CCAM Data Sharing Framework

Version 2

Published: 2024-02-15

Author: FAME support action

Acknowledgements

We would like to express our gratitude to everyone who has contributed to this document.

This document provides guidance for projects and organizations to share and re-use data related to testing new technologies in the fields of transportation, mobility, and automotive innovation. It has a particular focus on cooperative, connected and automated mobility (CCAM) testing and Field Operational Tests (FOTs). It is the combined effort of three EU support actions focused on networking and knowledge sharing: FOT-Net Data, CARTRE, and FAME.

The FOT-Net Data Sharing Framework (DSF) was first released in late 2016 by the FOT-Net Data consortium. Since then, these practices have been adopted wholly or partially by numerous European and international projects. In May 2018, the European General Data Protection Regulation (GDPR) was implemented, rendering some parts of the data protection chapter obsolete. Consequently, the CARTRE project updated the framework, primarily focusing on the data protection chapter but also making minor improvements to other sections.

The FAME EU project, which fosters engagement, cooperation, consensus-building, and knowledge sharing among CCAM stakeholders, is responsible for the latest update of this document. The primary goal of this update is to promote the sharing of test datasets among CCAM stakeholders, with the aim of facilitating more effective testing and development of new CCAM technologies.

The following organizations or individuals from organizations have contributed directly or indirectly to the CCAM Data Sharing Framework as members of at least one of the consortia of the FOT-Net Data, CARTRE, or FAME coordination and support actions:

Aalborg University	Denmark
AustriaTech	Austria
CEESAR	France
Chalmers University of Technology (SAFER)	Sweden
Daimler	Germany
European Road Transport Telematics Implementation Coordination Organisation (ERTICO)	Belgium
Eindhoven University of Technology	Netherlands
Stellantis (FIAT Research Centre (CRF))	Italy
Forschungsgesellschaft Kraftfahrwesen Aachen (FKA)	Germany
Noblis	United States
PTV	Germany
Renault (LAB)	France
Technical University of Aachen (IKA)	Germany
University of Leeds	United Kingdom
University of New South Wales	Australia
VIAS institute	Belgium
Vicomtech	Spain
Volvo Car Corporation	Sweden
Volvo Technology Corporation	Sweden
VTT Technical Research Centre of Finland Ltd	Finland

Table of Contents

1 Executive summary

Since the early 2000s, the transport and mobility sectors have seen significant growth in Field Operational Tests (FOTs) and Naturalistic Driving Studies (NDSs), worldwide. These studies have focused on testing the feasibility and societal impacts of vehicle-to-X connectivity and Advanced Driver Assistance Systems (ADAS), among other things. Over the past decade, the field of Connected, Cooperative and Automated Mobility (CCAM) has emerged as the leading area of research in the sector, where *Automated* is the key element (with or without *Connected* and *Cooperative* capabilities). While the initial CCAM tests were conducted on a small scale to ensure safety and test automation technology, the next wave of tests is set to match the previous FOTs in scope. In parallel, applications built on Cooperative intelligent Transport Systems standards and next generation ADAS, were deployed to the market.

Arranging tests on public roads is a resource-intensive task that often takes years to complete and requires multiple partners' involvement. These research and development projects produce valuable datasets that could be useful for other organizations and purposes beyond the original project plan. Consequently, there is a growing interest in data sharing to extract more results from the large data collection efforts.

To facilitate greater use of the collected test data, this document presents a data sharing framework, including recommendations facilitating accessibility and data interoperability on a European level. The framework integrates data sharing pre-requisites into project agreements from the start, suggesting procedures and templates. Using a common framework promotes harmonization across projects and ensures that basic criteria are met regarding data protection and possibilities for reuse for public and private organisations active in the CCAM domain.

The challenges, both legal and technical, when sharing data has limited its potential. As a response to that, European companies, authorities and research institutes, have set the focus on *federated data sharing*. The principle is that by establishing trust between stakeholders, within a specific domain or community, giving the data providers control of who has access to which data for which purpose, more data can be shared. The approach, together with using harmonized data formats and taxonomy, has the potential in unlocking relevant data for research purposes, explored in a CCAM-context in this document.

The proposed framework support organizations setting up new tests, in highlighting important data-related topics, enabling them to focus on the primary content of their research, such as research questions and study design. Furthermore, researchers, who want to reuse already collected datasets or multiple datasets in the same research project, can utilize a standard application procedure, rely on widely accepted training, and plan for the costs associated with using a particular dataset.

This document covers key aspects of data management, including detailed data and metadata descriptions, data-protection guidelines, insights on personal data and intellectual property rights training, and outlines of support and research services and financial considerations. It's designed as a comprehensive guide for both new and experienced researchers in the CCAM field, promoting effective data management practices and collaboration.

2 Introduction

This document presents a data sharing framework developed to facilitate sharing of data from Naturalistic Driving Studies (NDS), Field Operation Tests (FOT), pilots and living labs, with the aim of increasing re-use of data for research and validation purposes, and scaling of projects results from testing activities on public roads. It offers guidance on the following topics: (chapter 3) data sharing agreements; (4) data and metadata descriptions; (5) data protection; (6) training; (7) support and research services; (8) financial models; and (9) data governance procedures.

Tailored specifically for research and development projects, the framework facilitates the management and evaluation of Connected, Cooperative and Automated Mobility (CCAM) data, where *Automated* is the key element (with or without *Connected* and *Cooperative* capabilities). The scope of the framework is focused, but not limited to, data collected from automated vehicles tested on *public roads* (thus excluding water, rail or air transport modes). In many cases, a combination of higher and lower automation data is collected and analysed in a project, why we have lower automation in mind while developing the framework.

Out of the seven topics, five delve into administrative facets, while the chapters on data and metadata, as well as data protection, are technically inclined. These principles, though crafted within a CCAM FOT/NDS/pilot/living lab context, have demonstrated applicability across varied research domains, including engineering and life sciences.

