

Real and synthetic scenarios generated for the development, training, virtual testing and validation of CCAM systems



D1.3 Mid-term Data Management Plan

Document Type DMP
Document Number D1.3
Primary Author(s) Raffael Hipp | VUFO
Document Version / Status 1.1 | Final

Distribution Level PU (public)

Project Acronym SYNERGIES
Project Website www.synergies-ccam.eu
Project Coordinator IDIADA AUTOMOTIVE TECHNOLOGY SA
Grant Agreement Number 101146542
Date of latest version of Annex I against which the assessment will be made 2024-05-27



CONTRIBUTORS

Name	Organization	Name	Organization
Raffael Hipp	VUFO	Fadi Alakkad	UoW
Peter Miklis	VUFO		
Dorleta Garcia Melero	VICOM		
Sven Celin	AVL		
Katie Bruce	UoW		

FORMAL REVIEWERS

Name	Organization	Date
Fredrik Warg	RISE	2024-11-12
Thanh Bui	RISE	2024-11-12
Jordi Pont	IDI	2024-11-26
<i>David Storer</i>	<i>CLEPA</i>	<i>2025-11-21</i>
<i>Jordi Pont</i>	<i>IDI</i>	<i>2025-11-27</i>

DOCUMENT HISTORY

Revision	Date	Author Organization	Description
0.1	2024-10-01	Raffael Hipp - VUFO	First version of the document (Initial DMP, D1.2)
0.2	2024-11-22	Raffael Hipp - VUFO	Final draft version of the document
1.0	2024-11-26	Jordi Pont - IDI	Final version with feedback from reviewers integrated. Version to be submitted to EC.
1.1	2025-09-29	Raffael Hipp - VUFO	First version of the document (Mid-term DMP, D1.3)

TABLE OF CONTENTS

1	EXECUTIVE SUMMARY	7
2	INTRODUCTION	8
	2.1 Structure of the deliverable	9
	2.2 Partner consultations	10
3	SYNERGIES - PROJECT SUMMARY	11
	3.1 Introduction to the project	11
	3.2 Objectives	11
	3.3 Work packages and deliverables	12
	3.4 Project partners	13
4	ROLES AND RESPONSIBILITIES	16
5	DATA SCOPE AND DESCRIPTION	18
	5.1 Data sources and formats	18
	5.1.1 Data sources	18
	5.1.2 Data formats	19
	5.1.3 Methodologies and technologies	20
	5.2 Data size (expected)	20
6	DATA HANDLING	21
	6.1 Data labelling	21
	6.2 Versioning	22
	6.3 Standards	23
	6.3.1 Data protection standards	23
	6.3.2 Data formats and quality requirements	23
	6.3.3 Tools for data provision, storage and visualisation.	23
	6.4 Software and Tools	24
	6.4.1 Software for development and analysis	24
	6.4.2 Data handling and infrastructure	24
	6.4.3 Scenarios and Databases	25
7	DATA STORAGE AND DATA ARCHIVING	26
	7.1 Storage Strategy: A Federated Model	26
	7.2 Archiving and Retention Policy	26

8	DATA EXCHANGE AND DATA SHARING	28
8.1	Data flow	28
8.2	Data ownership	29
8.3	Data classification	29
8.4	Interoperability	31
9	DATA ACCESS AND RESTRICTIONS	32
9.1	Protection principles	32
9.2	FAIR data principles	33
9.3	Compliance with data protection laws	34
9.4	Ethical considerations	34
10	DATA SECURITY	36
11	DATA PUBLISHING AND LICENSING	37
11.1	Publishing	37
11.2	Licensing	38
12	COSTS	39
13	CONCLUSIONS	40
14	REFERENCES	42
A.	QUESTIONNAIRE	43
B.	ABBREVIATIONS AND DEFINITIONS	47

LIST OF FIGURES

Figure 1: SYNERGIES WPs and their main interactions.....	12
Figure 2: Countries of the SYNERGIES Project Partners.....	14

LIST OF TABLES

Table 1: SYNERGIES Project partners..... 14

Table 2: Data table to be completed for each dataset (data set information) 21

1 EXECUTIVE SUMMARY

The document provides guidelines and best practices for data management within the SYNERGIES project to ensure its successful implementation. It is based on the Consortium Agreement (CA) and the Grant Agreement (GA). The CA contains the data management provisions agreed between the project partners. The GA contains the data management provisions agreed with the European Commission. These agreements are supplemented in the Data Management Plan (DMP) with project-level guidelines for processes and procedures to ensure effective data management.

Responsibility for the implementation of the DMP lies with all project partners according to their roles and responsibilities. The Project Coordinator, Working Group Leaders and Task Leaders will ensure compliance with the DMP under the guidance and supervision of the Data Protection Officer.

Please note that the DMP in no way replaces the legal documents of the project and their application. In case of doubt, the legally binding documents take precedence.

The objective of this document is to outline the general data management plan for the project and the correct and appropriate handling of personal data that needs to be protected or that requires special treatment from an ethical point of view. When collecting, processing or using personal data by a project member, the General Data Protection Regulation (GDPR) and other legal requirements must first be taken into account. These requirements are explained in the DMP.

In addition, the FAIR (Findable, Accessible, Interoperable, Reusable) principles are highlighted in the DMP as an important guideline to ensure findability, accessibility, interoperability and reproducibility of data. The principles are summarized in the DMP, together with an explanation of the commitment to follow these principles. The SYNERGIES project intends to follow the FAIR principles wherever possible, taking into account commercial and practical constraints.

The way in which the data is managed and disseminated is critical to the success of the project. The project unites a variety of partners with different internal processes and objectives from different sectors such as business, technology and research. It is essential to establish and maintain a common understanding of how data will be handled, the conditions under which data will be published and the procedures for agreeing these conditions. In this way, a culture of trust can be built that improves collaboration between partners and, ultimately, project outcomes. The DMP provides guidelines and recommendations for good practice that can be considered by partners in their daily interactions and in the appropriate application of due diligence.

As soon as the project partners recognize the relevance of the data and results for the public, they will evaluate the possibilities of publication in accordance with the guidelines of this DMP and the requirements of the CA and GA. The DMP contains best practices and procedures for the provision of data within the project and for the archiving and availability of these data beyond the project period.

This document provides initial guidance that explains project implementation in a simple and practical way, supports compliance with the underlying legal obligations and promotes the application of best data management practices.

As the project progresses, the DMP will be continuously updated to enable efficient integration, creation and management of FAIR data.

2 INTRODUCTION

The Data Management Plan (DMP) is an essential part of the work package (WP) WP1 "Project Coordination and Technical Management" in the SYNERGIES project. WP1 ensures that the project is carried out successfully. Professional management and efficient coordination ensure that the project objectives are achieved. WP1 is responsible for the entire technical and administrative project management and deals with all issues relating to data management and innovation management.

For successful project realisation, guidelines for the handling of research data must be established for the entire duration of the project and beyond. These guidelines are developed and recorded in the data management plan. The DMP serves as an orientation framework for how the project data is collected, organised, stored and used efficiently by defining the methodologies and standards to be adhered to, including determining the scope of data collection, processing and generation. This document also specifies whether and how the data will be shared or made openly accessible, taking into account personal data protection. In addition, plans are also drawn up for data handling, data curation and long-term preservation of the data (after the project).

All procedures and guidelines for the entire data cycle (Data Generation, Data Storage, Data Processing, Data Sharing, Data Retention and Archiving, Data Deletion or Disposal) within a project are described and provisionally defined in the DMP (Deliverable D1.2). This information is updated regularly in the mid-term DMP (Deliverable D1.3) and the final DMP (Deliverable 1.4).

The aim of this DMP is to ensure the correct handling of project data in order to comply with both the legal framework and regulations (e.g. protection of personal data) and the practices of Open Science. In the context of the DMP, this refers to open data, which is explained in greater detail in chapter 9.1. The open science practices will be conducted in conjunction with Task 10.2 (Dissemination).

Importantly, the DMP has several functions including:

- **Creating Consistency:** data may be handled at all stages by multiple users and in multiple systems. To ensure that data is easy to handle and fit for purpose, a DMP should establish rules which create consistency across all project data.
- **Ensuring Data Quality:** By defining data collection methods, validation techniques and quality assurance processes, a data management plan helps to ensure the accuracy, consistency and reliability of project data.
- **Mitigating Risks:** A data management plan identifies potential risks such as data loss, security breaches or non-compliance with regulations. It establishes safeguards to mitigate these risks and ensure data integrity and data privacy.
- **Facilitating Reproducibility:** Transparent and well-documented data management practices allow project results to be reproduced, validated and utilized by other members of the scientific community, thereby promoting scientific progress and innovation.

2.1 Structure of the deliverable

This report is divided into the following chapters:

The introduction explains what a Data Management Plan (DMP) is, how it is structured and what it is used for.

Chapter 3 presents the SYNERGIES project and the objectives to be achieved during the project. The work packages and project partners are also introduced here.

Chapter 4 describes the most important positions and responsibilities of the project in relation to the data.

Chapter 5 describes the data used in the project. This includes where the data comes from, the format in which it is stored, the size of the data and how it can be classified.

Chapter 6 delves into data handling, detailing the naming conventions for data and the organization of the folder structure. It also outlines the procedures for version control. Additionally, this chapter lists the methods and software programmes necessary for data generation, presentation and processing.

Chapter 7 deals with data storage, both during and after the project. It details the locations, methods and amount of storage space needed throughout the project's duration. Additionally, it specifies the storage requirements for data archiving post-project, including the necessary storage space and the individuals responsible for managing this process.

