholders with the capacity to find, access, interoperate, and reuse data with none or minimal human intervention. They deliberately do not specify technical requirements but deliver a set of guiding principles that provide for a continuum of increasing reusability, via many different implementations. They describe characteristics and aspirations for systems and services to support the creation of valuable research outputs that could then be rigorously evaluated and extensively reused, with appropriate credit, to the benefit of creators and users[3].

Since the initiation of the FAIR principles in 2014, FAIR metrics[4], FAIR infrastructure[5] and FAIR tools[6] have been developed to aid the process of making data FAIR ("FAIRification"). In addition, a domain-independent FAIRification workflow emerged[7], considering the evident need to efficiently analyze sparse, heterogeneous and privacy-sensitive data from multiple sources across institutes and countries. With the advent of FAIR, the workflow was adapted to also cover the other three facets of FAIR: findability, accessibility and reusability, which heavily relies on metadata (i.e., data descriptors). Adopting these developments, this deliverable presents the FAIRification workflow that will be used as a guide and template for the needs of the FAITH project.

FAITH aims to provide the global reference data resource, data specifications and format and application platform in support of the next generation AI trustworthiness research in critical domains. All consortium partners have been committed to ensure that the data produced, collected, and processed, align with the FAIR Principles definition[8], thus are **findable, accessible, interoperable and reusable**. Through the life cycle of the FAITH project, the FAIR principles are followed as far as possible, for both the data available in the infrastructure and the AI models, while ensuring compliance with national and European ethic-legal framework.

[3] Mons, B., Neylon, C., Velterop, J., Dumontier, M., da Silva Santos, L. O. B., & Wilkinson, M. D. (2017). Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. *Information services & use*, *37*(1), 49-56.

[4] Wilkinson, M. D., Dumontier, M., Sansone, S. A., Bonino da Silva Santos, L. O., Prieto, M., Batista, D., ... & Schultes, E. (2019). Evaluating FAIR maturity through a scalable, automated, community-governed framework. *Scientific data*, *6*(1), 174.

[5] Weigel, T., Schwardmann, U., Klump, J., Bendoukha, S., & Quick, R. (2020). Making data and workflows findable for machines. *Data Intelligence*, *2*(1-2), 40-46.

[6] Thompson, M., Burger, K., Kaliyaperumal, R., Roos, M., & da Silva Santos, L. O. B. (2020). Making FAIR easy with FAIR tools: From creolization to convergence. *Data Intelligence*, *2*(1-2), 87-95.

[7] https://zenodo.org/records/3207809

[8] Wilkinson M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3:160018 doi: 10.1038/sdata.2016.18 (2016).

This section is based on the guidelines for effective data management during a Horizon Europe project, provided by the European Commission[9]. The FAIR component of the current DMP version comprises points in need for further clarification, which will be addressed during the project. The proposed FAIRification workflow, tailored to the specific requirements and needs posed by FAITH Large Scale Pilots (LSPs), is shown in Figure 1 and it is based on the GO FAIR[10] initiative.

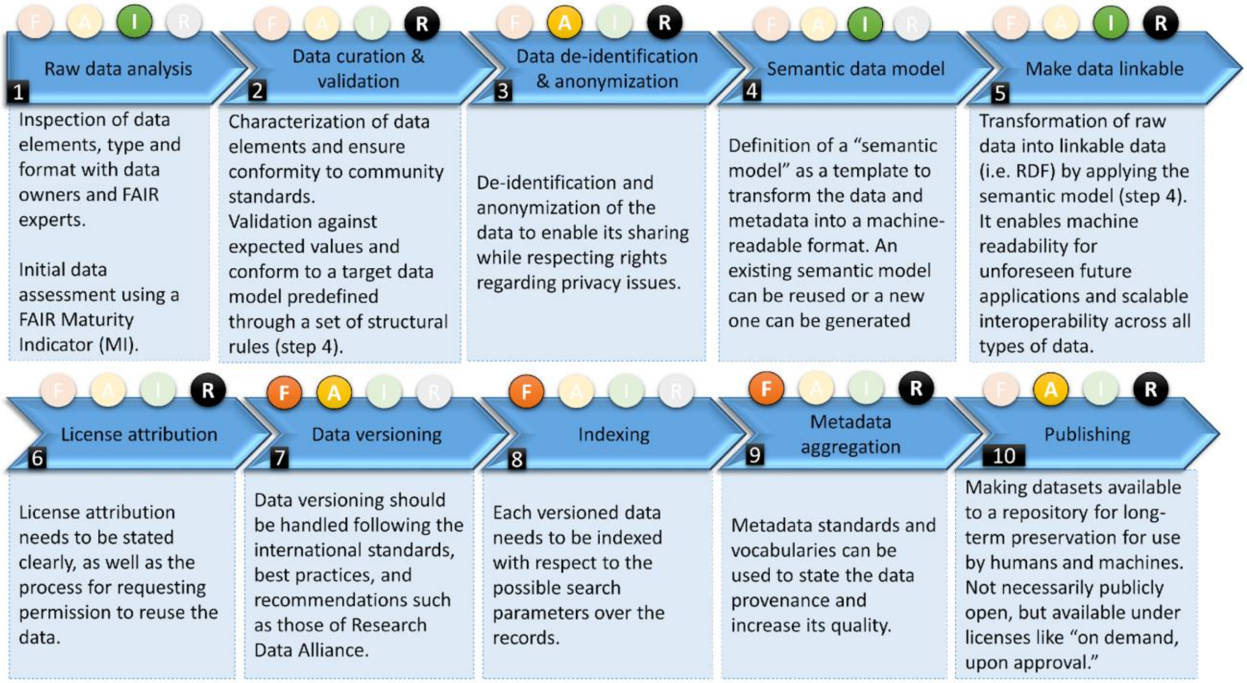
F A I R
1 Raw data analysis
Inspection of data elements, type and format with data owners and FAIR experts.
Initial data assessment using a FAIR Maturity Indicator (MI).
F A I R
2 Data curation & validation
Characterization of data elements and ensure conformity to community standards.
Validation against expected values and conform to a target data model predefined through a set of structural rules (step 4).
F A I R
3 Data de-identification & anonymization
De-identification and anonymization of the data to enable its sharing while respecting rights regarding privacy issues.
F A I R
4 Semantic data model
Definition of a "semantic model" as a template to transform the data and metadata into a machine-readable format. An existing semantic model can be reused or a new one can be generated
F A I R
5 Make data linkable
Transformation of raw data into linkable data (i.e. RDF) by applying the semantic model (step 4). It enables machine readability for unforeseen future applications and scalable interoperability across all types of data.
F A I R
6 License attribution
License attribution needs to be stated clearly, as well as the process for requesting permission to reuse the data.
F A I R
7 Data versioning
Data versioning should be handled following the international standards, best practices, and recommendations such as those of Research Data Alliance.
F A I R
8 Indexing
Each versioned data needs to be indexed with respect to the possible search parameters over the records.
F A I R
9 Metadata aggregation
Metadata standards and vocabularies can be used to state the data provenance and increase its quality.
F A I R
10 Publishing
Making datasets available to a repository for long-term preservation for use by humans and machines. Not necessarily publicly open, but available under licenses like "on demand, upon approval."