The CCAM Data Sharing Framework facilitates data sharing regardless of the size or content of a dataset. The framework is well-suited for large datasets, including both confidential/commercial data and personal data. Sharing large datasets imposes a greater effort in all the above-mentioned areas compared to a dataset with only a few signals and no video. However, when sharing smaller datasets, some sections of the framework might become less relevant, depending on the specific dataset. Nevertheless, each chapter provides advice and recommendations applicable to a wide range of situations.

This framework primarily addresses the exchange of semi-confidential data and doesn't centre on open-access data repositories or those accessed by license. Nevertheless, certain elements of the framework can be relevant to data repositories. For instance, the sections on data description or financial models may apply, as might aspects relevant to other CCAM scenarios like data from tests in restricted areas, test tracks, or simulators. The framework's main goal is to offer guidelines for bi-lateral data exchanges, whether through direct data transfer, remote desktop access, or federated data sharing.

The first release of the Data Sharing Framework was developed in FOT-Net Data project (2014–2016), with a primary focus on FOTs and NDSs. It was later updated for GDPR compliance in the CARTRE project (2017–2020). The current version, emerging from the FAME project (2022–2025), expands its reach to CCAM datasets and incorporates federated data sharing, alongside parallel deployment in the C-ITS domain.

2.1 Background

Over the past two decades, advancements in vehicle fleet data collection methodologies for research purposes emerged primarily from NDS and FOT. These studies were driven by two main factors: the need to better understand the causal factors behind incidents and accidents, and the progressive innovations in various driver assistance systems, leveraging cost-effective sensor, communication, and data server technologies.

It became essential to define and document best practices for carrying out these extensive trials, leading to the development of FESTA Handbook in 2008. The handbook has since received several updates, the latest being version 8 (FESTA, 2021). It covers the full process of running field operational tests: from formulating research questions and preparing the test, to analysing collected data to answer these research questions. While extensive, FESTA cannot cover data management and data sharing aspects in high detail. Therefore, in 2016, the first version of this Data Sharing Framework was released, later updated in the CARTRE project (applying General Data Protection Regulation (GDPR)) and ARCADE project (adding automated driving topics) (DSF, 2021).

By the mid-2010s, the focus of research turned towards automated driving technology, fuelled by advances in machine learning and neural networks. There was a true hype in the latter part of the decade aiming at a fast introduction of automated vehicles and services on public roads. The challenges were however many and, even though there has been significant progress, much is still to be done. The shift also brought new dimensions to data sharing. Now, the domain grapples with large datasets that include both developmental sensor data and driving behaviour data to assess traffic changes. New regulations have also come into play, like test permits mandating minimal data collection, but also voluntary agreements from industry participants and public authorities, e.g., to share data for testing and validation purposes, to faster introduce new and adapted vehicle functionalities on the market (C-ITS, 2023)

Data is vital for the development of automated vehicles. Initially, NDS and FOT data was used for building driver models and establishing baseline for driver behaviour across various scenarios. Soon after, data suited for training machine learning (ML) models began to be shared, primarily for research purposes, leading to an increase in datasets that address different aspects of driving. Currently, projects are collecting data from vehicles working on test tracks or in confined areas or conducting tests on public roads limited by conditions specified in the Operational Design Domain (ODD) of the system. now the aim is to extend the operational conditions for automated vehicles and initiate large-scale demonstrations or "living labs". The data from these projects will be invaluable in validating systems and assessing their impact on traffic safety, efficiency, the environment, and society at large.

In parallel, Cooperative Intelligent Transport Systems (C-ITS) were developed and deployed on the market, mainly with a large step as VW included C2C and C2X capabilities in the ID family (Gu, 2021). These “DAY1” features are based on a set of open and documented standards with full access of all participating stakeholders to testing and validation of these specifications, and the C-ITS components serving them. The EU, together with the strong European automotive industry, road operators and member states, launched the CCAM

partnership[1] in 2020 to address the needs, requirements, and strategies of automated driving. Starting in 2022, numerous projects under this programme (being part of Horizon Europe), as well as many national projects, have been advancing research in automated driving, with vehicles and road infrastructure elements with functions that enable fully fledged testing on public roads.

There are numerous challenges being addressed and data are a central element to overcoming them, the voluntary agreements between industry and public authorities mentioned above are widely supported and can cover the necessary data elements and types to address some challenges of testing and validation. However, no single entity can tackle these challenges by themselves – none possess the comprehensive data required to develop and validate the functions, systems, and vehicles. This framework introduces best practices for data exchange to facilitate and increase research data sharing.

2.2 Why share and re-use data?

Data sharing and re-use form an integral part of the European Strategy for Data (Data strategy, 2020). The Open Data Directive (EU) 2019/1024 (European Directive 95/46/EC Art. 2., 1995) sets guidelines for using data, including research data. If the data supplier also does research, sharing can lead to more collaboration and analysis. Sharing data can spark more research projects and raise chances of getting research funds.

Data sharing is essential for partners within a research project (a necessity), funding bodies, and organizations that focus on data exchange. There are today organisations openly sharing data at no cost, however, often limited to research purposes. This is common in AI and machine learning training datasets. They might share due to funding rules, to gain scientific prestige, or to showcase their technology.

Data providers invest significantly in collecting data and developing the related infrastructure and tools. Maintaining, assessing the quality, and sharing the data requires specialized staff to bring the data and tools to a level where they are easy to use. It is therefore essential to understand how to compensate data providers for their efforts. Providing some benefit would also increase the number of data providers willing to share their datasets.

The data provider often performs research on their own. Collaborations involving further analysis may create new funding opportunities and stimulate a large variety of research projects.

The original project collecting the data may only perform a narrow analysis based on their project's research objectives. From a funding organisation´s perspective, utilising the existing datasets for further analysis is an efficient return on investment. For project partners who already know the data, being able to further use it in new projects is good payback on invested efforts. During this additional phase of data reuse, the funding organisation could require that additional partners are brought in, to expand the data's reach.