Chapter 8 deals with data exchange both within the organisation and with external partners. This includes the definition of the data to be exchanged, the partners involved, the time and manner of data transfer. Compliance with the FAIR guidelines must be ensured, with the criteria of findability, accessibility, interoperability and reusability of the data as well as public access to scientific data taking centre stage. This chapter also discusses which metadata standard is used in this project. In addition, the methods and software components required for data access and utilisation are described.

Chapter 9 explains the data restrictions. First, the basic principles for working with data in the context of open data are described. The principles of lawfulness, fairness, transparency, data minimisation, accuracy, storage limitation, etc. apply. In addition, the provisions of the General Data Protection Regulation (GDPR), which contains the guidelines for personal data, must be considered. In addition, there may be restrictions due to ethical aspects, which are discussed here.

Chapter 10 focuses on data security, detailing methods for secure data storage. Key aspects include data recovery, data backup and the transfer of sensitive data. The chapter also covers crucial elements such as confidentiality, communication security, data integrity and availability. Additionally, it outlines the procedures for data backup, specifying the number of independent storage locations required, the frequency of backups and the methods for testing the backup system.

Chapter 11 describes how and when the data and results are published, and which licences are used for this.

Chapter 12 addresses the costs for data management, while Chapter 13 summarises the DMP.

2.2 Partner consultations

In order to gain a better understanding of the data used and required in the project, a questionnaire was sent to the SYNERGIES members to enquire about their data needs, and their data processing and management requirements (see attached questionnaire in the appendix A). This questionnaire will be sent to the SYNERGIES members at regular intervals to document evolutions.

As of this mid-term Data Management Plan (Deliverable D1.3), two rounds of questionnaire-based data collection have been completed. The first round of the project was conducted from project month M6 to M8, and the second from project month M13 to M16. The objective of the exercise was to document project-related developments and changes over time. The information obtained from the completed questionnaires is being systematically integrated into the respective chapters of the project documentation. Out of the 32 project partners listed under Chapter 3.4 "Project Partners", 26 actively participated in the partner consultations, providing valuable input.

3 SYNERGIES - PROJECT SUMMARY

3.1 Introduction to the project

The SYNERGIES project (Real and synthetic scenarios generated for the development, training, virtual testing and validation of CCAM (cooperative, connected and automated mobility) systems) is a European project funded by the EU research and innovation programme "Horizon Europe". SYNERGIES is represented by a consortium of 32 partners from 11 European countries, including leading vehicle manufacturers, suppliers, test laboratories and researchers.

The main objective of the SYNERGIES project is to develop a European scenario platform. This platform should fully enable the development, training, virtual testing and scalable scenario-based validation of CCAM systems. To this end, the Safety Assurance Framework developed in SUNRISE will be further developed in the following points:

- Expansion of the range of relevant scenarios through new scenarios for rural and urban environments that enable the descriptions of complex situations.
- Development of (semi)automatic AI (Artificial Intelligence) data processing tools for the extraction of scenarios.
- Provision of uniform access to a European Scenario Dataspace.
- Establishment of a marketplace to facilitate access to the tools for creating future scenarios.

In order to achieve this goal, five objectives were defined, which are described in more detail below.

3.2 Objectives

- **Objective 1: Deliver widely accessible scenarios.** Provision of scenarios for all stakeholders in the value chain through a scenario data space. Scaling up the SUNRISE framework by merging existing and newly developed scenario databases. The approach of this data space is to enable the provision of scenarios from day one by utilising existing initiatives and extending them with new initiatives. The aim is to build a dynamic EU-wide database that ensures efficient and up-to-date provision of relevant test scenarios.
- **Objective 2: Maximise the usability and coverage of the scenarios.** Maximise the usability and coverage of the scenarios to ensure the implementation of scenario-based validation. To achieve this, the requirements defined at the start of the project will consider a user-centred approach so that the SYNERGIES platform is specifically designed to satisfy the user needs of most relevant stakeholders. Additionally, SYNERGIES will enable the provision of scenario metadata that will allow a better understanding of the source, representativeness and limitations of the scenarios provided. AI tools will be developed and integrated into the toolchain to automatically obtain scenarios. The aim is to improve validation of CCAM systems through real and synthetic test scenarios covering as many traffic situations as possible that CCAM systems may face on European roads.

- Objective 3: Enable the use of heterogeneous and inclusive data sources for the generation of scenarios.** This is to be achieved by evaluating and utilising different data sources. The following data sources can be considered: accident data, vehicle data, infrastructure data, drone data. SYNERGIES will identify the most relevant data to be collected to obtain the scenarios required for the ODD (operational domain design) extension. The SYNERGIES platform includes a marketplace that provides all the tools needed to extract scenarios from raw data (and intermediate steps) to maximise the use of multiple sources. The objective is to employ the most suitable methods for capturing relevant traffic data, which will serve as a foundation for developing test scenarios across various traffic environments in line with the extended ODDs.
- Objective 4: Achieve acceptance and upscaling of the proposed solution** through a transparent, holistic approach that is interoperable with the main previous EU projects (HEADSTART, SUNRISE, L3Pilot, Hi-Drive) and existing scenario database initiatives. The SYNERGIES platform will build on the outcome of the SUNRISE project, which relies on an international network of experts. In addition, the SYNERGIES platform will start with access to a large amount of data recorded on European roads. To maximise the scaling potential of the SYNERGIES platform, the project will develop and provide the necessary tools (via the marketplace) to integrate new initiatives and expand the platform.
- Objective 5: Enable a solid AI foundation for CCAM** that includes the generation and enrichment of data (real and simulated) and scenarios to achieve improved data diversity, richness and completeness. Similarly, AI methods will complement the existing applications for identifying, extracting and generating scenarios and meet the requirements for ensuring safety and security. In addition, SYNERGIES will provide specific data sets for AI training and development. The dynamic scenario database will be used for AI development and training.

3.3 Work packages and deliverables

In order to successfully complete the SYNERGIES project and achieve the above-mentioned objectives, the activities are grouped into ten work packages (WPs). The following section explains which tasks are dealt with in the individual WPs and Figure 1 shows the links between them.

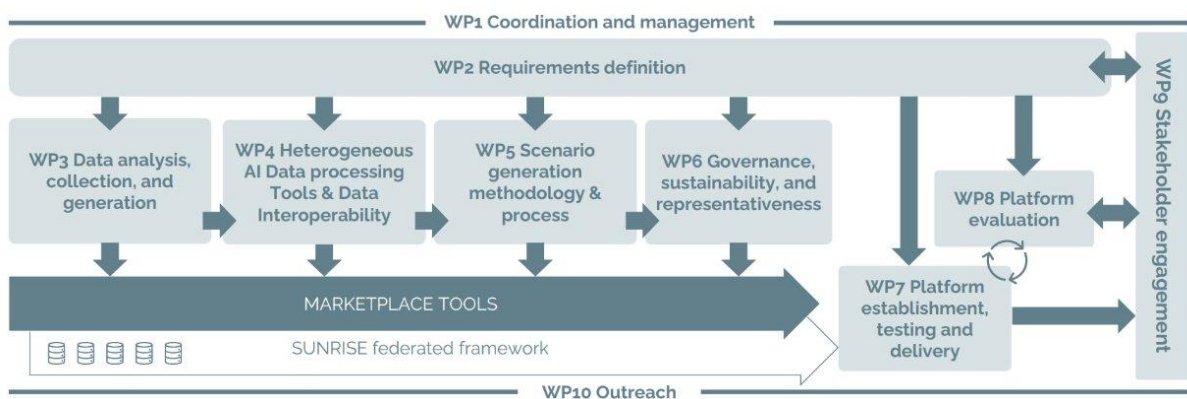


Figure 1: SYNERGIES WPs and their main interactions

The following is a brief overview and description of the individual WPs:

- **WP1 – Coordination and management:** Ensures the successful execution of the project and its objectives.
- **WP2 – Requirements definition:** Defines user-centric requirements by compiling and analysing the needs of the stakeholders.
- **WP3 – Data analysis, collection, and generation:** Defines best practices for data collection, identifies suitable existing datasets and collects new data.
- **WP4 – Heterogeneous AI data processing tools and data interoperability:** Creates trustworthy AI data processing tools that span from raw data to metadata, ready for deployment in the SYNERGIES platform.
- **WP5 – Scenario generation methodology and process:** Develops the overall methodology to identify, extract, enrich, and generate scenarios.
- **WP6 – Governance sustainability and representativeness:** Defines how to govern the scenario platform with the continuous updates of data.
- **WP7 – SYNERGIES Platform establishment, testing and delivery:** Integrates, tests, and delivers the developed SYNERGIES platform and ensures its functionality, reliability, and usability for end users.
- **WP8 – Platform evaluation:** Develops an evaluation framework that will enable the assessment of the platform's performance, the fulfilment of user requirements, and the collection of feedback for continuous improvement.
- **WP9 – Stakeholder engagement:** Ensures the involvement stakeholder involvement for project result acceptance and uptake.
- **WP10 – Outreach:** Maximises the impact of the project through effective dissemination of the project results and communication with external entities.

3.4 Project partners

The SYNERGIES consortium comprises 32 project partners from 11 European countries (Figure 2). The participating project partners include leading vehicle manufacturers, suppliers, test laboratories and research institutes.