Figure 1: The FAIRification workflow for the FAITH data adjusted from Sinaci et al.[11]. Each step is characterized based on the FAIR requirements it is addressing ((F)indability, (A)ccessibility, (I)nteroperability, and (R)eusability)

As shown in the image above, the process is completed in ten (10) steps. A synopsis of each stage will follow and then how these actions are adopted in the FAITH project. In more detail, the stages of FAIRification process are as follows:

1. ***Raw Data Analysis:*** Raw data analysis inspects the content of the data and tries to discover the represented concepts, the structure within and between concepts and the format in which the data elements are stored and serialized. In addition to data modeling, health data utilizes several coding schemes and terminology systems which needs to be considered during the raw data analysis step.
2. ***Data Curation and Validation:*** This step aims to increase the quality of the data set for research purposes. During data curation, data fields, types and values (metadata) are characterized and extracted. Curated data should be validated against known quantitative relationships and expected values and should conform to the semantic model defined for the FAIRification workflow, i.e. the

[9] https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm
[10] https://www.go-fair.org/go-fairinitiative/
[11] Sinaci AA et al. From Raw Data to FAIR Data: The FAIRification Workflow for Health Research. Methods Inf Med. 2020 Jun;59(S 01):e21-e32. doi: 10.1055/s-0040-1713684

predefined target data model through a set of structural rules. In addition, the data itself should conform to the semantic rules exposed by the data element or attribute itself.

3. Data De-identification and Anonymization: Once the data set has been curated, validated and associated metadata collected, the next step is to de-identify and/or anonymize the data set in order to enable it to be shared without involving data subjects' rights regarding privacy issues (Cf. 3.4.2, GDPR). The decision to apply de-identification and/or anonymization to the dataset will depend on the purpose for which the dataset has been developed. In addition to identifying direct attributes such as identifiers and names, identifying other elements of the datasets that may act as quasi-identifiers such as dates (such as birth, death, admission, discharge, visit and sample collection), locations (such as zip codes and regions), race and ethnicity, and in some cases rare diagnoses, should also be addressed.

4. Semantic Data Modeling: This step involves defining a "semantic model" for the data set, which describes the meaning of the entities and relationships in the data set accurately and in a way that are explainable by a computer. Depending on the data set, defining a correct semantic model can require significant effort, even for experienced data modelers. A good semantic model should represent a consensus view in a particular domain, for a particular purpose. Therefore, it is good practice to conform to existing models resulting from standardization efforts.

5. Make Data Linkable: Raw data can be converted to linkable data by applying the semantic model defined in the previous step. Currently, this is done using semantic web and linked data technologies. This step promotes interoperability and reuse, making it easier to integrate data with other data types and systems. However, the user should evaluate the feasibility of this step for the given data set. It makes sense to do for many types of data but may not be relevant for other types.

6. License Attribution: The use of license grants applied to health data sets must be subject to the applicable regulatory framework for each data owner, especially when sensitive data is involved. The importance of specifying clear license terms is required for dataset reuse. Therefore, the license attribution for the dataset should be clearly stated, as well as the process by which an external requester could request permission to reuse the dataset.

7. ***Data Versioning:*** Data versioning should be handled following the international standards, best practices, and recommendations such as those of RDA[12]. The standard recommended for data versioning is the Reference Model for an Open Archival Information System (OAIS), ISO 14721:2012[13]

8. Indexing: Indexing is an important step, since searching over these datasets is one of the ultimate goals of FAIRification. Each versioned data needs to be indexed with respect to the possible search parameters over the records.

9. Metadata Aggregation: This action is performed to state the dataset data provenance, increase its quality and understandability, thus enabling its findability and reusability in further research studies. There are many metadata standards and vocabularies already available to the scientific community. purposes.

[12] Rauber, A., Asmi, A., van Uytvanck, D., & Proell, S. (2015). Data citation of evolving data: Recommendations of the Working Group on Data Citation (WGDC). *Result of the RDA Data Citation WG*, *20*, 1-2.

[13] https://www.iso.org/standard/57284.html

10. Publishing: Data publishing is the process of making FAIR data sets available on a separate storage device for long-term preservation/retention. For datasets, this is not a trivial issue, as data types and sizes can vary significantly depending on the original sources. Publishing datasets to an external repository does not imply that the data is open, as some repositories make datasets available with permissions like "on demand, upon approval".

3.1 Making data Findable, including provisions for metadata

The first step in the FAIRification process is to ensure the data can be found easily inside large data pools. Thus, both metadata and data should be easily Identifiable by both humans and machines. All the deliverables will be listed on the FAITH website (www.faith-ec-project.eu), and the ways by which FAITH output can be accessed will be communicated via social media and other suitable channels to increase visibility of FAITH work. For public deliverables, a link will be available on the FAITH website pointing to the appropriate open repositories where the data is submitted.

3.1.1 F1. Assign globally unique and persistent identifiers to data and metadata

Digital resources, i.e., data and metadata, must be assigned a globally unique and persistent identifier as these identifiers remove ambiguity in the meaning of published data by assigning a unique identifier to every element of metadata. In FAITH project, for openly available data produced by the project, such as scientific publications, a Digital Object Identifier (DOI) will be issued directly by Zenodo once they are uploaded to the Zenodo repository[14]. The assignment and management of the persistent and globally unique identifiers is addressed in this first version of the DMP deliverable and will be updated through the whole lifecycle of the project.

3.1.2 F2. Describe the project data with rich metadata

Rich metadata, including descriptive information about the context, quality and characteristics of the data allows finding data based on the information provided by their metadata, even without the need of the identifier. To enable both global and local search engines to locate a resource, generic and domain-specific descriptors should be provided. Wherever possible, curated datasets will be published on Zenodo which makes them openly accessible and discoverable. The data that will be uploaded to Zenodo will inherit the metadata description of Zenodo, which is compliant with DataCite's Metadata Schema minimum recommended terms, with a few additional enrichments. The DataCite Metadata Schema for Publication and Citation of Research Data allow data to be understood and reused by other members of the research group and add contextual value to the datasets for future publishing and data sharing. Text files metadata will be automatically generated using the DataCite Metadata Generator after filing the form requesting intrinsic metadata. All published data

[14] http://about.zenodo.org/

sets will receive a DOI that will be referred to in any scientific publication that made use of this data set. A list of metadata elements required for data citation are given in the table.

Table 15: Citation and Discovery metadata for data repositories across common standards.