With the vast data coming from around the world, combining datasets can offer more reliable results than using just one. Studying specific groups, like older drivers, in various countries

[1] https://www.ccam.eu/

can reveal how traffic behaviour varies culturally. If extra research funds are tied to global partnerships and data sharing, it boosts the worldwide research community. Collaborating on research builds trust, encouraging more data sharing and broadening knowledge.

These are some of the general advantages of sharing and re-using datasets. It is important, though, to identify the special circumstances that create a win-win situation between the data provider and the researcher in each specific case. At the same time, we must protect the information of those who participate in the research.

Even though these incentives have been clear for some time, data owners and providers have hesitated to share. The risks, highlighted in the European Data Strategy's section on B2B data-sharing challenges, have also affected research datasets. To tackle this, the recent European approaches emphasize 'federated data sharing', a system enabling decentralized data control, and this approach is standardized by the GAIA-X initiative[2]. This method is introduced in this document as an alternative to traditional data exchange and remote data access. It offers data owners and providers greater control over their data and its utilization.

2.3 Data sharing approaches

This framework targets three different approaches to data sharing:

- *traditional data sharing* (i.e., transferring data, including the rights to use the data, from one actor to another)
- *on-site or remote data sharing* (i.e., granting a user access within a controlled environment, or an external user access to a remote desktop or server, with access to data and tools within an infrastructure governed and controlled by the data provider)
- *federated data sharing* (i.e., granting an external data user access to a small component within a trusted federated ecosystem, governed by the data provider, with control mechanisms for what data can be accessed and what is allowed to be extracted)

It should be noted that C-ITS peer-to-peer data exchange (C2C or C2X) is per se a federated data sharing approach (edge to edge), already deployed on the market (VW Golf and ID.x using ETSI ITS G5 / IEE802.p/bd). However, the usage of federated data sharing in this document concern data sharing between centralized data centre nodes.

Traditional data sharing (data download)

This approach includes transferring data to another stakeholder using any media or via network access. The data is being copied to the data users' environment and then used for any purpose agreed within the data provider in an agreement. This approach has been the most common, however, the volume and methods sharing the data have changed over the years. The methods include sharing a physical media (e.g., a hard drive), using or using secure network protocols (https, ftps, ssh).

[2] https://gaia-x.eu/

On-site or remote desktop

The benefits of remote desktop or on-site access is that the raw data must not necessarily be exchanged since the user connects to an environment in full control of the data provider. This means that any operation on the data is done within the environment and the data provider can decide on what data can be accessed. The environment is often restricted for data extractions (by different means depending on the sensitivity of the data), and if an extraction is accepted, the traditional data sharing approach is used.

The benefit is that the user can get access to relevant data, but the data provider is in full control. By allowing remote data access, the user need not be at the same office (or even country). The data user must, however, accept the conditions and data protection measures stipulated by the data provider.

Federated data sharing

Federated data sharing represents a method where various parties can share data without centralizing it. It is often associated with the concept of 'data spaces', which refers to collaborative environments for sharing data across different sources under governed conditions. Two commonly used terms are:

Federated Data Space (FDS), which refers to an integrated approach to managing and accessing data that is distributed across multiple systems or locations. In a federated data space, different data sources maintain their autonomy but are interconnected in a way that allows for unified data access and analysis.

Federated Database System (FDBS), which is a type of database management system that allows for the management and integration of multiple autonomous databases into a single federated database. Each participating database in a FDBS remains independent but can be accessed and queried as part of the federated system.

The concept of federated data sharing is an approach that could be seen as a compromise between the two previous methods of traditional data download and on-site or remote desktop access. However, it introduces a fundamental shift in the direction of data flows. The principle is that trust is established between actors in a network, and common principles for data access, description, and formats facilitate data exchange. This approach includes infrastructure and software tools, authentication and authorization, cybersecurity and data protection, taxonomy, data catalogues, and governance. These aspects have been described in European Data Strategy and GAIA-X technical specifications (GAIA-X, 2023) and implemented in various software platforms, such as international Data Space (IDSA, 2022), X-Road (X-Road, 2022), and Eclipse[3].

Compared to data download or sharing access to a common database, the principle of FDS is granting access to specific data based on an agreement that governs which data the user is allowed to access. This means that original data does not necessarily need to be sent to a user; instead, only the product of a computation is shared.

[3] https://eclipse-edc.github.io/docs/#/README

2.4 FAIR

A prerequisite for data reuse, both within a project/organization and externally, is that the data can be found, read, and interpreted. A structured approach to data creation can facilitate reuse and save time later. One set of relevant principles, generally favoured by the international research community and promoted by the European Commission, are the FAIR principles. FAIR stands for Findable, Accessible, Interoperable, and Reusable. The aim of these principles is to make data and metadata (data about data) machine actionable as well as human-readable. The principles can be viewed in Wilkinson et al. (2016).

These principles apply to three categories: data (or other digital objects), metadata (data about data), and infrastructure (such as a data repository or data space). The principles have some implications on how (meta)data is created and described, as well as how storage, search and access infrastructures are set up.

Findability

Digital resources (data and metadata) should be easy to find for humans and computers. Machine-readable metadata are essential to make data searchable and findable and, allow for easy transfer of metadata between services. Persistent identifiers for data and metadata make sure that data can be cited and shared effectively and without uncertainties.

Data creators can increase findability by selecting identifying a data repository/space early on and finding out about its data and metadata requirements, making sure that it also provides persistent identifiers and a catalogue service (internal or external to the repository or data space).

Accessibility

Access and authorisation need to be clearly defined for both user and machine. Infrastructures need to select open, free, and universally implementable standardized communication protocols which also allow for authentication and authorization procedures (if applicable). Metadata should remain accessible even if the data is no longer available (usually via a tombstone page[4]).

Data creators can facilitate accessibility by clarifying and describing the legal conditions for making data accessible, setting embargo periods (if necessary) after which data can be made available, and making sure that the selected repository/data space guarantees data longevity and availability.