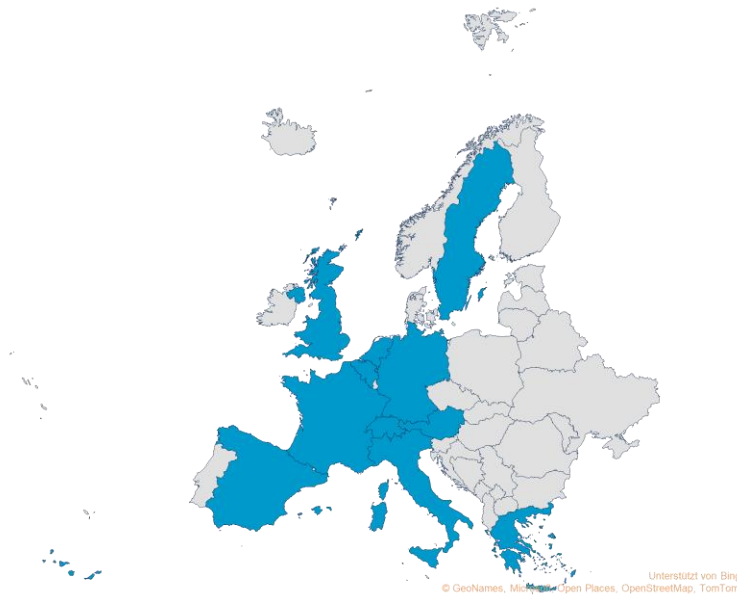


Figure 2: Countries of the SYNERGIES Project Partners

The participating project partners are listed below (Table 1).

Table 1: SYNERGIES Project partners

No.	Participant organisation name	Abbreviation	Country
1	IDIADA Automotive Technology, S.A.	IDI	ES
2	Bayerische Motoren Werke Aktiengesellschaft	BMW	DE
3	PSA Automobiles SA	PSA	FR
4	Renault SAS	REN	FR
5	Toyota Motor Europe NV	TME	BE
6	AVL List GmbH	AVL	AT
7	Centre Europeen d'études de securité et d'analyse des risques	CEESAR	FR
8	Association Européenne Des Fournisseurs Automobiles	CLEPA	BE
9	Fundacion para la promocion de la innovacion investigacion y desarrollo tecnologico en la industria de automocion de Galicia	CTAG	ES
10	Deutsches Zentrum fuer Luft - und Raumfahrt EV	DLR	DE

11	European road transport telematics implementation coordination organisation - intelligent transport systems & services Europe	ERT	BE
12	Faurecia Clarion electronics Europe	FCE	FR
13	Erevnitiko panepistimiako institouto systimatou epikoinonion kai ypologiston	ICCS	GR
14	Rheinisch-Westfaelische Technische Hochschule Aachen	RWTH	DE
15	Institut de recherche technologique System X	IRTSX	FR
16	Ivex NV	IVEX	BE
17	Mosaic factor SL	MOSAIC	ES
18	Partners for automated vehicle education Europe	PAVE	BE
19	RISE Research Institutes of Sweden AB	RISE	SE
20	SIEMENS industry software Netherlands BV	SIE-NL	NL
21	SIEMENS industry software NV	SIE-BE	BE
22	Nederlandse organisatie voor toegepast natuurwetenschappelijk onderzoek	TNO	NL
23	Technische Universiteit Eindhoven	TUE	NL
24	Universita degli studi di Genova	UNIGE	IT
25	The University of Warwick	UOW	UK
26	Fundacion centro de tecnologias de interaccion visual y comunicaciones	VICOM	ES
27	Virtual Vehicle Research GmbH	VIF	AT
28	Verkehrsunfallforschung an der TU Dresden GmbH	VUFO	DE
29	ZF Friedrichshafen AG	ZF	DE
30	* Torc Europe GmbH	TORC	DE
31	* Schweizerische Ruckversicherungs-Gesellschaft AG	SWRE	CH
32	* FEV IO GmbH	FEVIO	DE

4 ROLES AND RESPONSIBILITIES

It is recommended that, throughout the project, the Consortium Agreement (CA) and Grant Agreement (GA) are consulted regularly with respect to data processing. Additionally, key personnel involved in the project can offer valuable advice. Details about these key individuals are provided below.

- **Liaison and Collaboration Manager (LCM)**

The Liaison and Collaboration Manager supports and coordinates the collaboration between the project partners and external stakeholders (such as researchers, industry leaders and regulators) to guide the progress of the project. Continuous communication with key stakeholders will ensure that outcomes meet industry and regulatory expectations and maximize the impact of the project. By organizing workshops and collaboration platforms, feedback can be gathered and integrated into the project's development.

The role of Liaison and Collaboration Manager has been assigned to Iain Macbeth at ERTICO.

- **Dissemination Manager (DM)**

The Dissemination Manager (DM) of the project will be responsible for the dissemination and communication of the project results in cooperation with the project coordinator and the whole consortium throughout the project duration. Systematic dissemination of knowledge, research results and innovations will maximize the reach and impact of SYNERGIES and keep stakeholders informed about the progress of the project. Dissemination will be achieved through the publication of newsletters, reports, brochures, social media content and presentations. The DM will also plan and organize conferences, workshops and webinars. The DM will ensure that all EU directives and ethical guidelines are respected.

The role of Dissemination Manager has been assigned to Michael Karner at VIF.

- **Innovation and Exploitation Manager (IEM)**

The Innovation and Exploitation Manager (IEM) is responsible for ensuring the sustainable use of the project outcomes both during and after the project's completion. The IEM's role includes assessing the exploitation potential of identified innovations and developing strategies for their commercialization or utilization. In collaboration with the Steering Committee, the IEM develops and coordinates the necessary actions for the earliest possible market launch, taking into account the project objectives and legal framework conditions.

The role of Innovation and Exploitation Manager has been assigned to Stephane Dreher at ERTICO.

- **Data Protection Officer (DPO)**

The Data Protection Officer (DPO) ensures compliance with EU data protection laws, overseeing data privacy practices, conducting privacy impact assessments and advising on data protection obligations.

The DPO's tasks include advising the consortium on compliance with data protection laws. Another task is to support the implementation of data protection guidelines and

practices. The DPO also provides management with guidelines on data protection risks. As an advisor, the DPO helps to understand and manage the data protection obligations without directly managing the data protection procedures.

The role of Data Protection Officer has been assigned to Dorleta Garcia at VICOM.

- **Intellectual property rights Manager (IPRM)**

The Intellectual Property Rights (IPR) Manager provides advisory and support services related to intellectual property within the consortium. This role includes assisting with the identification, management and protection of intellectual property rights arising from the consortium's activities. The IPR Manager advises on Intellectual Property (IP) strategy, facilitates the handling of IP issues among consortium members and ensures compliance with applicable laws and regulations, without directly managing the IP assets. The IPR Manager only acts if the parties request assistance.

The role of Intellectual property rights Manager has been assigned to Mariola Hauke at CLEPA.

5 DATA SCOPE AND DESCRIPTION

In SYNERGIES, within WP3, data from previous EU projects, accident databases and other initiatives will be analysed to check their quality and suitability for scenario creation. In addition, new data will be generated from a variety of sources including vehicles, mobile roadside units, drones, simulation tools and generative AI methods. These datasets will be used throughout the project to extract scenarios, which will be stored in the Scenario Platform for the development, training, virtual testing and validation of CCAM systems.

5.1 Data sources and formats

Below, initial information is provided for the different research outputs and produced data, reflecting the status. Different types of data and research outputs will be produced in SYNERGIES:

- Datasets collected from equipped vehicles e.g. radars, cameras, LiDARs (not yet defined).
- Datasets collected from mobile roadside units e.g. cameras, LiDARs (CSV-like data).
- Datasets collected from drones incl. camera data (CSV-like data).
- Datasets generated from advanced simulation tools and generative AI methods (CSV-like data).
- Scenario Platform, including relevant scenarios (not yet defined): openly available.
- Documentation of specifications and architecture (text, graphical): SYNERGIES public deliverables.
- Data from end-user questionnaires (text) –content and form carefully assessed and GDPR-compliant.

To effectively track changes and updates, it is highly advisable to create a comprehensive log of any updates or revisions made to the datasets. This log should meticulously document each change, including the date of the update, the nature of the revision, and the individual responsible for the modification. By maintaining such a detailed history of changes, we can ensure transparency and accountability throughout the data management process. This approach not only helps in keeping track of the evolution of the datasets but also facilitates easier troubleshooting and auditing, thereby enhancing the overall integrity and reliability of the data.

A detailed description of the expected input and output is provided in the following chapters (chapter 5.1.1, 5.1.2, 5.1.3).

5.1.1 Data sources

A variety of data sources are utilized—partially—to support the generation of new datasets. These sources encompass both pre-existing information from earlier projects as well as a combination of publicly available and non-publicly accessible data. Publicly available data sources include a broad range of materials and inputs, such as:

- Police/Crash Reports
- Inputs from Crowdsourcing Platforms
- Open Data Sources
- Regulations (e.g. NCAP)
- Driving Tests
- Existing Scenarios

On the other hand, non-publicly available data sources include more sensitive or proprietary information. These primarily consist of:

- Internal datasets provided by Renault
- (In-Depth) Accident and Crash data

Despite the broad range of external sources, it is important to note that the majority of the data used in this project is generated rather than merely collected. This generated data includes:

- Synthetic Scenarios
- Collection of In-vehicle Data
- Video Observations (e.g. from drones or infrastructure)

5.1.2 Data formats

The following types of data and corresponding file formats are anticipated and considered essential:

- Internal/Proprietary Formats (e.g. YAML, S-CDF/OMEGA-PRIME)
- ASAM OpenX
 - Open Simulation Interface (OSI)
 - OpenSCENARIO (XOSC, XML)
 - OpenDRIVE (XODR, XML)
 - OpenLABEL (JSON)
- Vehicle Data (e.g. RTMaps)
- Description of infrastructure (e.g. HD map data, Lanelet2)
- Object-Level Data (e.g. OMEGA, HDF5, Hi-Drive formats, ROS)
- Weather Data (e.g. CSV, XML, JSON)
- Traffic/Interaction/Behavioural Data (e.g. XLSX, CSV, JSON, PCM)
- Data Evaluation Scripts (e.g. PY, R)
- Virtual Environment (e.g. UASSET)
- Raw Text (e.g. TXT, PDF)
- Video Data (e.g. H264, H265, AV1, AVI)
- Sensor Data (e.g. RTCM, CAN, MCAP, ROS, ROS2)

5.1.3 Methodologies and technologies

To make optimal use of the collected and generated data, the following section provides a comprehensive summary of the methods, technologies and software applications that are commonly used for processing, analysing and interpreting the data:

- Internal toolchains and custom scripts
- AI algorithms
- HD Maps
- SLAM software for odometry (Exwayz and OXTS solutions are also considered)
- Blurit for image anonymization
- PC Crash for accident reconstruction
- ADScene for the validation of various automated driving systems and advanced driver-assistance systems
- CARLA for simulation
- WebLabel for labelling
- AIMATS – Analysis and Investigation Method for All Traffic Scenarios
- CARLABS for interacting with users
- Software provided by sensor manufacturers

5.2 Data size (expected)

All data available in text or graphical form have very low size requirements (Kilobytes KBs/ Megabytes MBs). Concerning the datasets from vehicles, infrastructure, drones, simulation tools, which will be required for the verification of results and for reuse, the estimated size is in the order of Terabytes (TBs).