Citation Metadata	Dublin Core	Schema.org	Datacite	DATS
Dataset Identifier	Identifier	@id	Identifier	identifier
Title	Title	Name	Title	Title
Creator	Creator	Author	Creator	Creator
Data repository or archive	Publisher	Publisher	Publisher	Publisher
Publication Date	Date	datePublished	publicationYear	Date
Version	N/A	Version	Version	Version
Type	Type	Type	resourceTypeGeneral	type
Description	description	description	description	datatype dimension Material…
Keywords	Subject	Keywords	Subject	keywords
License	License	License	Rights	license
Related Dataset	isPartOf isVersionOf references	isPartOf citation	relatedIdentifier	isPartOf
Related Publication	bibliographicCitation	Citation	relatedIdentifier	publication

3.1.3 F3. Clearly and explicitly include in the metadata the identifier of the data that they describe

All the descriptions of a digital resources must contain an identifier of the resource being described. This is especially important where the resource and its metadata are stored independently, but persistently linked, which is generally considered good practice in FAIR. The association between a metadata file and the dataset is made explicit by mentioning dataset's globally unique and persistent identifier in the metadata. Where applicable for specific FAITH datasets the metadata creation will be based on widely used standards. To guarantee that the connection is annotated in a formal manner, the FAIRifier tool[15] may be used. Alternatively, the FAIR Data Point, which is based on the

[15] https://github.com/FAIRDataTeam/OpenRefine-metadata-extension/

Data Catalogue model (DCAT)[16], provides not only unique identifiers for potentially multiple layers of metadata, but also a single, predictable, and searchable path through these layers of descriptors, down to the data object itself. The most prominent metadata catalogues considered by the FAITH project for collecting, managing, analyzing, visualizing and sharing data are Egeria[17], MOLGENIS[18], CKAN[19] and InvenioRDM[20]. Among these metadata catalogs, the MOLGENIS metadata catalogue seems to be the most appropriate for the FAITH project given the already high adoption of the community, its modular and extensible design and the already available modules.

3.1.4 F4. Register or index the data in a searchable resource

Digital resources must be registered or indexed in a searchable resource. This resource provides the infrastructure by which a metadata record (F1) can be discovered, using either the attributes in that metadata (F2) or the identifier of the data object itself (F3)[23]. Metadata of each record uploaded to Zenodo is indexed directly in Zenodo's search engine, immediately after publishing. Metadata of each record is sent to DataCite servers during DOI registration and indexed there[21]. Datasets can also be indexed with metadata in common search indexes, such as Google Dataset Search via schema.org[22]. Keywords will be provided, based on standard terminologies, potentially enabling multilingual search, and search at various levels of detail. In addition, metadata may be shared via FAIRSharing[23] or other relevant findability service providers that deliver both human- and machine-readable access to metadata.

3.2 Making data **Accessible**

This principle refers to easy access to the data by the users, possibly including authentication and authorization. Since some FAITH data can be confidential they will be restricted in their use. Sensitive and personal data can be made accessible only following the GDPR requirements. As several repositories will be used to store data, the policy on how to grant access to restricted results will be developed over the course of the project.

3.2.1 A1. Data and metadata are retrievable by their identifier using a standardized communications protocol

FAIR data retrieval should be mediated without specialized or proprietary tools or communication methods and that the identifier (F1) follows a globally accepted schema tied to a standardized, high-

[16] https://www.w3.org/TR/vocab-dcat/
[17] https://egeria.odpi.org/
[18] https://www.molgenis.org/
[19] https://ckan.org/features/
[20] https://inveniosoftware.org/products/rdm/
[21] http://about.zenodo.org/principles/
[22] https://datasetsearch.research.google.com/
[23] https://fairsharing.org/

level communication protocol. Its purpose is to provide a predictable way to access a resource, regardless of whether unrestricted access to the content of the resource is granted or not. The most common example of a compliant standardized access protocol is the Hypertext Transfer Protocol (HTTP23). It offers useful features, including the ability to request metadata in a preferred format, and/or to inquire as to the formats that are available. It is also widely supported by software and common programming languages. The software developed during the project is hosted on a Git-like server. Documentation and a user guide with examples will be published as an online tutorial via the website and will accompany any release.

As specified to the IPR policy of the consortium, the collected data will be shared with the scientific community, when it does not concern sensitive information. Regarding clinical data, for confidentiality reasons, data is anonymised before being shared even if such sharing takes place only with the consortium. The direct access for some datasets is accomplished through a private repository. All the other datasets for FAITH project are securely uploaded to a centralized repository which is designed and developed during the project lifecycle. For data generated within FAITH the Data Owner/Data Provider should ensure that these are of high level of granularity and include adequate metadata description. Prior to the data collection any ethical and legal concerns will be assessed and respective counter measures taken. Moreover, for the data collected outside the project (existing – not generated internally) the necessary quality assessment will be performed and anonymization process will be performed if necessary. An email, telephone number, or Skype name of a contact person who can discuss access to highly sensitive data will be provided. This contact protocol will be clear and explicit in the metadata. Hence, even heavily protected and private data will be FAIR.

3.2.2 A2. Metadata are accessible, even when the data are no longer available

In case where the data record is no longer available, there must be a clear and precise way of discovering its historical metadata record. Collected, processed and generated research should be uploaded and preserved in wide acceptable formats to ensure long-time accessibility. The selection of the adequate data format needs to be carefully assessed in order to have a high chance of being usable in the future. For publicly accessible data through the open repository Zenodo, the consortium will use the provided metadata for individual records as well as record collections which are harvestable using the OAI-PMH protocol and retrievable through the public REST API as both are open, free and universal protocols for information retrieval on the web.

3.3 Making data **Interoperable**

The FAIRification process addresses data interoperability through defining a common language for both data and its metadata. This is achieved through a specific step within the FAIRification workflow. This step involves identifying the meaning and relationships between the data elements. More specifically a dictionary for your data must be create using shared and recognized vocabularies and ontologies.

3.3.1 I1. Data and metadata use a formal, accessible, shared, and broadly applicable language for knowledge representation.

The purpose of principle I1 is to achieve a "common understanding" of digital resources through a globally understood language for machines, with an emphasis on "knowledge" and "knowledge representation". To ensure automatic findability and interoperability of datasets, commonly used controlled vocabularies, ontologies and thesauri are used to describe the dataset. The most widely accepted framework to describe and structure (meta)data is, currently, the Resource Description Framework (RDF) extensible knowledge representation model which is the W3C's recommendation for how to represent knowledge on the Web in a machine-accessible format[24]. The RDF framework provides a common and straightforward underlying model and creates a powerful global virtual knowledge graph.

3.3.2 I2. Data and metadata use vocabularies that follow FAIR principles

Using "vocabularies" to refer to the methods that unambiguously represent concepts that exist in a domain, requires that the vocabulary terms used in the knowledge representation language (principle I1) can be sufficiently distinguished by a machine, in order to ensure detection of "false agreements" as well as "false disagreements". To make the data interoperable, standard open formats are used for storage. The necessary vocabularies and wide-open standards are used to design the data schema and store the data. Proprietary software and language-dependent formats will be avoided where possible. For data, such as tabular datasets, CSV, XML or JavaScript Object Notation (JSON) is used to format the metadata. In situations where wider standards, such as Dublin Core, are needed, we will provide proper mappings.