Interoperability

Data often need to be integrated with other data or within larger workflows. Making use of open, formal, standardized language, vocabularies, formats, etc., when creating data and metadata facilitates this interoperability.

To increase interoperability, data creators can make use of commonly used data formats, ideally openly described, and using openly, commonly used vocabularies for data.

[4] https://support.datacite.org/docs/tombstone-pages

Reusability

Rich metadata descriptions are key to reusability. This means good documentation and making data accessible via an infrastructure that provides (and requires) rich metadata. Data creators can enhance reusability by keeping well-documented data provenance, as well as selecting and using appropriate and openly described metadata standards. Set a licence for data when sharing the data.

These FAIR principles do not force any specific technical implementation. Neither are they a standard. Rather, they are a set of guidelines aimed at improving data reusability. Not all principles may apply or be possible to implement in all situations. These principles should be used to make data and infrastructures supporting them as FAIR as possible given existing factors and limitations. In practice, Science Europe's (2021) data management plan guide is a useful tool for implementing FAIR at project level.

Observe that *FAIR*-ness is not a binary state – data or metadata are not either FAIR or not FAIR. Implementing FAIR is usually an incremental process where (meta)data becomes increasingly FAIR as FAIR principles are implemented. Not all FAIR principles and steps apply to all types of data. See Appendix I for questions related to FAIR implementation.

Finally, FAIR does not imply openness: FAIR data can be restricted if necessary - the maxim "As open as possible, as restricted as necessary" is applicable in all cases.

2.5 Overview of the CCAM Data Sharing Framework

In the following chapters, we outline the CCAM Data Sharing Framework. Figure 1 shows the seven key areas it covers.

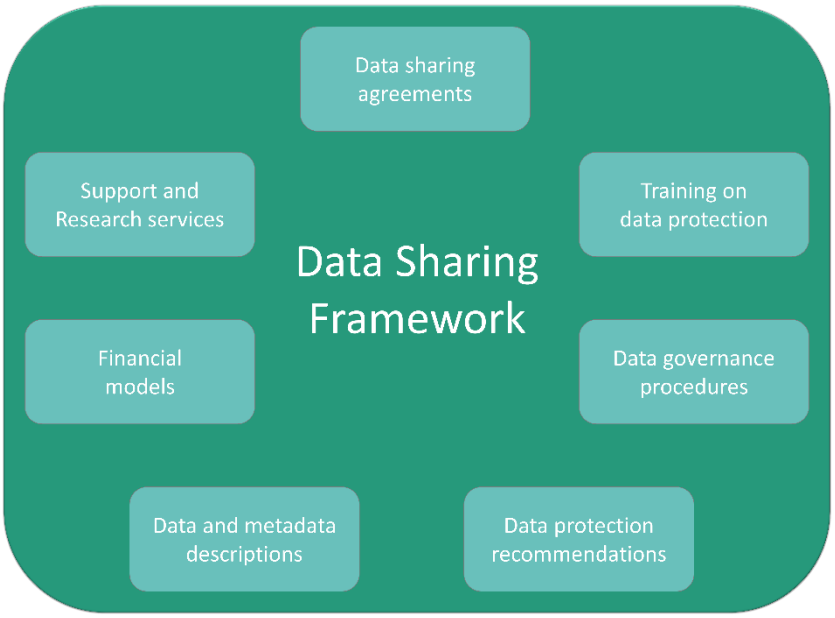

Figure 1: CCAM Data Sharing Framework

The CCAM Data Sharing Framework consists of:

- *Project agreement* content, including guidelines and checklists to incorporate the pre-requisites for data sharing in the agreements, which together with legal and ethical

constraints form the conditions for data sharing. The project agreements include the grant agreement (together with the description of the work), the consortium agreement, the participant agreement and data provider agreements.

- *Data and metadata description* recommendations, to facilitate the understanding of the context in which the data was collected and the validity of the data. These include a suggested structure for the documentation of data, divided into four categories: study design and execution documentation, descriptive metadata (e.g., how the data is calculated or sampling frequency), structural metadata (e.g., how the data is organised) and administrative metadata (e.g., access procedures).
- *Data protection* recommendations, focusing on personal and confidential data issues. It consists of security procedures and requirements for actors involved in data exchange, including detailed implementation guidelines.
- *Training* on security and human subject protection for all involved personnel. The guidelines cover four topics: who should be trained and when, what content should be part of the training (including detailed suggestions), how to do the training, and how to document it.
- *Support and research services*, proposing functions such as providing information/training to facilitate the start-up of projects, offering (for example) processed data for researchers less familiar with data, making analysis tools available or performing complete research tasks. This section also addresses the on/off boarding in a federated data sharing context.
- *Financial models* to provide funding for the data to be maintained and available, and data access services. Nine financial models are discussed, and a list of data management costs is provided.
- *Data governance procedures* which describe topics to consider when defining a protocol on allowance to use data and on/off-boarding a data space.

Each area has its own chapter that follows. It contains lessons learned from previous large European projects, recommendations, checklist and procedures, and raises questions for a project to be considered. Many of these topics that are addressed should be consider early on in the project phase, in order to success in utilizing the most of the research data available within this domain.

3 Data sharing agreements

The following chapter describes strategies for enhancing data sharing through common research project documents and agreements. These are particularly relevant in funded research involving multiple partners and stakeholders, where there is a collective interest in contributing to or accessing data. The three first agreements ((3.1 Funding agreement, 3.2 Consortium agreement and 3.3 Data management plan) are in the context of a project structure, the agreements described from 3.4 and onwards are more general agreements, also suited outside of a project consortium.

The initial process of setting up a project is crucial to enable data to be shared during and after a project. While agreements can be renegotiated, this process is often time-consuming and costly, especially in large consortia operating under pre-established conditions. Project agreements address various topics, with only a subset directly related to data sharing. Therefore, the time spent during the project application and at the beginning of the project to agree on the conditions for data access and use (including data re-use after the project) is well invested.

The funding agreement (including the work description) and the consortium agreement among project partners are the main documents to focus on initially.