6 DATA HANDLING

Data handling is an important aspect that presents various complex challenges. These challenges arise mainly from the diversity of data types, the integration of real and synthetic data, the coexistence with iterative AI development and operation cycles and the associated storage and computational costs. To ensure trustworthy data processing, SYNERGIES is committed to the following principles:

- **Data Provenance Registry:** Every transaction is meticulously recorded to establish relationships that enable traceability of data transformations, conversions, transfers and more.
- **Data Management:** We emphasize data availability, integrity, security and usability for multiple end users and organisations and actively work to avoid data silos.
- **Semantic/Linked Data:** Data enrichment by linking to registered concepts in an ontology-based layer.
- **Standardization of Data Interfaces:** Wherever possible, international standards and de facto standards for data formats are used. This approach minimizes the need for conversion layers and maximizes interoperability.

If datasets are shared between project partners as part of the project, it is recommended that they be annotated according to the table (Table 2) below.

6.1 Data labelling

The data should be annotated to ensure full transparency and to increase the likelihood that the data is shared with the correct permissions. The annotation should look like this (Table 2):

Table 2: Data table to be completed for each dataset (data set information)

Data set name	NAME
Data set classification	Sensitive/Public
Data set references	SYNERGIES_XX Each data set will have a reference that will be generated by the combination of the name of the project and the reference number
Data set owner	Owner organisation(s) and contact person(s)
Organisations the data owner has provided authorisation for usage	The name(s) of organisations that the data has been shared with, and the intended usage of the data

Source	Data set source (specify reuse if applicable)
WP's	WPs in the SYNERGIES project.
Data set description	Each data set will have a full data description explaining the data provenance, origin and usefulness. Reference may be made to existing data that could be reused.
Data format	All the format that defines data.
Data size	Size of the data set.
Derived data	Data extracted from the created data set.
Data sharing	Explanation of the sharing policies related to the data set, including any associated terms and conditions (i.e., data generated outside of the project)
Reuse of existing data	If applicable
Archiving and preservation	The preservation guarantee and the data storage during and after the project (databases, institutional repositories, public repositories...)

The data should be annotated in a prominent position. For example, as with standard Microsoft or similar applications, this should be on the first or second page/slide. When sharing data, such as cloud storage, simulation platforms, etc., it is recommended that the annotations are communicated in an appropriate manner. Either in the system itself or by prior notification before access is granted.

6.2 Versioning

The versioning of the dataset, folder structure and file names should include the following information in the title: the creator or contributor, the creation date and the access conditions. This will make the data easier to find. Each structure should be organized as follows:

- DATE: YYYY-MM-DD (ISO8601)
- TIME: hh:mm:ss (ISO8601)
- TITLE: Data set title
- OWNER: Author using email address
- PARTNER: Partner short name
- WP-TASK: DoA Linked Task, e.g. WP3T2.2
- MAJOR: Major Version Number "XX."
- MINOR: Minor Version Number ".YY"
- DIST: DoA Distribution level

6.3 Standards

As part of the partner consultation via questionnaires the project partners' requirements and expectations with regard to data handling standards were systematically surveyed. The objective is to ensure that the data used and generated in the project is handled in a consistent and interoperable manner.

6.3.1 Data protection standards

The following data protection standards apply to all data processing activities within the project, and compliance with these standards is mandatory.

- The General Data Protection Regulation (GDPR)
- National legal requirements and specifications by responsible authorities

These standards are mandatory for all partners and form the basis for the lawful collection, processing and disclosure of personal or sensitive data. More detailed information can be found in Chapter 10 "DATA SECURITY".

6.3.2 Data formats and quality requirements

To ensure interoperability and efficient collaboration, the following file formats and quality criteria were specified as framework conditions:

- Use of internationally recognised standard formats, including those from the industrial sector
- A common data format for data exchange
- OpenLABEL for object list data and sensor descriptions
- ROSBags/RTMaps or other formats for raw data recordings

In addition, the following methodological requirements were formulated:

- Definition of common standards for
 - The recording of driver data as part of scenario identification
 - The exchange of scenarios (including a minimum parameter set)
 - Visualisation of data.
- Definition of quality criteria for driving data and the contents of the scenario database.
- A standardised classification of functional and logical scenarios that covers at least 70 % of the project's use cases.

6.3.3 Tools for data provision, storage and visualisation.

The following tools and platforms were recommended or identified as standard:

- Using a version control system such as Git facilitates collaborative management of data and source code.
- The use of browser-based tools for the purpose of querying data libraries and visualising statistical analyses.

6.4 Software and Tools

The questionnaire was also used to record the tools and software solutions used to manage the data generated and used in the project. The objective is to facilitate a shared understanding of the technical tools employed in the project and to support the utilisation of interoperable, future-proof solutions.

The results below provide an overview of the tools and software solutions that have been used or are to be used, grouped by area of application.

6.4.1 Software for development and analysis

A wide range of development environments and analysis tools are used to process, analyse and simulate project data.

- Programming languages and development environments include:
 - Python (including open-source machine learning libraries such as PyTorch and TensorFlow)
 - RStudio
 - MATLAB
 - Microsoft Excel
 - Unreal Engine
- Other tools:
 - In-house developments or internal software solutions
 - Open-source software for data processing (e.g. data visualisation or machine learning)

6.4.2 Data handling and infrastructure

Various tools are used or required to store data in a structured way, to version it and to make it traceable.

A key tool is the version control system Git, which is used for the following purposes in particular:

- Storage and management of data (e.g. XML and JSON) and source codes
- Traceable versioning and logging of changes to data records
- Collaborative processing and further development of data

Some project partners also use their own data management solutions or central platforms to organise and share data.

The SYNERGIES members also specified the following requirements for the tools used to manage project data:

- Versioning tools to ensure traceability and data provenance
- Cooperative platforms for joint data processing and exchange within the consortium.

Using this infrastructure ensures that project data is processed consistently, and is both traceable and accessible.

6.4.3 Scenarios and Databases

The following tools and platforms were identified for managing, querying and processing scenario data:

- Scenario databases and platforms:
 - Scenario.Centre: a proprietary platform for scenario management
 - ADScene: a database with an input and query tool for scenarios.
 - SafetyPool and other crowdsourcing platforms for collecting and providing traffic scenarios.
- Scenario processing and identification:
 - Use of API-based tools and algorithms for scenario identification and uploading them to the databases.
 - Tools for filtering and processing specific data formats:
 - Time series data (e.g. from GIDAS PCM)
 - XML files (e.g. OpenDRIVE, OpenSCENARIO)
 - JSON files (e.g. AIMATS).
 - Open-source library:
 - TASI: a Python library for analysing traffic data and interpreting traffic situations (<https://zenodo.org/record/14514547>).

The information collected will form the basis for consistent data handling throughout the project. It will also support the development of joint processes and the selection of interoperable tools and formats, ensuring the sustainable and quality-assured utilisation of project data.

7 DATA STORAGE AND DATA ARCHIVING

To make the platform even more functionally rich, SYNERGIES supports robust data archiving and storage solutions. Accordingly, the project DMP includes comprehensive data protection, metadata documentation, and facilitates data interoperability, according to the FAIR principles. This approach also enables the ability to efficiently retrieve and use the data while ensuring the long-term preservation and accessibility of the valued dataset.

As foreseen in the original DMP, consultations with the partners provided detailed information on storage and archiving solutions, done in end-2024 and mid-2025.

7.1 Storage Strategy: A Federated Model

The project SYNERGIES will work with different storage-solutions. There is no single central repository for all data; and therefore, storage will be provided by partner-controlled infrastructures and dedicated project platforms.

Primary storage locations and methods which the partners reported included:

- **On-premise:** Several partners have their own servers, including AVL, CEESAR, DLR, RISE, and Vicomtech. These are controlled, maintained, and backed up within the policies of the partner's IT department, usually on redundant storage, such as RAID.
- **Cloud Platforms:** Many partners use commercial cloud services for storage. The main ones used are Microsoft Azure and Amazon Web Services (AWS). Data stored in these environments is encrypted and protected via access controls managed by the partner. For structured data, data sovereignty is achieved using, among others, private clouds and database infrastructures like PostgreSQL. Collaboration happens daily using platforms like SharePoint and Nextcloud.
- **The SYNERGIES database** shall be developed as a central repository for the final scenarios and key project outputs.

Source code is stored within specific Git repositories - GitHub and GitLab.

