D7.2 RESEARCH DATA MANAGEMENT PLAN

Revision: v.1.1

Work package	WP 7
Task	Task 7.1
Due date	30/11/2021
Submission date	30/11/2021
Deliverable lead	TU Delft
Version	1.1
Authors	Aaron Ding, Cornelis van de Kamp & Nicolas Dintzner (TU Delft)
Reviewers	Nikolay Tcholtchev (FOCUS) Shen Wang (UDC) Huber Flores (TARTU) Jason Pridmore (EUR) João Fernando Ferreira Gonçalves (EUR) Nicolas Kourtelis (TID) Souneil Park (TID) Miguel Garcia (AUS)

Grant Agreement No.: 101021808
Call: H2020-SU-DS-2020
Topic: SU-DS02-2020
Type of Action: RIA

Abstract	The Data Management Plan describes how data is collected, documented, archived, shared and reused during and post project, in secure and privacy-safeguarding ways.
Keywords	Data; Security; Privacy; Management;

Document Revision History

Version	Date	Description of change	List of contributor(s)
V0.1	15/11/2021	Initial version prepared	van de Kamp
V0.2	19/11/2021	Feedback PI and TUD Data Stuart incorporated	Ding, van de Kamp & Dintzner
V1.0	23/11/2021	Revision by all WP leaders	See reviewers above
V1.1	29/11/2021	Final version ok-ed by TUD data steward	Dintzner (TU Delft)

DISCLAIMER

COPYRIGHT NOTICE

Project funded by the European Commission in the H2020 Programme		
Nature of the deliverable:	R	
Dissemination Level		
PU	Public, fully open, e.g., web	✓
CL	Classified, information as referred to in Commission Decision 2001/844/EC	
CO	Confidential to SPATIAL project and Commission Services	

* R: Document, report (excluding the periodic and final reports)

DEM: Demonstrator, pilot, prototype, plan designs

DEC: Websites, patents filing, press & media actions, videos, etc.

OTHER: Software, technical diagram, etc

SPATIAL project is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101021808.

EXECUTIVE SUMMARY

The purpose of this deliverable is to provide Data Management Plan for SPATIAL. The Data Management Plan describes how data is collected, documented, archived, shared and reused during and after the project, in secure and privacy-safeguarding ways. Collection and processing of data raising ethical questions has been carefully outlined in the Ethics-related deliverables (D8.1 and D8.5). Aspects on open access for scientific research are explained in Deliverable D7.3 Project Management Plan.

The following governance measures will be adopted to take care of implementation of the presented ways of working:

- WP leaders are responsible for adhering to the Data Management Plan specifications for their respective work package;
- For the overall project, TU Delft will be responsible for complying with the Data Management Plan;
- All consortium partners are responsible for making sure personnel working on the project have read the Data Management Plan and adopted the principles.

Various types of data will be collected in SPATIAL, such as requirements, risks, threat models, case studies, (AI) model test and simulation results, data from surveys, and statistic data. Also, raw data will be generated in numerical simulations of the different algorithms and explainable AI schemes and frameworks to be applied in the course of the project.

We consider the Data Management Plan to be a living document that can be updated over the course of the project. Some decisions in this document may be adjusted when new insights are obtained in case they require further specification of the tools and platforms, along with the research in the Work Packages. The Management Committee is the body to decide on changes.

SPATIAL project is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101021808.

TABLE OF CONTENTS

SPATIAL project is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101021808.

ABBREVIATIONS

D	Deliverable
DOI	Digital Object Identifier
TUD	Delft University of Technology
EU	European Union
FAIR	Findability, Accessibility, Interoperability, and Reuse (of digital assets)
AIS	Automatic Identification System
QA	Quality Assurance
WP	Work Package

SPATIAL project is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101021808.

1 SECTION: DATA SUMMARY

In this chapter the means of data collection for each Work Package (WP) are being described, as indicated by the Work Package leaders. For each WP the following items (as far as applicable) are concerned: type, format and size of the data, way of documentation, reproducibility and usage of pre-existing data.

Each work package leader will be informed of the requirements of the project (openness, FAIR data). They will decide how, given the confidentiality and security of the data they process, what is the best operational approach to accomplish those objectives depending on the nature of the collected data (industrial partner's data, personally identifiable information).