It has become increasingly common during the project to draft a data management plan, which outlines how data will be collected, managed, and shared among project partners. Later, during the project and its data collection (which may include personal and confidential data), participant agreements and any agreements with data providers are important to define the conditions and scope for data re-use and management. Participant agreement and data provider agreements are important in the case of data re-use.

It may be necessary to have an agreement between actors in a common data space or a data repository. The agreement outlines the terms and conditions for data sharing between different actors within the data space, defining each actor's roles and responsibilities and specifying the conditions for joining, accessing and using data. Ensuring all actors are aligned on data sharing terms and conditions is crucial.

There may be overlap on the matters relating to data sharing between these documents. Where there is overlap between project documents, it is imperative that the terms do not conflict or contradict each other.

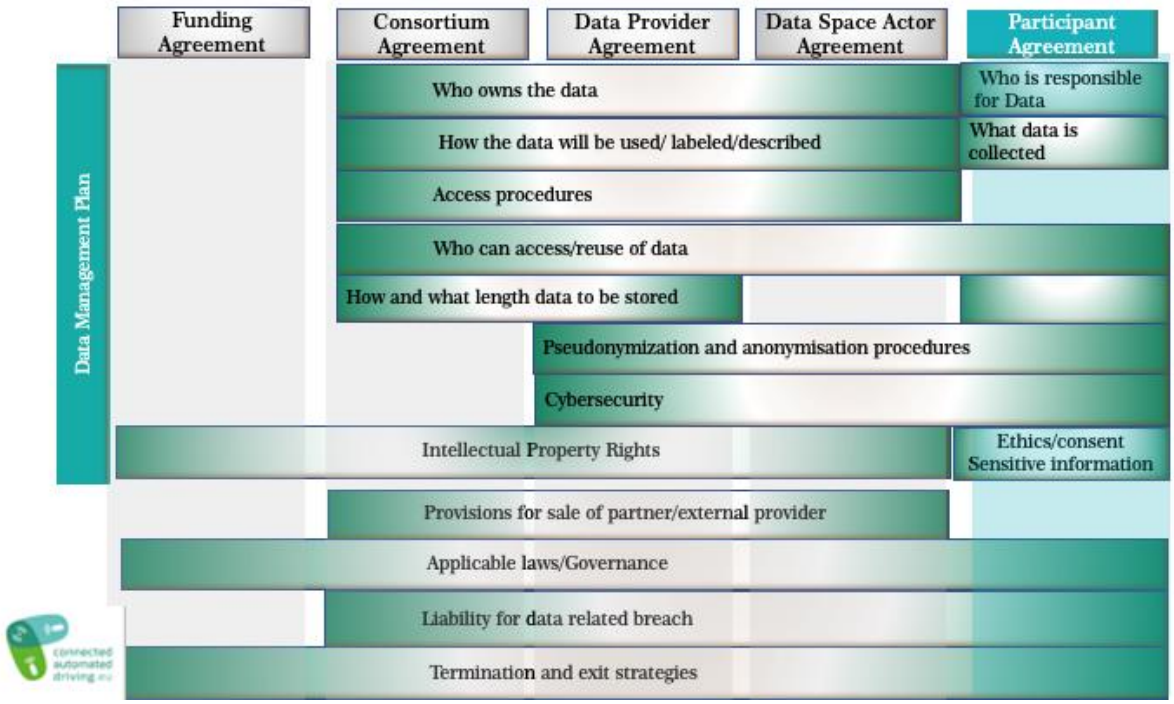

Figure 2: Project documents and how they may overlap on data topics.

This chapter discusses which topics to concentrate on, from a data-sharing perspective, for each specific project document.

3.1 Funding agreement including the description of the work

The level of preparation required for a funding agreement to facilitate data sharing and re-use will vary depending on the funding body. Projects funded by Horizon 2020 and Horizon Europe are already subject to specific requirements for data sharing and data availability, as these principles are integral to their funding agreements. Such projects align with EU policies promoting open access and open data.

For instance, projects participating in the European Commission's Open Research Data pilot[5] (initiated in 2017 under Horizon 2020), were required to upload datasets in an open file format to an appropriate data repository before the project ended. While the pilot included exceptions for projects collecting personal data, it marked a significant step towards the EU's commitment to open data. Subsequent EU open data initiatives, such as those under the European Open Science Cloud (EOSC), further expand the scope and align with FAIR data principles.

The application process and subsequent steps prior to the commencement of the research should involve clarifying the topics. The funding agreement should also address the research objectives, description and scope of the work, timescales for deliverables, re-use of data, funding budget parameters and reporting requirements.

[5] https://www.openaire.eu/what-is-the-open-research-data-pilot

3.2 Consortium Agreement

The consortium agreement (CA) sets out the legal relationship between the consortia partners, detailing their responsibilities and agreed terms of collaboration. The CA should also include the terms governing data sharing and re-use of the data. The issues which should be clarified and between the partners or parties to the agreement are included below in Table 1, which provides guidance in identifying the issues relevant to sharing and re-using the data after the project. The text of the CA should be broad. Finer details relating to data sharing can be reflected in a data management plan. Before the project ends, it is important to have a comprehensive written agreement for how the data should be handled after the project.

Table 1: Data-sharing topics within the consortium agreement

Topic	Issues to be considered
Ownership and access to data and data tools	• Who owns the data? • Is it necessary to add a specific ownership clause for the collected data? • How could the data be used and on which conditions? • Will all partners have access to all/part of the data? • May the data be licensed to third parties? • May third parties have access to the data and on what conditions? • Will any data become publicly available (with or without any license) and how will this data be provisioned? • Are there constraints related to personal data, especially video and tracked location data or location data in general? • Are there future agreements with data providers to take into account? • Who will own the analysis tools and on what conditions are they licensed during and after the project? • Does any partner have any intellectual property rights over tools which may impact their use? • How can data be re-used if the data is owned by one partner and this partner ceases operation or leaves the project? • Will the data generated be subject to proposed frameworks such as the EU Data Act or EU AI act which may require that data be provided to users and third parties
Storage and download of data	• How will the data be stored during and after the project – centrally, distributed, by a third party, or combinations? • What are the general requirements for data protection and how are they assured? • Shall all/part of the data be downloadable for all partners and if so, under which conditions?