The volume of data varies a great deal with type, confirming initial estimates. Consultations with partners make it clear that whereas metadata, reports, and sample data may only be Megabytes to Gigabytes, raw datasets are expected to be substantial, ranging from hundreds of Gigabytes to Terabytes.

7.2 Archiving and Retention Policy

Data archiving and retention policies are not centralized but are the responsibility of the data-owning partner, as per the arrangements under GA and CA. Decisions on what data to keep and for how long are driven by three key drivers:

1. **Legal Requirements and GDPR:** The Legal Departments and DPOs from the partners (such as AVL, CEESAR, Siemens, Vicomtech) determine the retention periods in view of the GDPR, considering legal requirements at the national level. Personal data is deleted after anonymization.

2. **Internal Partner Policies:** Each partner has its own archival policies. Several partners (such as DLR and RISE, for example) have a standard 10-year retention period for research data. Others cite a legal obligation of 5 years post-project.
3. **Project Relevance:** The project team (such as the Coordinator and WP Leaders) along with specific platform teams (for example, ADScene management team) determines what data is relevant to retain according to project goals.

Long-term preservation is ensured by storage on partners' persistent, backed-up servers (e.g., DLR, TNO), by publishing datasets to public repositories (such as Zenodo) and through the planned sustainability of project platforms (such as ADScene), the SYNERGIES platform. As a minimum, data will be available for the lifetime of the project; however, most partner strategies point to longer-term availability, 5-10 years or more.

To further enhance the platform's capabilities, SYNERGIES incorporates robust data archiving and storage solutions. The project's data management strategy includes comprehensive data protection measures, metadata documentation, and support for data interoperability, in compliance with the FAIR principles. This approach not only facilitates efficient data retrieval and usage but also ensures the long-term preservation and accessibility of valuable datasets.

As promised in the initial DMP, partner consultations (conducted end-2024 and mid-2025) have provided specific details on storage and archiving solutions.

8 DATA EXCHANGE AND DATA SHARING

The way in which data is exchanged and disseminated is of the most importance. The project brings together a wide range of partners with different internal processes and objectives, from commercial and technical companies to research organisations. It is essential that a common understanding of how data is shared, the conditions under which data is made public and the processes of agreement of such conditions is established and maintained. In doing so, a culture of trust can be established, which in turn will enhance the collaboration of the partners, and ultimately the project outcomes. Whilst the day-to-day interactions between partners and employing the appropriate level of due diligence is what builds such trust and fundamentally, data will be exchanged and shared in line with the requirements of the GA, CA and all applicable data protection legislation, this section will provide guidance and recommendations on best practice.

Key guiding principles for data exchange and protection include:

- All data exchanges must comply with the General Data Protection Regulation (GDPR) and all relevant national legal requirements.
- Data should be shared in a manner that promotes openness, transparency, credibility, and consistency among all partners.
- Data Processing Agreements (DPAs) must be signed and approved before any data is shared between partners, as required by the Grant Agreement (GA) and Consortium Agreement (CA).
- Any exchange of raw data will be evaluated on a case-by-case basis, and all exchanges must adhere to ISO 27001 [8] information-security standards.
- As a publicly funded project, SYNERGIES is committed to sharing as much data as possible within the constraints of GDPR and intellectual-property rights.
- As a publicly funded project, SYNERGIES shares maximum data within GDPR and IP constraints.
- All partners are expected to follow core information-security principles, including maintaining appropriate access rights, data availability, data correctness, and traceability.
- Camera data containing personal identifiers must be anonymised before sharing. Only non-identifiable data (e.g., LiDAR, Radar, GNSS/IMU) may be shared externally.
- Data may be exchanged primarily for research purposes, and public release will be considered on a case-by-case basis following legal and ethical review.

8.1 Data flow

It is expected that most of the data will either flow into or be generated within the technical and co-operation work packages. In all cases it is expected that all partners will handle data with 'the same degree of care with regard to the Confidential Information disclosed within the scope of the Project as with its own confidential and/or proprietary information, but in no case less than reasonable care' (CA, clause 10.5).

The main data flow is as follows:

1. Data collection (WP3)
2. Data processing and scenario detection (WP4)

3. Scenario generation (WP5)
4. Data analysis (WP6).

The data-flow process includes:

- Capturing data, annotating at the object level, identifying and extracting scenarios, generating scenario datasets, and integrating them into the SYNERGIES platform.
- Recording data locally, processing it to generate object lists and metadata, and extracting relevant scenarios for analysis.
- Translating all data into a Common Data Format (CDF) to guarantee interoperability between partners
- Applying a defined sequence for handling personal data:
 - Raw → Object → Anonymised → Scenario → Potential Sharing, ensuring full compliance with GDPR

8.2 Data ownership

The data originator is the organisation that originally created the data.

Unless otherwise and explicitly stated it should be assumed the data originator owns the data and has the authority and responsibility to define who that data can be shared with, for what purpose and its classification.

The data originator should follow the best practice guidelines in this DMP and within their organisation when sharing data, including clear classification and labelling.

- Each partner owns the data they collect or generate during the project.
- Reuse of data, especially for commercial purposes, requires prior consultation and approval by the data owner.
- Data owners define internal sharing procedures in coordination with their technical and legal departments, as required by the GA and CA.
- For certain databases, users must register before obtaining access; administrators manage access rights, and the Steering Committee must approve private container access.
- Data producers may share datasets with SYNERGIES partners for research purposes or release them publicly after proper review and approval.
- If the data is collected within the SYNERGIES project, then the data can be used by all SYNERGIES partners.

8.3 Data classification

The simplest and most effective way of ensuring data is handled with care is a simple classification system.

The SYNERGIES project has two levels of classification:

1. Sensitive or confidential
2. Public

Sensitive or confidential

Data classified as sensitive or confidential includes information that is restricted under the Grant Agreement (GA), Consortium Agreement (CA), GDPR or intellectual-property rights.

Key points regarding sensitive/confidential data:

- **Ownership and Sharing:** The data originator retains ownership and decides who can access the data, for what purpose, and under which conditions. Data should not be shared with partners or externally without approval from the originator
- **Labelling:** Data must be clearly labelled before sharing. It is recommended to follow the labelling guidelines from Chapter 6.1, Table 2, and to highlight sensitive/confidential data in bold red for visibility. Partners will comply with the agreed labelling strategy once finalised in the working groups.
- **Traceability:** The origin and handling of each dataset must be traceable. This includes:
 - o Unique data model for driving data
 - o Unique scenario classification
 - o Relevant parameters and tags
 - o Format-specific labelling (e.g., GIDAS PCM for time series, OpenDRIVE/OpenSCENARIO for XML, AIMATS for JSON, MCAP files)
- **Security:** Sensitive data must be managed according to ISO 27001 [8] standards, with proper access control, availability, correctness, and traceability.
- **Personal Data:** Data containing personal identifiers, such as camera footage, must be anonymised before sharing. Only non-identifiable data (LiDAR, Radar, GNSS/IMU) may be shared externally.

Public

Data classified as **public** refers to information that is intended for external dissemination and can be accessed freely, though some datasets may require registration or have terms of use.

Key points regarding public data:

- **Approval:** Any data intended for public release must be approved by the data originator to ensure GDPR compliance, IP clearance, and ethical considerations.
- **Anonymisation:** Public data must be free of personal identifiers. Camera footage must have blurred faces and license plates; LiDAR, Radar, and GNSS/IMU data may be shared after consultation. Scenario datasets may also be shared publicly following legal review.
- **Decision Process:** The Steering Committee, in consultation with relevant work packages, approves public release. Data exchange agreements specifying purpose, duration, and restrictions should be executed where applicable.
- **Documentation:** Public datasets will include metadata, readme files, and technical documentation to ensure interoperability, traceability, and reusability.
- **Standards and Interoperability:** Data will follow community-endorsed standards and practices where possible, such as ASAM OpenLabel v1.0.0, OpenSCENARIO, OpenDRIVE, AIMATS, and the SYNERGIES Common Data Format. FAIR principles will be applied to facilitate reuse.

8.4 Interoperability

Interoperability ensures that data collected, processed, and shared within the SYNERGIES project can be effectively reused, integrated, and understood by all partners, as well as potentially by third parties after the project. Partner feedback and best practices have informed the following guidelines:

Data Standards, Formats, and Methodologies

- **Common Data Format (CDF):** All project data will be translated into the SYNERGIES Common Data Format to guarantee consistency and interoperability across partners.
- **Community and Industry Standards:**
 - **ASAM OpenLabel v1.0.0:** for labelling and metadata of object-level data.
 - **OpenSCENARIO / OpenDRIVE:** for scenario and simulation description (XML format).
 - **AIMATS (JSON):** for data exchange and storage of simulation results.
 - **PCM (GIDAS):** for time-series data.
 - Partners will follow additional standards where applicable or develop project-specific formats when no standard exists.
- **Metadata and Ontologies:**
 - Metadata will be derived from scenario definitions, ODD (Operational Design Domain), and ontology structures as defined in [WP5](#).
 - Key information will include scenario type, origin, relevant parameters, and traceability tags.
- **Documentation:** Data will include README files, technical documentation, variable definitions, units of measurement, and references to related deliverables, ensuring data re-use and validation of analyses.

Best Practices for Interoperability

- **FAIR Principles:** Data will follow the FAIR principles (Findable, Accessible, Interoperable, Reusable) wherever feasible, enabling effective data sharing both internally and externally.
- **Traceability:** All datasets will have clear references to their origin, including partner, project task, and generation process, ensuring that third parties can understand context and reuse appropriately.
- **Data Validation:** Partners will implement validation processes, such as consistency and causality checks, spot checks, or scenario review workflows to ensure that data conforms to agreed standards.
- **Version Control and Repository Management:**
 - Data will be managed through controlled repositories (e.g., GitLab) with appropriate access rights.
 - Private and public containers will be clearly defined, with the Steering Committee approving any public access or external sharing.
 - Versioning ensures that any changes to the data or metadata are traceable.
- **Interoperability with Existing Data:** Where possible, datasets will include references to external or prior research datasets to enable comparative analysis and integration.