3.3.3 I3. Data and metadata include qualified references to other data or metadata

A "qualified reference" is a reference to another resource (i.e., referencing that external resource's persistent identifier), in which the nature of the relationship is also clearly specified. This principle therefore relates to the good practice to clearly distinguish between metadata (files/containers) and the resources they describe. The scientific links between the datasets will be described, specifying if one dataset builds on another data set, if additional datasets are needed to complete the data, or if complementary information is stored in a different dataset. Furthermore, all datasets will be properly cited (i.e., including their globally unique and persistent identifiers).

[24] https://www.w3.org/RDF/

3.4 Making data **Reusable**

The goal of FAIRification process is to optimise the reuse of data. To accomplish this, metadata and data should be well-described and appropriately licensed so that they can be replicated and/or exploited in different settings. In the context of FAIRification, reusability refers to the ease with which data can be used for new research purposes beyond its original intent. FAIR principles ensure data is presented in a way that allows others to understand, integrate, and analyze it efficiently, fostering new discoveries.

3.4.1 R1. Data and metadata are richly described with a plurality of accurate and relevant attributes

The focus of R1 principle is to enable machines and humans to assess if the discovered resource is appropriate for reuse, given a specific task. This reiterates the need for providers to consider not only high-level metadata facets that will assist in generic search (as described by principle F2), but also to consider more detailed metadata that will provide much more "operational" instructions for re-use. For datasets and the scientific articles freely available in the Zenodo repository, the re-use principle is inherently supported since: (1) Each record contains a minimum of DataCite's mandatory terms, (2) License is one of the mandatory terms in Zenodo's metadata and is referring to an Open Definition license, (3) Data downloaded by the users is subject to the license specified in the metadata by the uploader, and (4) Zenodo is not a domain-specific repository

3.4.2 R2. Data and metadata are released with a clear and accessible data usage license.

Digital resources and their metadata must always include a license that describes under which conditions the resource can be used. In order to facilitate reuse, the license chosen should be as open as possible. The FAITH team recognizes the importance of software licensing from the outset of the project. Therefore, data and metadata is released with a clear and accessible data usage license, like MIT (for software codes) or Creative Commons (for datasets) since these are the best option between unrestricted access and the promotion of a fair community practice that acknowledges the provenance of data. In the cases where specific service information cannot be publicly shared, the reasons will be mentioned in their metadata descriptions (e.g., ethical, personal data, intellectual property, commercial, privacy-related, security-related).

Provenance descriptions are implemented following community specific templates exploiting MOLGENIS framework and the common data and metadata mode created for the specific purpose. In addition, metadata registries, shall be used to choose domain-specific standards taking into full consideration the relevant inter-domain interoperability requirements. A common data and metadata model has been designed and exploited through MOLGENIS framework. Given the uncertainty of the potential value of the developed tools in the future, the exact licenses to be used in case-by-case situation will be refined through the course of the project.

3.5 Implementation of data FAIRification

In the context of the FAITH project, we are committed to integrating the FAIR (Findable, Accessible, Interoperable, Reusable) principles across the entire data modeling lifecycle. This implementation extends from the collection and curation of data to the preparation of training data, AI model training and validation, and the eventual deployment of AI models. As such, data are findable, as persistent and globally unique ids are assigned to each dataset coupled with standardized metadata. In addition, the publication of datasets at Zenodo gives visibility to the data and allow us to obtain a DOI (Digital Object Identifier), which serves as a globally unique reference (Permanent ID Url) for each dataset. This standardization of data facilitates seamless data exchange and collaboration across the healthcare and AI research domains. Finally, the FAITH data are also reusable with all the steps of the data creation processes to be meticulously recorded. Moreover, privacy protection is prioritized through advanced anonymization strategies and implementation of data access licenses, for enhancing data reuse in the future.

3.6 Evaluation of data FAIRness

A final step in the post-FAIRification phase is to assess the FAIRness of the data. This process may include: 1) an evaluation to check whether the original objectives have been achieved (if not, some of the steps in the workflow may need to be revisited), and 2) checking the FAIR status of the data and metadata using FAIRness assessment tooling. Thus, tools developed for conducting FAIRness evaluations can be based on either discrete/ open-answer evaluation questionnaires or semi-automated evaluation models[25]. Communities have already published documents that can guide implementation choices. Some examples are "the FAIR metrics"[26] and the follow-up Maturity Indicators[27] and the FAIR Convergence Matrix[28]. In, addition, self-assessment models for measuring the maturity level of a dataset have also been developed such as the RDA FAIR Data Maturity Model[29]. On top of that, some evaluator services not only run the FAIR metrics evaluations, but also are intended to deliver a certified report regarding compliance with the FAIR Principles and the resulting level of FAIRness[30]. In FAITH the evaluation of data FAIRness is an activity of high importance and shall be achieved by utilizing the rich collected metadata through the MOLGENIS data and metadata model.

[25] de Miranda Azevedo, R., & Dumontier, M. (2020). Considerations for the conduction and interpretation of FAIRness evaluations. *Data Intelligence*, *2*(1-2), 285-292.

[26] Wilkinson, M. D., Sansone, S. A., Schultes, E., Doorn, P., Bonino da Silva Santos, L. O., & Dumontier, M. (2018). A design framework and exemplar metrics for FAIRness. *Scientific data*, *5*(1), 1-4.

[27] Wilkinson, M. D., Dumontier, M., Sansone, S. A., Bonino da Silva Santos, L. O., Prieto, M., Batista, D., ... & Schultes, E. (2019). Evaluating FAIR maturity through a scalable, automated, community-governed framework. *Scientific data*, *6*(1), 174.

[28] H.P. Sustkova et al. FAIR convergence matrix: Optimizing the reuse of existing FAIR-related resources. Data Intelligence (2020); https://www.go-fair.org/today/fair-matrix/

[29] https://www.rd-alliance.org/groups/fair-data-maturity-model-wg

[30] https://www.gofairfoundation.org/certification/

4 General Data Protection Regulation (GDPR)

4.1 General

EU citizens are granted the rights to privacy and data protection by the Charter of Fundamental Rights of the EU. Article 7 states that "everyone has the right respect for private and family life, home and communications", whereas Article 8 regulates that "everyone has the right to the protection of personal data concerning him or her" and that processing of such data must be "on the basis of the consent of the person concerned or some other legitimate basis laid down by law." These rights are developed in detail by the General Data Protection Regulation (GDPR), Regulation 2016/679/EC, which is in force in every Member State since 25th of May 2018[31].

The European Union's General Data Protection Regulation (GDPR) stands as a cornerstone of data privacy in the bloc. Its reach extends to the realm of scientific research as well, specifically impacting programs like Horizon Europe, the EU's flagship initiative for funding research and innovation. This section delves into the analytical exploration of how GDPR regulations are applied to FAITH project.

The GDPR aims to further protect the personal data of individuals and their free movement within the EU. The GDPR applies to all entities that are either fully established in the EU or have branches established in the EU that process personal data as part of their activities, regardless of where the data is processed. It also applies to entities established outside of the EU, which offer goods/services to individuals in the EU or monitor the behaviour of such individuals within the EU.