	• Shall all/part of the data be downloadable for third parties and if so, under which conditions? • Is there a time limit for requesting data for download? • Is there a time limit for keeping the data?
Access methods	• Can the data be downloaded, remotely accessed, accessed using connector within a data space, or only accessed at the premises of any partner? • Shall a specific access procedure be used, and if so by whom? • Who shall govern the access procedure? • How will the data be accessed?
Cybersecurity	• How is the data infrastructure protected from both a technical and organisational perspective? • Is there a plan to prevent and respond to a data leak, phishing attempt, or attack (i.e., encrypting files)?
Areas of use	• Shall it be possible to use the data for education, research and commercial purposes? • Are there special conditions for commercial use? Can predefined licenses be applied? • In which research/commercial areas could the data be used? (i.e., safety, mobility, etc.)
Post-project re-use of data	• Which partner is responsible for maintaining the data after the project? • Shall a non-partner be the provider of the project data after the project? • Which application procedure shall be used? • Who will grant access to the data after the project? • Are there conditions, such as legal and ethical constraints and availability of funding for data storage and access services, to be considered? • Are there time limits after which the data need to be deleted?
Post-project financing	• How will the storage and support services for data re-use be financed after the project? • Known or to be decided? • How will this funding be distributed?

3.3 Data management plan

To ensure coherent and up-to-date data documentation and management, the creation of a Data Management Plan (DMP) is recommended. The function of a DMP is to collect relevant information about data and its handling throughout the project. While agreements may

document some aspects regulating data use and handling throughout a project, a DMP collects relevant information into one document, and also allows for a more detailed description of practical aspects of data management. A DMP can be updated as the situation changes. The use of existing DMP templates is encouraged, such as the European Commission's Horizon Europe DMP template[6]. If a funding agency is involved, a DMP may be recommended or even required.

Data Management Plans are instrumental in facilitating the adoption of the FAIR (Findable, Accessible, Interoperable, and Reusable) principles. While these principles offer a high-level framework, their implementation can vary. The Horizon Europe DMP template helps translate these abstract principles into concrete action by posing specific, relevant questions.

Example Questions from the Horizon Europe DMP Template:

- Making data findable: Will rich metadata be provided to allow discovery?
- Making data accessible: Will the data be deposited in a trusted repository?
- Making data interoperable: What data and metadata vocabularies, standards, formats or methodologies will you follow to make your data interoperable to allow data exchange and re-use within and across disciplines?
- Increase data re-use: How will you provide documentation needed to validate data analysis and facilitate data re-use (e.g., readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, units of measurement, etc.)?

Questions about data collection, personal and sensitive information, data ownership, rights, and access may have been dealt with in other project-related documents, but not all. Other questions, about provenance and metadata standards, are closely linked to the next chapter.

Additionally, we recommend using Science Europe's Practical Guide to the International Alignment of Research Data Management – Extended Edition (Science Europe, 2021) as a guide for Data Management Plans.

3.4 Participant agreements including consent forms

The participant agreement (sometimes described as consent forms) detail the objectives of the research project to the participant and outlines the commitments required on the part of both the project and the participant. Participants should be made aware of their legal rights regarding their participation and their data. The participant agreement must clarify these issues to enable the participant to make an informed decision about whether to take part or not. Information can be provided within an annexure to the agreement, or often these are provided by way of a separate document known as an information sheet. In this case the consent form will then refer to the information sheet. A person can only give informed consent if they understand the nature of the research and what is being asked of them. Research involving human participants can involve asking personal information and personal

6 https://enspire.science/wp-content/uploads/2021/09/Horizon-Europe-Data-Management-Plan-Template.pdf

opinions, or asking participants to be actively involved in a research activity such as driving a vehicle. It is important to be clear on the purpose for which their data will be used during and after the project, and other rights which they have such as the right to withdraw.

From a data-sharing standpoint, participant agreements or information sheets should cover the topics included in Table 2:

Table 2: Data-sharing topics in participant consent/agreements

Topic	Issues to be considered
Data collection	What data is to be collected, being specific about audio and video or any personal/sensitive data
Data storage	How, where and for what length of time will the data be stored
Governance and administration	Who is responsible for the data? Who do the participants contact to take action in respect of their data? How can they withdraw?
Access	Who will have access to the data
Data use	What will the data be used for including areas of future research
Security and data protection	Procedures for anonymisation and pseudonymisation

3.5 Data provider agreements

Data providers could be companies providing sensor systems, map data, weather data or other services that the project needs to enhance the dataset. The organisations may or may not be part of the project.

Data provider agreements are a contract between a data provider and a data consumer (in this case, the project consortium) that reflects the terms and conditions of the data provision. These agreements are required to ensure both parties understand their rights and obligations and may include provisions related to data ownership, specification of permitted uses of data, aspects of privacy protection, confidentiality, or other technical specifications (e.g., including data formats or licenses). It is important to be aware of topics that can affect future research due to possible restrictions in data use, as well as complications that may arise should the data provider change ownership or cease trading.

The aspects outlined in Table 3 should be considered when considering contracting data from third parties:

Table 3: Data provider agreements

Topic	Issues to be considered
Data collection & storage	What data is to be collected? What, where and for how long data will be stored?

Data access & use	What methods and procedures are used to access the data and how will data be transferred or shared? What will the data be used for including areas of future research? Are there restrictions on the data use or special conditions for using the data after the project?
Data governance	Who is responsible for the transfer of the data? Who owns the data? Who has access to the data (project partners/third parties)? What are the provisions dealing with the ownership of a third-party (e.g., in case of notable changes in ownership (like merges), or even if a business is shut down)?
Legal	Which is the applicable jurisdiction in the event of a dispute? How is notification of financial and other conflicts of interest handled?
Data privacy, security and protection	What data is confidential? Is there any personal data? What measures are implemented to ensure data security? How is the right to withdraw implemented (incl. destruction of data and associated procedures)? How can data providers keep adequate records, and supporting documents relating to the data subjects?
Visibility	Does any output from the dataset refer to any identifier of the dataset, and/or the Grant or funding statement?
Ethics and classified information	Sensitive information with security recommendations must comply with additional requirements imposed by the funding authority.