9 DATA ACCESS AND RESTRICTIONS

In the context of the SYNERGIES project, effective data management is crucial for the successful development, training, and validation of Cooperative, Connected, and Automated Mobility (CCAM) systems.

The unique challenges of CCAM such as its safety-critical nature, reliance on complex sensor fusion, and the involvement of personal mobility data make a robust data strategy paramount. The sheer volume, velocity, and variety of data from sensors like LiDAR, radar, and cameras present significant technical and ethical hurdles. This chapter outlines our approach to data access and restrictions, emphasizing a necessary balance between the principles of open data and the legitimate intellectual property, commercial sensitivities, and confidentiality requirements of the partners. This approach is informed by direct stakeholder feedback and aims to foster a transparent and collaborative environment while respecting the specific constraints of each data provider. For example, while anonymized traffic scenarios might be shared openly to benefit the wider research community, the underlying raw sensor data or proprietary algorithms developed from it may constitute significant intellectual property.

The SYNERGIES project aims to enhance the usability and impact of the research outcomes, the intention being to create a central point for data *discovery* (e.g., through a metadata catalogue), sharing, and access, alongside partner-specific repositories (e.g., private GitLab instances). This federated model allows partners to maintain control and data sovereignty over their valuable data assets which may be terabytes in size and subject to strict internal policies while still enabling consortium-wide collaboration under agreed-upon rules. This approach favours moving *computation* to the data, rather than moving all data to a central location.

9.1 Protection principles

The SYNERGIES project is grounded in the belief that open access to research data is essential for innovation and collaboration, wherever feasible. We adhere to key principles that govern our data management practices [\[1\]](#):

- **Lawfulness and Fairness:** All data collection and processing activities will comply with applicable laws and regulations, ensuring that data subjects are treated fairly. This includes, but is not limited to, the ePrivacy Directive and relevant national-level acts concerning telecommunications and vehicle data.
- **Transparency:** We commit to being transparent about how data is collected, used, and shared, providing clear information to stakeholders. For example, this involves providing clear, layered data notices to any test-vehicle participants or citizens in a data collection zone, explaining in simple terms what is being collected and why.
- **Purpose Limitation:** Data will only be collected for specified, legitimate purposes and not further processed in a manner incompatible with those purposes. Data collected for validating a perception algorithm, for instance, will not be repurposed for unrelated driver behaviours marketing without explicit, separate consent and technical separation.
- **Data Minimization:** We will collect only the data necessary for our objectives, reducing the risk of unnecessary exposure. This principle also links to efficiency

and cost; collecting, storing, and processing terabytes of unnecessary data is a significant financial, environmental, and technical burden.

- **Accuracy:** Efforts will be made to ensure that data is accurate and kept up to date, with mechanisms in place for regular review. In a safety-critical domain like CCAM, inaccurate training data (e.g., a mislabelled 'stop sign' or 'pedestrian') could have catastrophic consequences, making this principle a core safety and ethical requirement.
- **Storage Limitation:** Data will be retained only as long as necessary for the purposes for which it was collected, after which it will be securely deleted if required by law or specific agreements. Stakeholder feedback indicates a preference for retaining anonymized data for longer periods (e.g., 5-10 years) to maximize research value. This retention period allows for longitudinal studies, re-validation of models against older data, and compliance with potential regulatory or funding-body audit requirements.
- **Integrity and Confidentiality:** We will implement appropriate technical and organizational measures to protect data against unauthorized access, loss, or damage. This will involve clear access control mechanisms, with data access decisions resting with the respective data owner. Practical measures will include encryption at rest and in transit, strict role-based access control (RBAC) policies, and secure enclaves for data processing where necessary.
- **Accountability:** The project will maintain records of data processing activities and ensure compliance with these principles. This is evidenced through auditable records, Data Protection Impact Assessments (DPIAs) for high-risk processing, and a comprehensive, living Data Management Plan (DMP) that is updated as the project evolves.

9.2 FAIR data principles

To further enhance data accessibility and usability, the SYNERGIES project embraces the FAIR data principles [\[2\]](#):

- **Findable:** Data will be assigned unique identifiers and metadata to facilitate easy discovery. This metadata catalogue will be a key feature of the SYNERGIES platform, detailing data origin, sensor types, recording conditions (e.g., weather, location), format, and, critically, the access conditions and contact person.
- **Accessible:** We will ensure that data is available in a format that is easy to access and use, while respecting any necessary restrictions. Based on partner feedback, data may be shared via multiple channels, including the central SYNERGIES Platform, partner-managed repositories (e.g., Github/Gitlab), or dedicated research data infrastructures like Zenodo. Accessibility does not always mean 'open'; it means the procedure for access is clear, whether that involves downloading from a public repository, authenticating via a secure API, or completing a formal data-sharing request.
- **Interoperable:** Data will be structured in a way that allows it to be integrated with other datasets, promoting collaboration across projects. Given the diversity of sensor formats and proprietary software, the project will promote interoperability by agreeing on common data models, shared ontologies for scenario description (such as ASAM OpenSCENARIO), and providing conversion tools where possible.
- **Reusable:** Clear licensing and documentation will be provided to enable others to reuse the data effectively. Where possible, data will be provided in common,

open formats readable by standard software (e.g., text editors, Python, MS Excel) or domain-specific tools (e.g., ASAM OpenLabel, ROSBags, CARLA, Foxglove studio) to lower the barrier for reuse. This requires clear licensing that distinguishes between, for example, licenses for internal consortium R&D and broader public-release licenses (e.g., Creative Commons, ODbL). Documentation will also include 'datasheets for datasets,' explaining a dataset's origins, limitations, and appropriate use cases.

9.3 Compliance with data protection laws

The SYNERGIES project is committed to complying with the General Data Protection Regulation (GDPR) and other relevant data protection laws, such as the UK data protection act or the Swiss federal act on data protection. This includes:

- **Personal Data Handling:** Any personal data collected will be processed lawfully, with explicit consent obtained where necessary. Data subjects will be informed of their rights regarding their data. Stakeholders have repeatedly identified GDPR as a primary legal framework, necessitating robust anonymization and data handling protocols before any data can be shared, even within the consortium. The challenge of robustly anonymizing sensor data (e.g., blurring faces/license plates in video, or removing PII from LiDAR point clouds) is non-trivial. The project will investigate and apply state-of-the-art techniques, such as k-anonymity or differential privacy, or rely on robust pseudonymization where full anonymization would destroy data utility.
- **Data Protection Impact Assessments:** We will conduct assessments to identify and mitigate risks associated with data processing activities. These assessments will be conducted before initiating any new large-scale data collection from human subjects or in public spaces, ensuring that privacy is built-in 'by design and by default'.
- **Partners which are established in a non-EU country** undertake to comply with their obligations under the GA and to respect general principles (including fundamental rights, values and ethical principles, environmental and labour standards, rules on classified information, intellectual property rights, visibility of funding and protection of personal data). This commitment is crucial for ensuring that data can flow securely within the consortium, maintaining a consistent, high standard of data protection and building trust between all partners, regardless of their geography.

9.4 Ethical considerations

Ethical considerations are integral to our data management strategy. We will ensure that:

- **Stakeholder Engagement:** All relevant stakeholders, including data subjects, will be engaged in discussions about data use and sharing. This extends beyond just data subjects to include public authorities, municipalities where data is collected, and the wider CCAM community, ensuring the project's 'social license to operate'.
- **Respect for Privacy:** We will prioritize the privacy of individuals and communities, ensuring that data collection methods are respectful and non-intrusive. This is not just a legal obligation but a cornerstone of public trust. If the public perceives

CCAM development as an invasive surveillance technology, it could significantly slow or halt its adoption.

- **Transparency in Research:** We will communicate openly about our research objectives and methodologies, fostering trust among stakeholders. This means publishing methodologies, not just favourable results. It also involves being open about the limitations of the data (e.g., 'data was only collected in clear weather') and the models derived from it, to prevent over-generalization of findings.

10 DATA SECURITY

Data security is primordial for the SYNERGIES Project, ensuring the protection of sensitive information and maintaining the integrity and confidentiality of data throughout its lifecycle. In order to maintain data security, the following principles are taken into account:

- **Confidentiality:** sensitive information will only be accessible to those authorized to access the data
- **Integrity:** the data stored needs to maintain its consistency over its lifecycle.
- **Availability:** data needs to be accessible and available to users when they need it.
- **Accountability:** the data will be traceable, its origin will be known and will be included in the data inventory that will be provided in the intermediate DPM.
- **Transparency:** data handling practices will be open and clear.
- **Data minimization:** only the data needed specifically for the purposes of the project will be gathered.
- **Resilience:** the data storage systems will be resistant to disruption, cyber-attacks and system failures. The intermediate version will provide the common practices by the data hosts.

Adherence to these security principles will ensure that the lifecycle of the data is properly followed.