Therefore, since the 25th of May 2018, not only applicants, beneficiaries, contractors or subcontractors receiving funding from EU programmes such as H2020, COSME, EMFF, LIFE, the SME Instrument or EEN, but also trainers and experts, must comply with the GDPR. Any natural or legal person who collects or in any way uses for professional purposes the personal data of individuals must comply with the new rules, as is the case with any other EU or national legislation they are subject to.

The GDPR applies only to the processing of personal data. Since the EU data protection legislation only deals with the processing of personal data, the distinction of personal and non-personal data (which includes anonymous data) is crucial for all activities of the project. Article 4(1) GDPR defines personal data as:

> *any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one 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*.

Personal data, as defined by the GDPR, is any information related to an identified or identifiable natural person, i.e., names, identification numbers, emails, postal addresses, phone, location data, pictures, signatures, etc. This excludes information about companies, anonymised or statistical data,

[31] https://eur-lex.europa.eu/content/news/general-data-protection-regulation-GDPR-applies-from-25-May-2018.html

which is not personal data. Processing means any operation performed on the personal data, such as collecting, recording, storing, organising, filing, using, combining, disclosing, transferring, or erasing manually or automatically, i.e., collecting contact details of participants to an event, sending newsletters by email, publication of participants lists or pictures with persons related to an event, subscription to e-services etc.

Moreover, more privacy-friendly approaches such as utilizing synthetic data and dummy or fake data will be considered by partners for FAITH research activities, noting that legal categorization of those data (whether anonymous data or personal data) is not very clear from a legal point of view. Furthermore, anonymization or implementing any privacy-friendly process is likely to eliminate any possible risks to the rights and freedoms of data subjects, as it is no longer possible to associate the data with a data subject. Therefore, the use of non-personal data is strongly advised whenever this is possible and considered functional for the purposes of the processing. In the FAITH project, the Partners decided to exclusively work with anonymized data to the extent the research objective can be achieved in this way. **Therefore, GDPR will not be applicable for specific FAITH research activities that do not rely on processing personal data**. However, the anonymization process, where personal data is transformed from personal to non-personal data, is considered data processing; therefore, this step falls within the scope of the GDPR. Overall, FAITH Partners will ensure compliance with the GDPR rules when they are processing personal data for research purposes, even in the case of carrying out 'anonymisation ' operations. In the following sections, the legal basis for the anonymization will be briefly described.

4.2 Legal requirements of anonymization

The requirements for anonymization are not definitively outlined in the GDPR, which is why the possibility for incomplete anonymization was pointed out as one of the main risks in the DPIA. The main legal conditions for a GDPR-compliant anonymization are described below.

As mentioned above, the GDPR is only applicable to the processing of personal data as set out in Article 1(1). Recital 26 GDPR explicitly states that:

> *The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable. This Regulation does not therefore concern the processing of such anonymous information, including for statistical or research purposes*.

The Article 29 Data Protection Working Party ('WP29') describes anonymization as a "*technique applied to personal data in order to achieve irreversible de-identification*".[32] However, the process of anonymization is often confused with the processes of pseudonymization. Personal data which has undergone pseudonymization and which could be attributed to a natural person with the use of

[32] *Donnelly/McDonagh*, Health Research, Consent and the GDPR Exemption, European Journal of Health Law 26 (2019) 100; Opinion, 05/2014 on Anonymisation Techniques, adopted 10 April 2014, 0829/14/EN, 7.

additional information, cannot be described as anonymized for GDPR purposes[33] as it allows for the re-identification of the data subject. Many of the pseudonymization techniques commonly used in health research, e.g. key coding of data, may not be sufficient to take the processing outside of the scope of the GDPR requirements.[47] There is a degree of uncertainty as regards the borderline between pseudonymization and anonymization for GDPR purposes.[47] The primary distinction between pseudonymization and anonymization is that the former allows for the re-identification of the data subject, whereas the latter process does not.

In assessing the possibility and probability of the re-identification of the data subject, Recital 26 of the GDPR requires an objective assessment of the measures which are likely to enable re-identification, such as the costs of identification, the time required for identification and the available technology and foreseeable technological developments. To fulfil the above-mentioned legal requirements and ensure anonymization, as long as the objective of research activity can be achieved with anonymized data, the partners are not storing any information which allows for the identification, whether directly or indirectly, of the concerned data subject. The Partners shall utilize the anonymization tools, which are agreed upon within the Consortium. If the anonymization is not meaningful for the pursued research activity, then FAITH Partners will rely on pseudonymization techniques and secure encryption methods to process needed personal data to achieve the research objective in compliance with the applicable data protection rules including the GDPR.

4.3 Data protection principles and envisaged measures

The GDPR establishes a set of principles and requirements that the FAITH consortium shall comply with as also promised in the FAITH consortium Agreement(Section 4.5) and FAITH Grant Agreement. FAITH Partners will comply with all the requirements specified by the GDPR to ensure full respect for the principles relating to the processing of personal data. In particular, the project shall adhere to the data protection principles of:

- **Lawfulness, fairness and transparency:**
    - In accordance with these principles, data must be processed with respect to the law, proportionally to the aim foreseen and transparently towards the data subjects concerned. FAITH will process personal data in compliance with the GDPR and other national or European applicable legislation that applies in the context of the project. FAITH will also process data fairly by balancing the data processing needs of the consortium and the rights and interests of the data subjects. FAITH shall also process data in a transparent manner, by providing information to the natural persons concerned about the collection, use and storage of their data as well as the extent of these operations, following the informed consent procedures decided by FAITH Partners. All research processes and procedures will be transparent to all stakeholders.

[33] *Donnelly/McDonagh*, Health Research, Consent and the GDPR Exemption, EJHL 26, 100.

- **Purpose limitation:**
    - The collection and processing of personal data should be limited to specified, explicit and legitimate purposes. Following this principle, FAITH will take appropriate technical and organisational measures to ensure that, by default, only personal data which are relevant to the envisaged research are collected and processed. Personal data will be used in FAITH to pursue research objectives and tasks indicated in the FAITH Consortium and Grant Agreements. Among others, this also includes:
        - To disseminate FAITH results to stakeholders
        - To communicate news and information about the project to the public, the media and civil society organisations.
        - To support stakeholders in the exploitation of FAITH results.
- **Data minimisation:**
    - This principle entails the need for FAITH partners to ensure that personal data being processed is adequate (i.e., sufficient to properly fulfil the stated purposes of the project), relevant (i.e. the personal data has a rational link to FAITH research purposes) and limited to what is necessary (i.e. FAITH partners shall not hold more than what is necessary for the purposes of the research). FAITH shall take appropriate technical and organisational measures to ensure that, by default, only personal data which are relevant to the envisaged research are processed. In compliance with the data minimisation principle, FAITH will assess whether the same purposes can be achieved by collecting less data than initially intended and, where that is the case, apply the narrower collection option available.
- **Accuracy:**
    - Personal data shall be accurate and, where necessary, kept up to date. In accordance with this principle, FAITH will take every reasonable step to ensure that the data being processed is accurate and kept up to date. Accordingly, and having regard to the purposes for which the data are processed, FAITH partners shall erase or rectify data without delay.
- **Storage limitation:**
    - In line with this principle, FAITH shall not keep personal data for longer than is necessary for the purposes of the project. To address this, FAITH will ensure that personal data from volunteers is kept for as long as necessary and, where appropriate, in an anonymised or pseudonymised manner. Personal data collected during the FAITH research will be deleted incline with the agreed storage duration.
- **Integrity and confidentiality:**
    - According to this principle, FAITH partners (as data controllers of the processing) must have appropriate security measures in place to protect the personal data held. Once the personal data has been used for its intended purposes within the project, it will be deleted to avoid accidental risk of future disclosure (unless required to be kept for legal

or contractual purposes). Such measures and procedures are intended to safeguard ethical values, such as protecting the rights and autonomy of individuals to the fullest.