3.6 Data space actor agreement

Where project data is to be shared or accessed via a Federated Data Space (FDS), the terms for contributing to the FDS and the terms under which parties may access data should be clarified in an agreement.

Data agreements and their management require the definition and identification of user roles under a data sharing framework, such as the European Gaia-X federated secure data infrastructure. Participants in Gaia-X are categorized as:

- Provider: operates resources in Gaia-X and offers them as services, defining the service offering including terms and conditions and technical policies.
- Consumer: searches service offerings and consumes service instances in Gaia-X to facilitate digital offerings for end-users.

- Federator: in charge of the federation services and the federation itself. Federators enable the interaction between providers and consumers.

In addition, the Gaia-X Trust Framework establishes a minimum baseline to participate in Gaia-X ecosystems. This baseline is defined as a set of rules about common governance and interoperability across individual ecosystems while allowing users to have full control over their activities.

Nevertheless, in its current form (GAIA-X, 2023), the Gaia-X Architecture Document present a generic model for automated contracts, but the realisation of these contracts is not within its scope. Existing examples on automated contracts management are the Cloud Adoption Framework by Microsoft (Data contracts, 2022).

The topics which may be covered within a data space actor agreement are similar to those outlined above in other types of project agreements. However, the execution of those topics is likely to be quite different, as a FDS does not require the physical transfer of data, nor does it rely on a centralised repository. Consequently, the prerequisites for a data space actor emphasise suitable infrastructure, advanced data-sharing interfaces and authentication methods, and a high standard of security and data protection protocols.

Table 4 below outlines the considerations for a data space actor agreement.

Table 4: Data space actor agreement

Topic	Issues to be considered
Purpose of data sharing	In a project as Provider: What is the purpose of sharing the dataset? In a project as Consumer: How will datasets will be used?
Scope of data sharing	In a project as Provider: Which datasets will be shared? In a project as Consumer: Who will have access?
Data Ownership	In a project as Provider: Who is responsible for ensuring the data's accuracy, citation guidelines? In a project as Consumer: How can a data provider be acknowledged in work, on which the datasets are based?
Intellectual Property Rights	In a project as Provider: Are there any IPR matters in the dataset and does this impose any restrictions on its use?
Liability	In a project as Provider or Consumer: Who is responsible in the case of a data leak or unauthorised misuse?
Data security and privacy	Project as Provider or Consumer: Ensure the data is secure from misuse and that data is protected in accordance with relevant data protection laws. Ensure adequate infrastructure is in place.
Governance	Project as Provider or Consumer: Define which jurisdiction will govern any dispute and any regulations on data protection.

4 Data and metadata descriptions

CCAM projects and studies can collect large amounts of raw data, especially when continuous data-logging is favoured over event-based data collection. Moreover, often this data is merged with data from other (external) sources. In general, these studies also generate considerable amounts of derived data.

Derived data can manifest in various forms, each tailored to specific requirements. For example, this data might closely resemble the original raw data but presented in a different format. Examples of this are in-vehicle signal values decoded from raw CAN frames, or a subset of raw data from selected, shorter driving scenarios, being stored in a database.

Alternatively, derived data may involve refined versions of raw measurements – cleaned, filtered, and perhaps discretized. Different data sources can be combined into a new dataset, for example by combining weather or GPS data to sensor data. They can be a derived measure, where several pieces of information have been combined to compute a new, more directly interpretable measure (e.g., time headway is the distance to the forward vehicle divided by speed; traffic density is calculated from traffic volume and speed).

Data streamed in a continuous flow can also be transmitted and captured, in point-to-point communications or broadcasted. Such data streaming can produce significant volume of data which may often be consumed on the fly. They but also made persistent into storage systems for further analysis.

Lastly, they can be aggregated data, including aggregated time-series data obtained using a data-reduction process, in which the most important aspects of the dataset have been summarised. The summarised data generally consist of a list of relevant events or driving situations and their associated attributes, the result of a mix of algorithm and annotation-based processes.

Depending on the aims and methodology, simply re-using data in their most transformed/aggregated form may be sufficient. Occasionally, and when not prevented by intellectual property agreements (e.g., in the case of CAN data provided by vehicle manufacturers), it may be necessary to go back to the original, raw form. In most cases, however, cleaned-up, derived, annotated data will be the most useful.

Whichever form of data is used, the core of data sharing is that the data provided are valid, or at least documented to a level where an assessment of the level of validity can be performed. This is potentially problematic if the data re-user was not part of the project and does not know in detail how the tests were performed, which sensor/version was used or how the data were processed from the raw data. The main problem is usually that the data are insufficiently described.

Data re-use requires precise knowledge about the data. Therefore, it is vital to have extensive and high-quality metadata (see definition below), providing the following information:

- the purpose and context of the data (basic project information, a description of the data, purpose of collection, responsible and contact persons or organisations)
- the provenance: the conditions in which data has been collected, how data have been stores, cleaned up, processed and aggregated

- how they can be accessed: conditions for and method of access
- usage restrictions and licence information

A well-documented dataset inspires trust when being used and reduces the risk of less confident conclusions – something that all stakeholders benefit from.

In addition, before researchers/analysts/business developers even start to use a dataset, it must be identified as potentially interesting and then selected as relevant for their purpose. These first steps only require a subset of the aforementioned documentation, which gives an overview sufficient to compare several datasets but is compact enough to ensure efficiency both in terms of creation and consultation. This results in the choice of items to be documented in a data catalogue.