Following partner consultations, it is confirmed that data hosts within the consortium employ a multi-layered set of common security practices to implement the principles above:

- **Access Control and Authentication:** Strict user management is implemented. This includes authentication (e.g., login/password, multi-factor authentication via "Authenticator" or Okta Verify) and authorization (e.g., Active Directory, group policies, NTFS rights, API tokens) to ensure only authorized personnel can access data.
- **Encryption:** Partners employ encryption for data both at rest (e.g., AES256 encryption, encrypted hard disks) and in transit (e.g., HTTPS, password-secured data transfers).
- **Backup and Redundancy Policies:** Backup policies are managed by each partner or cloud provider. Common practices include daily backups (for low-volume data), duplication of raw data to separate backup appliances, hardware redundancy (e.g., RAID), and the use of native cloud backup services (for AWS/Azure).
- **Standards and Policies:** Partners adhere to their internal data security policies (e.g., RISE data policy) and international standards such as ISO 27001 [\[8\]](#).
- **Physical Security:** For on-premise hardware, physical security measures such as locked server racks are applied.
- **Sensitive Data Access Governance:** Access to sensitive data is governed by partners' internal processes, often involving their Data Protection Officer (DPO). For specific platforms like ADScene, a dedicated data access committee exists. For personal data, informed consent is obtained for sharing and preservation.

11 DATA PUBLISHING AND LICENSING

11.1 Publishing

This process is thoroughly described in D10.1 (Dissemination and Communication Plan), which has already been submitted. To avoid duplication and potential ambiguity, it is not reiterated here. The information can be found in D10.1 Annex A.

Key aspects for publishing data:

- **Timing:**
 - Data may be published during development or at defined project milestones.
 - Publication can occur as soon as the data has been validated for quality, completeness, and regulatory compliance.
 - SYNERGIES database data will be released publicly once partners confirm quality.
 - Scenario data may be made available through public containers, depending on data owner preferences and Steering Committee approval.
- **Platforms:**
 - Open datasets will be published via partner repositories such as GitHub
 - Driving data publication may require an amendment to the Consortium Agreement to define terms/conditions.
- **Conditions:**
 - Only anonymized or sample data can be disseminated publicly.
 - Raw data containing personal identifiers will not be released publicly due to GDPR.
 - Data owners retain the right to define access levels, conditions, and any restrictions for public release

Guiding principles for public dissemination:

Public dissemination within SYNERGIES follows these principles:

- **Compliance:** All dissemination respects GA, CA, GDPR, intellectual property, and ethical/legal obligations, as well as institutional policies (e.g., RISE).
- **Data Owner Control:** Data owners determine the level and conditions of public access, ensuring that legal rights and partner preferences are respected.
- **Coordination:** Driving data publication is subject to Steering Committee decisions and WP agreements. VICOMTECH disseminates datasets alongside scientific publications, newsletters, social media (LinkedIn), and project websites.
- **Transparency:** Metadata and documentation accompany published datasets to facilitate re-use, traceability, and reproducibility.

Data utility outside the project:

SYNERGIES data is expected to be valuable for a wide range of stakeholders, including but not limited to:

- Regulators, technical services, and certification bodies.

- OEMs, suppliers, and tiers in the automotive ecosystem.
- Tool developers, AD developers, and AI solution providers.
- Researchers, universities, RTD departments, and academia in the CCAM field.
- Cities, local authorities, and human-in-the-loop research committees.
- Transport operators, safety experts, and vehicle simulation experts.

Raw data and scenarios generated within the project are particularly relevant for the development, testing, and validation of automated driving systems and related solutions.

11.2 Licensing

The terms under which data is licensed are agreed at the time of licensing but are always in accordance with the obligations set out in the Grant Agreement. Rules for licensing should be applied as following:

- **Metadata Licensing:** Where possible, metadata will be provided under a **CCo license** for the public domain in accordance with GA requirements, ensuring the widest re-use.
- **Data Licensing:** Data will be licensed using standard re-use licenses that are compatible with the Grant Agreement obligations. Licensing terms will be clearly documented to indicate permissible uses, attribution, and any restrictions.
- **Data Sharing Agreements:** External data sharing may require agreements or contracts specifying the purpose, duration, and restrictions of use.

12 COSTS

The costs of managing data in line with FAIR principles, including costs relating to archiving and long-term storage have been considered and included in the proposal by each partner. Financial management for the project shall be in line with the terms of the GA and CA and in as much as adherence to the terms of the GA and CA allows, partners shall bear the costs associated with making the data owned by the partner findable, accessible, and re-useable.

Costs for making data FAIR: Within project planned budget as defined in Grant Agreement. Specific costs identified include driving data format modification costs, cloud storage costs and management. Most partners indicate no direct costs or costs covered within indirect costs (e.g., IT and DPO at 10% of direct costs). Sample data stored in Git repositories requires no specific budget allocation.

Budget allocation for storage: As defined in Grant Agreement. Some partners requested storage budget, but it was removed during proposal phase. For some, data already stored and paid (no SYNERGIES budget needed). Yes, for storage budget but not for related maintenance service. Majority indicate no specific budget allocation; sample data stored with source codes in Git repositories. Most partners indicate no dedicated storage budget.

Long-term archiving costs: Partners have no funding to cover long-term archiving; ideally a platform with own funding would be used. Costs depend on archive strategy and data volume. Cloud costs rising due to energy costs; only evaluation possible. Maintenance and storage costs not known currently. Potential storage on HDDs (offline) planned but costs not yet determined. Sample data in Git repositories requires no specific allocation.

Cost coverage: Partners will cover costs themselves. Some have absolutely no money allocated. Costs covered by license fees. Most often covered by future research projects in same direction. Internal expenses at partner organizations (such as VICOMTECH). Sample data storage in Git repositories requires no specific budget allocation.

Other costs: IT and DPO part of indirect costs (partially funded by project at 10% of direct costs). Processing costs, web storage, maintenance and update of interface. Access management, dataset/software updates (no cost estimate available). Most costs not yet identified.

13 CONCLUSIONS

The document provides guidelines and best practices for data management within the SYNERGIES project, ensuring its successful implementation. More than a set of rules, this Data Management Plan (DMP) represents the foundational strategy for managing the project's single most critical asset: its data. It serves as the operational link between the project's legal framework the Consortium Agreement (CA) and the Grant Agreement (GA) and the technical execution of the work packages, particularly [WP3](#) (Data Analysis), [WP4](#) (AI Data Processing), [WP5](#) (Scenario Generation), and [WP7d](#) (Platform Establishment). The CA contains the data management provisions agreed between the project partners, and the GA contains the data management provisions agreed with the European Commission. These agreements are supplemented in the DMP with project-level guidelines for processes and procedures to ensure effective data management.

Responsibility for the implementation of the DMP lies with all project members according to their roles and responsibilities. This is not a passive 'compliance' task but an active, ongoing data governance challenge that requires a consortium wide culture of diligence. The Project Coordinator, Working Group Leaders, and Task Leaders will ensure compliance with the DMP under the guidance and supervision of the Data Protection Officer. Furthermore, this DMP is a critical tool for the Innovation and Exploitation Manager (IEM), as the clear definition of data access, IPR, and licensing (Chapter 11.2) is the prerequisite for any successful post-project exploitation and commercialization of the SYNERGIES platform.

Importantly, the DMP in no way replaces the legal documents of the project and their application and, in case of doubt, the legally binding documents take precedence. This DMP should be considered to be the document that operationalizes the legal abstractions of the GA/CA, translating them into specific technical and procedural controls, such as the data classification system (Chapter 8.3) and the access restriction protocols (Chapter 9).

The objective of this document is to outline the general data management plan for the project and the correct and appropriate handling of personal data that needs to be protected or that requires special treatment from an ethical point of view. However, the project faces a dual-mandate: it must not only protect personal data (PII) under the General Data Protection Regulation (GDPR) but also protect the high-value, commercially sensitive intellectual property (IP) of the partners, such as raw sensor data and proprietary algorithms. When collecting, processing or using personal data by a project member, the GDPR and other legal requirements must first be considered. These requirements are explained in the DMP. The federated model and clear access-control mechanisms described are designed to manage both of these "sensitive data" challenges simultaneously.

In addition, the FAIR Principles are highlighted in the DMP as an important guideline to ensure findability, accessibility, interoperability, and reproducibility of data. The principles are summarized in the DMP, together with an explanation of the commitment to follow these principles.

The SYNERGIES project intends to follow the FAIR principles wherever possible, considering commercial and practical constraints. This commitment is not merely academic; the FAIR principles are the technical enabler for the project's "solid AI foundation" (Objective 5). By standardizing on formats like ASAM OpenX (Chapter 6.3.2) and promoting a common metadata catalogue (Chapter 9.2), the DMP directly supports the interoperability (Chapter 8.4) needed to build, train, and validate AI models across heterogeneous datasets from 32 different partners.

The way in which the data is managed and disseminated is critical to the success of the project. The project unites a variety of partners with different internal processes and objectives from different sectors such as business, technology and research. This diversity of stakeholders (e.g., competing OEMs, suppliers, and public research institutes) makes a one-size-fits-all, "share-everything" model impossible.

It is essential to establish and maintain a common understanding of how data will be handled, the conditions under which data will be published and the procedures for agreeing these conditions. In this way, a culture of trust can be built that improves collaboration between partners and, ultimately, project outcomes. This "culture of trust" is built upon the DMP's *technical guarantees* namely, the federated platform model (Chapter 9) that respects data sovereignty by moving computation to the data, and the granular classification system (Chapter 8.3) that allows partners to share data with specific WPs without ceding control.

The DMP provides guidelines and recommendations for good practice that can be considered by partners in their daily interactions and in the appropriate application of due diligence.

As soon as the project partners recognize the relevance of the data and results for the public, they will evaluate the possibilities of publication in accordance with the guidelines of this DMP and the requirements of the CA and GA. This evaluation, led by the IEM, will assess data for multiple exploitation pathways beyond simple public release, including the creation of open benchmark datasets, assets for a sustainable post-project marketplace, or licensed data products for commercial use.

The DMP contains best practices and procedures for the provision of data within the project and for the archiving and availability of these data beyond the project period.