TABLE 1: PER WP IMPORTANT ASPECTS TO MANAGEMENT OF INVOLVED DATA

WP		A. Purpose of the data collection and relation to project objectives B. Types and formats of data C. Will existing data be re-used? D. The origin of the data E. The expected size of the data F. Data utility
WP1 Requirement and threat modelling	A.	Existing data (literature) from various sources on AI is collected and analysed for the development of the SPATIAL project and the WP5 SPATIAL deployments. The work will generate new literature reviews.
	B.	MetaData: .txt format Other: .xlsx., .docx Literature references (.bib and/or tabular format: xlsx or csv) List of available data sets and tools (tabular format: xlsx or csv)
	C.	This work package is mostly focused on gathering available information from existing sources – including scientific literature and research materials. We will identify, within this work package, the existing requirements from the SPATIAL industrial use cases and combine those with requirements that can be deduced from scientific literatures, articles and standards.
	D.	Own collections, desktop research and surveys conducted within the SPATIAL project use cases, other sources, literature. Online resources (article and research data databases).
	E.	<1GB
	F.	Research data resulting from this research will be of interest to other researchers, policy makers, practitioners in the sectors and the developers of the SPATIAL innovations.
WP2	A.	The purpose of the data collection is to validate the developed resilience and accountability features based on the existing intelligent cybersecurity solutions. The related SPATIAL objectives are:

SPATIAL project is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101021808.

Resilient accountability metrics and embedded algorithmic accountability		objectives 1: To develop systematic verification and validation software/hardware mechanisms that ensure AI transparency and explainability in security solution development; objective 2: To develop system solutions, platforms, and standards that enhance resilience and preserve privacy during the training and deployment of AI in decentralized, uncontrolled environments; objective 3: To define effective and practical organizational adoption and adaptation guidelines to ensure streamlined implementation of trustworthy AI solutions.
	B.	Machine-generated time series data: the real-time and historical performance of computation, storage, and communications of the whole monitored intelligent cybersecurity systems. Publicly available data such as MNIST and CIFA-10 that is widely accessible and acceptable in the AI research domain.
	C.	Existing data will be re-used to analyse the effectiveness of developed XAI solutions and the effectiveness when integrated with existing intelligent cybersecurity solutions with recorded data – currently, we cannot determine which data.
	D.	Automatically generated from the devices on the testbed. A publicly available dataset that is already widely used in the AI research community.
	E.	Maximally at a 10GB scale
	F.	Research data resulting from this research will be of interest to other researchers, policy makers, practitioners in the sectors and the developers of the SPATIAL innovations.
WP3 System architecture, consistency and accountability for AI, Validation and Testing	A.	Existing datasets would be considered for initial feasibility experiments of the project. Licences and proprietary rights will be respected and acknowledged for its usage. Afterwards the work will generate its own data, which will include data from systems and applications, e.g., sensor, audio, performance; and potential participants using the system. Measurements are collected through a combination of controlled laboratory experiments and field trials in everyday environments. In addition, in accordance with regulations on protection of private data, the persons participating in the field experiments will be asked for consent before data collection and the data will be treated according to the current regulations. The developed algorithms and models will be designed to consider privacy of individuals, and they will operate on anonymous data that is abstracted to an aggregate level. All studies will be submitted to review and guidance from relevant institutional review boards of the consortium members.
	B.	System and experimental data will be located into a control version system (CVS) in order to explore its evolution during the project. CVS will allow any researcher analysing the data to study it at different points in time. In this manner, if the data gets corrupted at any time, any researcher can go back to a previous valid data point. In addition, our replication system that collects the data will ensure having multiple unmodified copies.

SPATIAL project is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101021808.