- **Accountability:**
  - Accountability is the grounding principle of the GDPR. It requires the controller to adhere to and demonstrate compliance with the Regulation and the above principles. Accountability in the context of FAITH may translate, for example, into carrying out the necessary evaluations about legal and ethical procedures for FAITH research activities entailing the use of personal data. Accordingly, FAITH partners will ensure the accountability and responsible handling of the personal data processed within the FAITH Project.

4.4 Involvement of research volunteers

As outlined above, certain FAITH research activities (including specific FAITH LSPs) might involve adult volunteers for research purposes and the processing of personal data. The FAITH consortium has considered dedicated procedures tailored to the FAITH research operations in particular, including LSPs to ensure that the fundamental rights of all human participants are treated with due care. As indicated above, the FAITH consortium does not envisage the extensive collection or processing of personal data to achieve its objectives but will mostly collect the data to pursue research objectives identified in the FAITH Grant and Consortium Agreements (e.g., within the context of certain LSPs) as well as the contact information of all research participants for the management of the project and the dissemination activities.

5 FAITH Open Data

5.1 Open access to scientific publications

Open Science means sharing knowledge and tools as early as possible, not only between researchers and between disciplines but also with society at large. Open access publishing (also called 'gold' open access), meaning that an article is immediately provided in open access mode by the scientific publisher, or self-archiving (also called 'green' open access) meaning that the published article or the final peer-reviewed manuscript is archived by the researcher- or a representative -in an online repository before, after or alongside its publication will be adopted. Authors copyrights agreements will determine whether scientific publications, resulted from the project, will adopt the gold or the green model. However, in the case copyright agreements are not violated (e.g. in the case of peer reviewed journals and international conference proceedings), the consortium will favor whichever model guarantees wider dissemination of the project results.

5.2 EU Recommendations on open data access

The following recommendations concerning open data access has been reported:

Commission Recommendation 2012/417/EU on access to and preservation of scientific information, , where EU Member State has nominated a National Point of Reference, with the task of reporting on the implementation of open access in the Member States. The 2012 Recommendation on access to and preservation of scientific information (2012/417/EU) was part of a package that outlined measures to improve access to scientific information produced in Europe and to bring them in line with the Commission's own policy for Horizon 2020[34].

Commission Recommendation (EU) 2018/790on access to and preservation of scientific information[35], that explicitly reflects developments in areas such as research data management (including the concept of FAIR data, i.e. data that is Findable, Accessible, Interoperable and Re-usable), Text and Data Mining (TDM) and technical standards that enable re-use incentive schemes. It reflects ongoing developments at the EU level of the European Open Science Cloud, and it more accurately considers the increased capacity of data analytics of today and its role in research. It also clearly identifies as two separate points the issue of reward systems for researchers to share data and commit to other open science practices on the one hand, and skills and competences of researchers and staff from research institutions on the other hand[36].

[34] http://ec.europa.eu/research/openscience/pdf/openaccess/background_note_open_access.pdf#view=fit&pagemode=none

[35] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32018H0790

[36] http://ec.europa.eu/research/openscience/index.cfm

6 Risks

The FAITH project identifies and addresses a comprehensive range of data management risks to ensure the integrity, security, and ethical handling of data throughout its lifecycle and the LSPs implementation. Firstly, compliance with data protection regulations, particularly the General Data Protection Regulation (GDPR), is paramount. Non-compliance can result in substantial fines and reputational damage. Risks include improper handling of personal data, inadequate data subject consent procedures, and failure to uphold data subjects' rights such as the right to access, rectify, and erase their data. Mitigation strategies involve rigorous adherence to GDPR guidelines and ensuring all data processing activities are transparent and consensual. In FAITH the data security risks are another critical concern. These encompass unauthorized access to sensitive data, data breaches, and data loss. Such incidents can lead to significant legal repercussions and loss of stakeholder trust. To mitigate these risks, robust data security measures will be implemented by the LSP leaders and participants, including encryption, secure access controls, regular security audits, and ensuring data is stored in trusted repositories with redundant backups to prevent data loss.

Ethical risks in data management also require close attention. These include risks related to the ethical collection, storage, and usage of data, particularly when involving vulnerable populations or sensitive information. Ethical risks might lead to public backlash and undermine the ethical integrity of the project. To address these, all data collection and processing activities will comply with established ethical standards and guidelines. Ethical review boards should oversee the data management practices, and informed consent must be obtained from all participants whenever necessary.

Technical risks related to data integrity and interoperability can affect the project's and large scale pilot outcomes. These risks include data corruption, incomplete data sets, and challenges in integrating data from diverse sources. Thus FAITH consortium will set the necessary infrastructure and methods to ensure data integrity by implementing strict version control, regular data validation checks, and maintaining comprehensive metadata for all datasets to facilitate interoperability and reproducibility of research findings.

Moreover data sharing and open access risks are significant in the context of adhering to FAIR (Findable, Accessible, Interoperable, Re-usable) principles. Risks include inadequate anonymization of data leading to potential re-identification of individuals, and failure to provide long-term accessibility and usability of data. In FAITH mitigation involves employing advanced anonymization techniques, clearly defined data sharing policies, and ensuring data is deposited in reputable, long-term repositories with persistent identifiers such as DOIs.

By addressing these diverse risks through a combination of legal, ethical, and technical measures, the FAITH project can ensure the responsible management of data, safeguarding both the project's integrity and the rights and privacy of data subjects

7 Ethical aspects

FAITH Partners will comply with the EU Regulation passed by the European Parliament, Establishing the Horizon 2020, Article 19 ('Ethical principles') sets out that:

> "All the research and innovation activities carried out under Horizon 2020 shall comply with ethical principles and relevant national, Union and international legislation, including the Charter of Fundamental Rights of the European Union and the European Convention on Human Rights and its Supplementary Protocols. Attention shall be paid to the principle of proportionality, the right to privacy, the right to the protection of personal data, the right to the physical and mental integrity of a person, the right to non-discrimination and the need to ensure high levels of human health protection."[37]

FAITH Grant Agreement re-emphasizes the obligation that all beneficiaries need to comply with ethical and research integrity principles. Failing to incorporate these values would not only indicate irresponsible research that results in outputs of questionable value that may be seen as unreliable and high-risk but would also constitute a breach of a beneficiary with the abovementioned obligations and may lead to significant adverse effects for the human subjects involved.