The SYNERGIES project recognizes that effective data access and management are vital for achieving its objectives. By adhering to the principles of open data where possible, ensuring compliance with data protection laws, and prioritizing ethical considerations, the aim is to create a robust framework that supports innovation and collaboration in the field of CCAM systems. This DMP intends to strike that critical balance, enabling collaboration while protecting confidentiality, thus creating a secure dataspace where the project's ambitious goals can be met.

This document provides initial guidance that explains project implementation in a simple and practical way, supports compliance with the underlying legal obligations and promotes the application of best data management practices. As the Mid-Term DMP (D1.3), this version already represents a significant evolution from the initial plan, incorporating the findings from two rounds of partner consultations (Chapter 2.2) and solidifying the project's technical choices on standards and data formats (Chapter 6.3).

As the project progresses, the DMP will be continuously updated to enable efficient integration, creation and management of FAIR data. This "living document" approach is essential. The updates in the final DMP (D1.4) will not just be a formality; they will form the definitive blueprint for the long-term governance and sustainability (Chapter 6) of the SYNERGIES scenario platform, transitioning it from a project-based asset to a lasting, valuable resource for the entire European CCAM community.

14 REFERENCES

- [1] European Data Protection Board (EDPB). (2021). "Guidelines on Examples regarding Data Protection by Design and by Default." Available at: <https://edpb.europa.eu>
- [2] Wilkinson, M. D., et al. (2016). "The FAIR Guiding Principles for scientific data management and stewardship." *Scientific Data*, 3(160018). DOI:10.1038/sdata.2016.18.
- [3] SYNERGIES Consortium (2024). Grant Agreement.
- [4] SYNERGIES Consortium (2024). Consortium Agreement.
- [5] OECD. (2021). "Recommendation of the Council concerning Access to Research Data from Public Funding." DOI:10.1787/9789264252875-en.
- [6] International Organization for Standardization (ISO). (2019). "ISO 19115-1: Metadata – Part 1: Fundamentals." Available at: <https://www.iso.org>.
- [7] Smith, J., & Dodoiu, T. (2023). *D1.3: Data Management Plan*. SUNRISE. Available at: <https://ccam-sunrise-project.eu/deliverable/d1-3-data-management-plan/>.
- [8] International Organization for Standardization (ISO). (2022). 'ISO/IEC 27001:2022 Information security, cybersecurity and privacy protection – Information security management systems – Requirements.' Available at: <https://www.iso.org>.

A. QUESTIONNAIRE

Data Description

- What is the origin/provenance of the data (generated/ re-used)?
- What types and formats of data will the project generate/re-use?
- Are you using a file format that is standard in your field? If not, how will you document the alternative you are using?
- What existing data do you expect to use within the project (re-use)?
- What methodologies or software for data collection or production are you going to use?

Data Management

- What data management standards do you expect/require for the project?
- What tools or software are required to manage the data?

Data Storage

- Where and how is the data stored?
- What is the expected volume of the data that you intend to generate or re-use?
- Who decides and how, what data to keep and for how long?
- Will the data be identified by a persistent identifier? If so, how (e.g. Digital Object Identifiers)?
- What metadata will be created?
- What general standards will be followed? If metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.
- Outline the approach to search keywords. Will keywords be provided in the metadata to optimise the possibility of discovery and subsequent re-use?
- Will metadata be provided in such a way that it can be harvested and indexed?
- Will documentation or references to any software required to access or read the data be included? Will it be possible to include the relevant software (e.g. in open-source code)?

Data Archiving

- How will long term preservation be ensured?
- Until when will data be available?

Data Exchange and Sharing

- What are the guiding principles for data exchange internally and externally?
- What is the expected data flow in the project?
- Who owns the data and has responsibility to define who data may be shared with?
- Under what conditions will data be made public?
- What are the processes for agreeing the conditions under which data will be made public?

- How will data that is intended to be shared externally be classified and labelled?
- What data and metadata vocabularies, standards, formats or methodologies will you follow to make your data interoperable to enable data exchange and re-use?
- Will you follow community-endorsed interoperability best practices? If so, which ones?
- Will your data include qualified references to other data (e.g. other data from your project, or datasets from previous research)?
- How will you provide documentation needed to validate data analysis and facilitate data re-use (e.g. readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, units of measurement, etc.)?
- Will the data produced in the project be useable by third parties, in particular after the end of the project?
- Describe all relevant data quality assurance processes.
- Specify the length of time for which the data will remain re-usable.
- Specify when the data will be made available for re-use.
- If applicable, specify why and for what period a data embargo is needed.

Data Access and Restrictions

- Will the data used in the project be open data?
- If data of the project is open, what is the way in which it would be shared?
- If the data of the project is not open data who will be able to access it (all partners, only ones that provided the data, etc.)?
- If the data of the project is not open to public, for how long will the partners have rights to access it?
- Will every request for sharing the data be individually assessed?
- Will the data, after the duration of the project, become publicly available?
- Would the partners be able to use the data after the project is completed?
- Will the data from other partners be available even after the project is concluded?
- After the project conclusion, will the partners be obliged to destroy the data?
- How is the destruction of the data checked and enforced?
- What are the regulations and enforcements for data deletion, and will there be governing body that checks if the deed is done?
- How will we do the data access? How will be the governing party to decide the data access?
- How is the data provided by the partners weighed since not all partners will provide same amount of data?
- Who is dedicated and ensure that restrictions are met?
- On which level data sharing will be done between partners?
- Are there, or could there be, any ethics or legal issues that can have an impact on data sharing?
- Will the data be accessible through a free and standardized access protocol?

- If certain datasets cannot be shared (or need to be shared under restricted access conditions), explain why, clearly separating legal and contractual reasons from intentional restrictions.
- If an embargo is applied to give time to publish or seek protection of the intellectual property (e.g. patents), specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.
- Will metadata contain information to enable the user to access the data?
- What tools or software are required to read or view the data?
- Will documentation or reference about any software be needed to access or read the data be included? Will it be possible to include the relevant software (e.g. in open source code)?
- How long will the data remain available and findable?
- Will metadata be guaranteed to remain available after data is no longer available?
- Specify where the data and associated metadata, documentation and code are deposited.

Data Security

- Where is data stored?
- What security measures are there?
- Is there a back-up policy? If so, how it work?
- Will there be any paper-based documentation stored? How will it be stored?
- How will the identity of the person accessing the data be ascertained?
- Specify how access will be provided in case there are any restrictions
- What are the guidelines for the secure storage and transfer of sensitive data?
- Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)?
- Will informed consent for data sharing and long term preservation be included in questionnaires dealing with personal data?

Data Publishing and Licensing

- How and when will the data be published and/or licensed to permit the widest re-use possible?
- What are the guiding principles and agreed processes regarding public dissemination?
- To whom might your data be useful ("data utility") outside your project?
- Will the metadata be provided with a CC0 licence for the public domain in accordance with the grant agreement?
- Will your data be licensed using standard reuse licenses, in line with the obligations set out in the Grant Agreement?

Costs

- What are the costs for making data findable, accessible, interoperable and re-useable in your project (e.g. direct and indirect costs related to storage, archiving, re-use, security, etc.)?

- Is there budget allocation for storage?
- Describe the costs for long-term archiving and are they available?
- How will these costs be met?
- Are there any other costs? If so, what costs and how much?

B. ABBREVIATIONS AND DEFINITIONS

Term	Definition
AD	Application Development
AI	Artificial Intelligence
AIMATS	Analysis and Investigation Method for All Traffic Scenarios
API	Application Programming Interface
ASAM	Association for Standardization of Automation and Measuring Systems
AWS	Amazon Web Services
CA	Consortium Agreement
CARE	Community database on road accidents resulting in death or injury
CCO	Creative Commons
CDF	Common Data Format
CSV	Comma-separated values
CCAM	Cooperative, connected and automated mobility
DM	Dissemination Manager
DMD	Driving Monitoring Dataset
DPA	Data Processing Agreements
DPIAs	Data Protection Impact Assessments
DPO	Data Protection Officer
DoA	Description of the Action
DMP	Data Management Plan
EC	European Commission
EDPB	European Data Protection Board
EU	European Union
FAIR	Findable, accessible, interoperable and reusable
GA	Grant Agreement
GDPR	General Data Protection Regulation
GIDAS	German In-depth Accident Study
GNSS/IMU	Global Navigation Satellite System/Inertial Measurement Unit
HDD	Hard Disk Drive
HTTPS	Hypertext Transfer Protocol Secure
IEM	Innovation and Exploitation Manager
IGLAD	Initiative for the global harmonization of accident data
IPRM	Intellectual property rights Manager
IPR	Intellectual property rights
IP	Intellectual property
IT	Information Technology

ISO	International Organization for Standardization
JSON	JavaScript Object Notation
KB	Kilobyte
LCM	Liaison and Collaboration Manager
LiDAR	Light Detection and Ranging
MB	Megabyte
NCAP	New Car Assessment Programme
NTFS	New Technology File System
ODD	Operational Domain Design
OECD	Organisation for Economic Co-operation and Development
OEM	Original Equipment Manufacture
OSI	Open Simulation Interface
PCM	Pre-Crash-Matrix
PDF	Portable Document Format
PII	Personally Identifiable Information
R&D	Research and development
RAID	Redundant Array of Inexpensive Disks
RBAC	Role-Based Access Control
RDF	Resource Description Framework
ROS	Robot Operating System
RTMaps	Real-Time Multi-sensor Applications
TASI	Traffic Data Analysis and Situation Interpretation
TB	Terabyte
TBD	To Be Determined/Decided
TXT	Text file document
UC	Use case
UK	United Kingdom
WP	Work Package
XML	Extensible Markup Language