	C.	Modification statistics will be also applied over the data on time. We also plan to apply MD5 algorithm over the data periodically, in order to prevent undesirable direct and indirect modifications. Image data, video footage and binary are also handled tracked though MD5 signatures and stored using their own proprietary formats.
	D.	Data generated within SPATIAL will be documented along with the data files, and with the source code that generated it, where possible. Wherever applicable, peer-reviewed publications will complement data documentation.
	E.	Any experimental data and any source code will be stored at the University of Tartu until publication, and thereafter freely available to the research community upon request and agreement of all members in the consortium. When no IPR or confidentiality issues are involved, the data will be submitted to public data repositories or international storage services such as EUDAT Data Infrastructure or CRAWDAD
	F.	Research data resulting from this research will be of interest to other researchers, policy makers, practitioners in the sectors and the developers of the SPATIAL innovations. Datasets released within the project will be anonymized before release to ensure no information that can be used to identify the person or sensitive information is released. Additionally, licenses for datasets will be chosen to reduce the risks of re-identification.
WP4 User Engagement, Acceptance and Practice Transitions	A.	Data will be collected to analyse communication patterns, practices and experiencesto identify the socio-technical contexts that can impact algorithmic development and adoption of SPATIAL outputs.
	B.	Audio/video recordings of interviews, meetings, communication texts (e.g. Discussions on EMDesk). Observation notes and field diaries. Possible formats (.docx; .wav; .mp3; .mov; .mpeg
	C.	n/a
	D.	Data will be collected in use case demonstrations and SPATIAL partners, through online and offline channels.
	E.	5 to 10 GB (transcriptions of interviews and meetings, video/audio recordings)
	F.	Data is useful to identify patterns and practices that would otherwise be missed without any form of written, audio or video record. Only this sort of data collection allows for the level of in-depth analysis required to reach WP4 goals.
WP5 Deployment and Demonstration	A.	Data sets will be developed in order to execute pilots in realistic scenarios. The data will be used to test the solutions developed in the prior WPs and demonstrate their applicability and usefulness.
	B.	The pilots require data from IoT devices, sensors, mobile & edge network infrastructure, virtual organizational network in the cloud. They will take the form of .csv or binary files.

SPATIAL project is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101021808.

	C.	n/a
	D.	As this work package includes multiple use case demonstrations, the data will be collected in multiple different settings. The source of the data potentially includes 4G/5G/IoT testbed, edge infrastructure of a network operator, and cloud platforms.
	E.	The expected size could be very large (TB) depending on the ultimate use case scenarios.
	F.	Research data resulting from this research will be of interest to other researchers, policy makers, practitioners in the sectors and the developers of the SPATIAL innovations. While clear description of the data will be provided in general, in case the data includes sensitive information or involve confidential matters, a sample of emulated data will be provided instead of the real data.
WP6 Impact, Outreach and Collaboration	A.	Building a database of stakeholders for promotion and dissemination of project results purposes.
	B.	Excel file including personal information about stakeholders mainly name, surname, email, phone and address
	C.	n/a
	D.	Contact data are coming from connections established by the consortium, public presentations, meetings with stakeholders, lists on the internet, etc.
	E.	Several MBs
	F.	Dissemination, creation of mailing campaigns, etc.

2 SECTION: FAIR DATA

2.1 MAKING DATA FINDABLE, INCLUDING PROVISIONS FOR METADATA

In the context of this project, both Zenodo and 4TU.ResearchData may be used. Given the international reach of Zenodo, we will prefer that repository. Should it be insufficient for the project (dataset size), then we will fall back to 4TU.ResearchData.

Data underlying the research publications will be made openly available through the 4TU.ResearchData or Zenodo. Both Zenodo and 4TU.ResearchData for Research Data are

SPATIAL project is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101021808.

trusted and certified research data repository, with a Data Seal of Approval certification. Datasets will be accompanied by rich metadata (adhering to the DataCite metadata standard) to ensure that all datasets are **findable**. In addition, to further aid their discoverability, keywords describing the datasets will be added.

The 4TU.Center for Research Data, and Zenodo are also using schema.org metadata, meaning that all datasets are indexed in the Google Dataset Search. Datasets will be assigned Digital Object Identifiers (DOIs), to make them **citable** and persistently **available**.

4TU.ResearchData provides long archive of dataset with both open and restricted access. For data sets which should be archived, but cannot be made open will be stored in a closed repository (such as DANS-Easy), under the responsibility of the associated Workpackage leader.

2.2 MAKING DATA OPENLY ACCESSIBLE

The SPATIAL Open Access ZENODO repository is available at this link.

We will follow the "as open as possible, as closed as necessary" principle, and given the commercial nature and personal aspect of the data processed, specific attention will be given to open data sharing throughout the project.

All data that can be made publicly available will be made so, under an open license (such CC-BY).

Data will be shared in openly accessible format, when possible (text files, and CSV format), easing future re-use.

Confidential data used in the project will not be shared openly. We will provide example datasets, or sample of the data, openly, to facilitate the validation and ease reproducibility of the work.

2.3 INTEROPERABILITY

We will identify in Work Package 1 standards that can be used to format the research data, as to follow industry or community standards.