The FAITH project's partners must carry out the research actions in compliance with:

a) ethical principles (including the highest standards of research integrity) and

b) applicable EU and national law and conform to the ethics standards and guidelines of H2020.

Therefore, the FAITH consortium considers ethics as an integral part of research from beginning to end, and ethical compliance is seen as pivotal to achieve real research excellence. Ethical compliance will furthermore facilitate public trust in the FAITH solution and increase credibility in the project's outputs. While there are specific data protection and human rights laws that this project will fully adhere to ensuring the protection individuals, there are however no specific EU laws regarding ethics. In the words of the European Data Protection Supervisor: "Ethical thinking and deliberation come before, during, and after the law".[38]

The fact that our research is legally permissible does not necessarily mean that it will be deemed ethical. Therefore, the responsibility lies within each participating partner to conduct their tasks in ways that include respect and protection of human values, which are intrinsic to the existing legislation and fully adhere to the highest ethical standards, as set out for example in The European Code of Conduct for Research Integrity.[39].

[37] Regulation (EU) No 1291/2013 of the European Parliament and of the Council of 11 December 2013 establishing Horizon 2020 - the Framework Programme for Research and Innovation (2014-2020) and repealing Decision No 1982/2006/EC Text with EEA relevance

[38] https://edps.europa.eu/sites/edp/files/publication/19-03-25_reuters_interview_en.pdf

[39] ALLEA - All European Academies, The European Code of Research Integrity (2017) https://www.allea.org/wp-content/uploads/2017/05/ALLEA-European-Code-of-Conduct-for-Research-Integrity-2017.pdf

Notably, this implies compliance with the following essential principles:

- Honesty
- Reliability
- Objectivity
- Impartiality
- Open communication
- Duty of care
- Fairness and
- Responsibility for future science generations.

This means that the coordinators must ensure that persons carrying out research tasks:

- present their research goals and intentions in an honest and transparent manner;
- design their research carefully and conduct it in a reliable fashion, taking its impact on society into account;
- use techniques and methodologies (including for data collection and management) that are appropriate for the field(s) concerned- taking into account the specificities of the large-scale pilots;
- exercise due care for the subjects of research — be they human beings, animals, the environment or cultural objects;
- ensure objectivity, accuracy and impartiality when disseminating the results;
- allow — as much as possible and taking into account the legitimate interest of the beneficiaries — access to research data, in order to enable research to be reproduced;
- consider the potentially sensitive nature of the research and refrain from publishing or disseminating research results, datasets or protocols that might be misused, infringe on the rights of others or cause harm to the persons involved;
- make the necessary references to their work and that of other researchers;
- refrain from practising any form of plagiarism, data falsification or fabrication;
- avoid double funding, conflicts of interest, misrepresentation of credentials or other research misconduct.

To the extent it fulfills the research objectives, the consortium uses anonymization techniques in order to remove all identifiers, which could allow reidentification of an individual. If anonymization does not fulfill the pursued research objective, pseudonymization and encryption techniques will be utilised to ensure secure and responsible data processing operations. Overall, FAITH Partners will take all necessary technical and organisation measures to use the least intrusive methods to collect, process, and store the data and to ensure data security with the ultimate aim of compliance with the regulations and ethics principles. However, to mitigate the remaining risk of insufficient anonymization, the access and terms of use of the data repository are strictly regulated by the Terms & Conditions of the repository.

8 References

In the current deliverable the necessary references are introduced in the document as footnotes.

Appendix I – Dataset description template

The following table depicts the template format where the different datasets are described including valuable metadata for an efficient Data Management Monitoring.

Table 16: DS.#.# table template

Data identification: *DS#.#_* <dataset name>	
Generic description:	
<Provide a short description of the dataset>	
Origin of data:	
<Describe here the origination of the data that compile the specific dataset>	
Nature and scale of data:	
<file format of dataset>	
To whom the dataset could be useful:	
<Describe who could utilise/exploit the specific dataset>	
Related scientific publication(s)	
<Is the dataset related to a scientific publication? Is the latter Gold or Green Open Access?>	
Indicative existing similar data sets (including possibilities for integration and reuse):	
<Are there public available datasets similar to the specific one? If yes provide details.>	
Partners activities and responsibilities	
Partner owner of the data	-
Partner in charge of the data analysis	-
Partner in charge of the data storage	-
Related WP(s) and task(s)	-
Standards and metadata	
Standards, format, estimated volume of data	

	<Are there Standards that the dataset complies to? What is the dataset format? What is the estimated size of it?>
Data exploitation and sharing	
Data access policy/ Dissemination level: confidential (only for members of the Consortium and the Commission Services) or Public	<Classification of the Dissemination Level>
Data sharing, re-use, distribution, publication (How?)	-
Personal data protection: are they personal data? If so, have you gained (written) consent from data subjects to collect this information?	<Does the dataset include personal sensitive data?>
Access Procedures	-
Embargo periods (if any)	-
Archiving and preservation (including storage and backup)	
Data storage (including backup): where? For how long?	<Where are the data stored? What type of backup process is planned?
Indicative associated costs for data archiving and preservation	-
Indicative plan for covering the above costs	-
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 securely stored in trusted repositories for long-term preservation and curation?	
Ethics	
Describe any potential ethical issues during data collection, storage, processing and archiving, together with the ethical approval procedures related to the project.	
If the research activities involve children, patients, members of vulnerable populations, use of embryonic stem cells, privacy and data protection issues or research on animals and primates, the	

ethical principles and relevant national, EU and international legislation must be complied with.	

Appendix II – Tools and services for data FAIRification

Tool/Service	Reference	Short Description
Raysync	https://www.raysync.io/	Raysync.io offers features that can aid in maintaining data integrity during transfers, which is crucial for ensuring data hasn't been tampered with. Raysync.io utilizes a proprietary protocol for high-speed file transfers and implements robust security measures like access controls and encryption to safeguard data throughout the process. This combination helps minimize the risk of data corruption and unauthorized access, contributing to overall data trustworthiness.
Registry of Research Data Repositories	https://www.re3data.org/	re3data.org plays a vital role in the findability aspect (F) of FAIR data. re3data.org functions as a comprehensive registry that indexes information about research data repositories around the world. By searching re3data.org, researchers can discover datasets relevant to their field, ensuring they find data that adheres to FAIR principles. This promotes data discovery and accessibility, which are crucial first steps for data reuse and verification of its integrity.
Zenodo	https://zenodo.org/	Zenodo is a general-purpose open-access repository developed under the European OpenAIRE program and operated by CERN. It allows researchers to deposit research papers, data sets, research software, reports, and any other research related digital artifacts. For each submission, a persistent digital object identifier (DOI) is minted and supports various data and license types. One supported source is GitHub repositories. Zenodo is compliant with the data management requirements of Horizon 2020 and Horizon Europe, the EU's research and innovation funding programmes. Zenodo acts as a launchpad for FAIR data as It facilitates all four FAIR principles. Zenodo assigns a unique Digital Object Identifier (DOI) to your data, making it easily discoverable through search engines and research platforms. It also provides open access to your data, allowing anyone to download it with minimal restrictions. Zenodo encourages the use of standard data formats and rich metadata descriptions, promoting compatibility with various analysis tools. Finally, by making data openly accessible and well-described, Zenodo fosters data reuse by other researchers, accelerating scientific progress.
OpenRefine	https://github.com/OpenRefine/OpenRefine	OpenRefine shines as a powerful tool for data wrangling, which is a key step in the fairification process. It allows you to clean, transform, and enrich your data, making it more findable, accessible, interoperable, and reusable (FAIR).