However, we are not currently aware of such standards.

All data sets will come with documentation, describing how the information is organized and the vocabulary used in the documents, in the form of README files.

We will work with open data format as much as possible (CSV/Text files), and when not possible, we will provide in the documentation information regarding the necessary tools required to manipulate the data.

SPATIAL project is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101021808.

All datasets will be published in a research data repository (4TU.ResearchData), and will thus follow the DublinCore and DataCite metadata standard, easing findability of the datasets.

2.4 INCREASE DATA RE-USE (THROUGH CLARIFYING LICENCES)

Datasets will be licensed under a CC-BY license which requires attribution/credit for the original creation, while at the same time ensures broadest possible re-use. Datasets will be made publicly available as much as possible at the time of the publication of corresponding research papers resulting from this study.

The 4TU.Center for Research Data (where the datasets will be deposited) ensures data quality and curation (manual curation at the time of deposition, and automated curation and checks for data integrity after the deposit). Research data will be available for at least 10 years from the time of data deposition. Zenodo does not offer such quality assurance. Should we upload data in Zenodo, we will ensure that all datasets are properly documented.

3 SECTION: ALLOCATION OF RESOURCES

The 4TU.Center for Research Data is able to archive 1TB of data per researcher per year free of charge for all TU Delft researchers. In practice, this is applicable as long as one of the authors is from the TUD. For non-TUD researchers the costs-free limit is 10 GB per year. We do not expect to exceed this limit and therefore we expect no additional costs for long-term preservation. Zenodo has a limit of 50GB per datasets, and no limits to the number of uploads. Both Zenodo and 4TU.ResearchData are readily available.

Regarding personnel resources:

- WP leaders are responsible for adhering to the Data Management Plan specifications for their respective work package;
- For the overall project, TUD will be responsible for complying with the Data Management Plan;
- All consortium partners are responsible for making sure personnel working on the project have read the Data Management Plan and adopted the principles.
- All consortium partners have access to the EMDESK tool used to manage this project – and which contains the latest version of this document.

4 SECTION: DATA SECURITY

This section addresses data recovery as well as and how secure storage and transfer of sensitive data will take place. During the course of the project, data will be stored on servers maintained and automatically backed up by the ICT of each partner institution. Only team members have access to the designated server. The storage security is ensured by the ICT of each partner's institution.

SPATIAL project is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101021808.

The TU Delft support teams (data steward, privacy and ICT) will provide additional advice, if needed, on data storage during the project.

SURFDrive will be used for sharing data among collaborators. SURFDrive is a cloud service for Dutch education and research institutes. It complies with all Dutch and European privacy legislation. The data is stored safely in the Netherlands and is never made available to third parties. This data storage solution offers secure storage and transfer.

Additional information about access to the SURFDrive, ZENODO and EMDESK accounts can be found in the D7.3 Project Management Plan.

5 SECTION: ETHICAL ASPECTS

During this project, human participants will be interviewed and as such SPATIAL will involve personal research data. Details of all the steps undertaken to ensure appropriate levels of data protection are outlined in the dedicated ethics section (Section 34) of the GA and ethics deliverables (D8.1 H - Requirement No. 1, D8.3 POPD - Requirement No. 3). The ethics deliverables focus particularly on:

- The procedures and criteria that will be used to identify/recruit research participants;
- Templates of the informed consent forms and information sheets (in language and terms intelligible to the participants);
- Accordance of the data collection and processing with the 'data minimization' principle;
- Description of the technical and organizational measures will be implemented to safeguard the rights and freedoms of the data subjects/research participants;
- Description of the security measures that will be implemented to prevent unauthorized access to personal data or the equipment used for processing;
- Anonymization/pseudonymisation techniques that will be implemented;
- Information on the informed consent procedures with regards to data processing.

Furthermore, we comply with applicable EU data protection regulations and directives (see for details: https://ec.europa.eu/info/law/law-topic/data-protection/eu-data-protection-rules_en#abouttheregulationanddataprotection).

6 SECTION: CONCLUSION

This deliverable describes how SPATIAL data will be used, collected, and processed, taking as starting principles the applicable EU guidelines. Throughout SPATIAL more specific insights may gained regarding important aspects of the data involved (see Table 1). As such, this deliverable may be updated in due time in consultation with the management committee.

SPATIAL project is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101021808.