		OpenRefine tackles issues like inconsistencies, missing values, and incorrect formatting, all of which can hinder data quality and ultimately, its trustworthiness. By using OpenRefine, you can ensure your data is well-structured, documented, and adheres to established standards, making it easier for others to understand, utilize, and verify its integrity. This FAIRifier enables a post-hoc FAIRification workflow: load an existing dataset, perform data wrangling tasks, add FAIR attributes to the data, generate a linked data version of the data and, finally, push the result to an online FAIR data infrastructure to make it accessible and discoverable. Literal values in a dataset can be replaced by identifiers either manually or by embedded, customizable script expressions. The interoperability of the dataset can be improved by connecting these identifiers into a meaningful semantic graph-structure of ontological classes and properties using the integrated RDF model editor. A provenance trail automatically keeps track of each modification and additionally enables "undo" operations and repetition of operations on similar datasets. A FAIR data export function opens up a metadata editor to provide information about the dataset itself.
Apache Spark	https://spark.apache.org/	Apache Spark can be a valuable tool in supporting FAIR data practices by facilitating data processing tasks that contribute to reusability. Spark excels at large-scale data processing, allowing researchers to efficiently clean, transform, and analyze datasets. This can help ensure data consistency and quality, which are important aspects of data integrity. Additionally, Spark's ability to work with various data formats improves data accessibility and interoperability, making it easier for others to reuse the data for further analysis.
Dryad	https://datadryad.org/stash	Dryad is an international open-access repository of research data, especially data underlying scientific and medical publications. Dryad is a curated general-purpose repository that makes data discoverable, freely reusable, and citable. Dryad serves as a repository for tables, spreadsheets, flat files, and all other kinds of published data for which specialized repositories do not already exist. All data files are are attributed a DOI and are made available for reuse under the terms of a Creative Commons Zero waiver. Dryad's metadata is supported by a Dublin Core metadata application profile.
Dublin Core Metadata	https://www.dublincore.org/	The Dublin Core Metadata Initiative (DCMI) contributes to data fairification by promoting interoperability and

Initiative (DCMI)		reusability (IR) of data. DCMI provides a set of widely adopted vocabulary terms for describing data. When datasets use these standard terms consistently, it becomes easier for machines and humans to understand the data's meaning and context. This standardization facilitates data discovery and integration with other datasets, allowing for broader use and verification. While DCMI doesn't directly address findability or accessibility (FA) of data, its focus on interoperability lays the groundwork for FAIR data practices.
FAIR Data Point (part of Data FAIRport)	https://www.dtls.nl/fai r-data/find-fair-data-tools/	FAIR Data Point (FDP) is software that allows data owners to expose metadata and data in a FAIR manner. It offers a graphical user interface (GUI) for human clients and an application programming interface (API) for software clients. FDP acts as a building block for data FAIRification as it provides a standardized format for publishing rich metadata about datasets. This metadata describes the data itself, its origin, access conditions, and any relevant details that contribute to understanding and using the data.
Data Stewardship (DS) Wizard	https://ds-wizard.org/ https://github.com/ds - wizard	The DS Wizard is based on FAIR Data Stewardship, in which each data-related decision in a project acts to optimize the FAIRness of the data, explicitly guiding researchers in order to make their results FAIRer. The Data Stewardship (DS) Wizard acts as a guide for researchers and data stewards to achieve FAIRification of their data. It functions through a series of questions that prompt users to consider various aspects of data management, all geared towards making data Findable, Accessible, Interoperable, and Reusable. By following the DS Wizard's prompts, researchers can establish best practices for data documentation, storage, and sharing, ultimately contributing to the trustworthiness and long-term value of their research data.
FAIR Search Engine (part of Data FAIRport)	https://www.dtls.nl/fair-data/find-fair-data-tools/	The FAIR Data Search Engine harvests the metadata available on FAIR Data Points or compatible data repositories, indexes them, and provides a search interface.
ORKA (part of Data FAIRport)	https://www.dtls.nl/fair-data/find-fair-data-tools/	The Open, Reusable Knowledge graph Annotator (ORKA) supports easy human curation of knowledge graphs by offering graph annotation as a service and capturing the provenance of the annotator and the original statement.
FigShare	https://figshare.com/	Figshare is an online open access repository where researchers can share and preserve their research outputs, including figures, datasets, images, and videos. The files can be uploaded in any format and items are attributed a DOI. It is free to upload content and free to access, in adherence to

		the principle of open data. All files are released under a Creative Commons license. The main hosting mechanism for the platform is Amazon S3.which supports backup and preservation via a distributed cloud computing network. Figshare positions itself as a strong proponent of FAIR data as it functions as both a repository and a facilitator for data fairification by offering the following functionalities. It allows researchers to deposit their research outputs, including datasets, in a persistent and publicly accessible location. This enhances the findability (F) and accessibility (A) of the data. Figshare encourages the use of rich metadata standards when uploading data. This detailed information description contributes to the interoperability (I) of the data, making it easier for others to understand and integrate it with their own research. Finally, Figshare enables researchers to track different versions of their data and assign appropriate licenses. This transparency fosters trust in the data's integrity and reusability (R), allowing others to confidently use and build upon the findings.
FAIRsharing	https://fairsharing.org/	FAIRsharing is a web-based, searchable portal and FAIR-supporting resource that provides an informative and educational registry on data standards, databases, repositories and policy, alongside search and visualization tools and services that interoperate with other FAIR-enabling resources. FAIRsharing guides consumers to discover, select and use standards, databases, repositories and policy, and producers to make their resources more discoverable, more widely adopted and cited. Each record in FAIRsharing is curated in collaboration with the maintainers of the resource themselves. FAIRsharing acts as a bridge between data producers and consumers in the realm of data fairification. It functions as a comprehensive resource that maps the landscape of scientific data standards, repositories, and policies. It increases the visibility and discoverability of data standards, repositories, and policies relevant to FAIR data management. This empowers researchers to find the appropriate tools and resources to make their data FAIR. It fosters connections between data producers and consumers by making them aware of established standards and best practices. This shared understanding ensures data is deposited in repositories that meet FAIR criteria, promoting findability, accessibility, and interoperability. By promoting established data standards and policies, Fairsharing.org indirectly contributes to the reusability of data. Researchers can find data that adheres to common formats and protocols,

		facilitating its integration with other datasets for further analysis. In essence, Fairsharing.org acts as a facilitator, promoting practices and resources that empower researchers to make their data FAIR and ensure its long-term value and trustworthiness.