The aim of this chapter is to address these issues and provide methods for efficiently describing a dataset and the associated metadata. It suggests good practices for documenting a data collection and datasets in a structured way.

4.1 Definitions

This document defines **data** as '**any facts, statistics or digital material whose value might be used during analysis and impact its result**'. Data can have format to enable its reading or storage, but on its own it may not have meaning.

The ISO/OECD definition (ISO 20546, 2019) is similar where 'data is the reinterpretable representation of information in a formalized manner suitable for communication, interpretation, or processing'.

Information, however, refers to processed, organized or structured data, or data that has been given a semantic load, or a value under a certain business domain. Also information has a definition by ISO/OECD: 'Information refers to knowledge concerning objects, such as facts, events, things, processes or ideas, including concepts, that within a certain context has a particular meaning'.

A **data element** is an atomic unit of data that has precise meaning or precise semantics (e.g., represented as a column in a table or an array of values).

A categorization of data is proposed in the following chapter.

Metadata is data that defines and describes other data. This document defines metadata as **'**data that describe properties of other data, in the form of information about its origin, purpose, version, type of content, or any other information that may simplify data storage, consumption or processing'.

This document presents four different categories of CCAM metadata, each providing a different kind of information about the data. These are described below and in Figure 3:

1. *Study design and execution* documentation, which corresponds to a high-level description of the data collection - its initial objectives and how they were met, description of the test site, etc.
2. *Descriptive* metadata, which precisely describe each data element, including information about its origin and quality.
3. *Structural* metadata, which describe how the data are organized.

4. *Administrative* metadata, which set the conditions for accessing the data and how access is to be implemented.

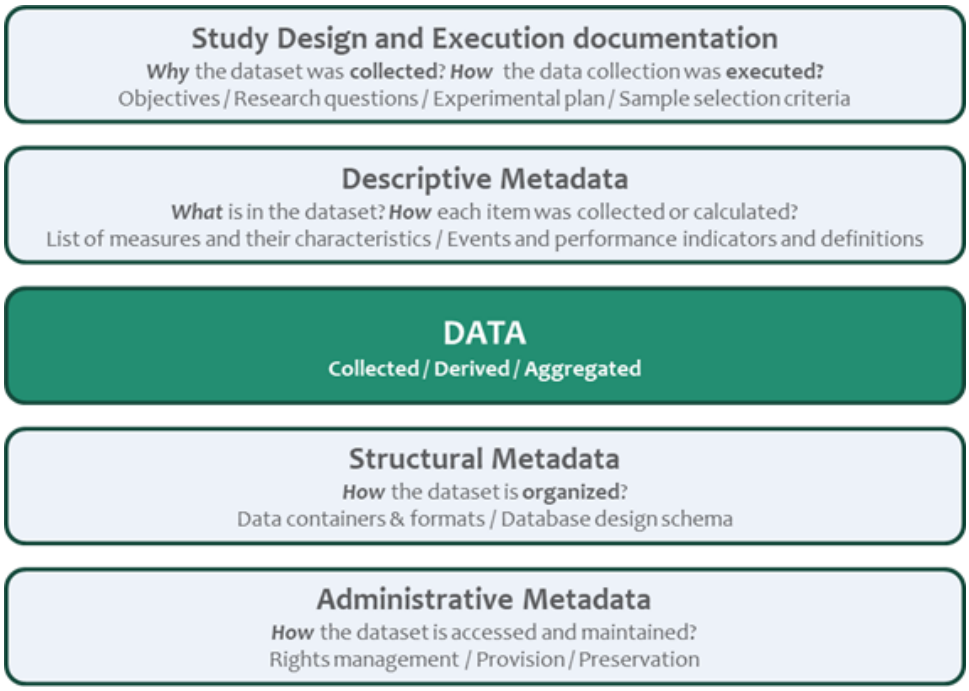

Figure 3: Types of metadata in relation to data

In the following sections, data categories are introduced and classified, including examples and recommendations for which data should be systematically collected (see 4.2). The four metadata classes are more precisely defined, and what should be documented (according to good practice) for each of them is described (see 4.3).

The recommendations are based on best practices from NDS/FOT projects, as well as CCAM projects, and are applicable to other types of studies and frameworks where data is used.

4.2 Data categories

Data can take many successive forms, from raw collected data to very high-level aggregated data, with many steps in between. A dataset is not only the result of data collection, but also of an iterative process, comprising pre-processing, integration of different data sources, calculation of derived measures and manual and/or automatic data reduction. Aggregated data are usually the easiest to use but may only be suitable for analysing research questions similar to the initial study. In contrast, raw data can meet a larger variety of needs, but usually requires a deep technical understanding of the data collection process and sufficient data storage and operational capacity to be used in a relevant and efficient way. A trade-off, using intermediary states of the data, generally must be found, illustrated in Figure 4.

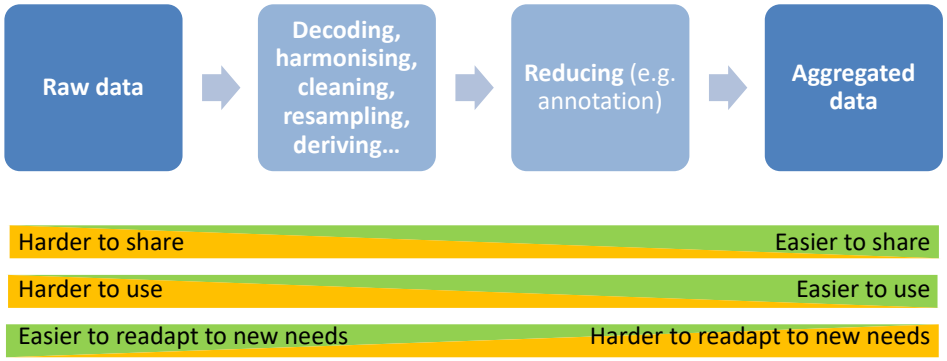

Figure 4: The trade-off between usability, usefulness, and availability

As a result, a data re-use case will typically require a combination of very different forms of data. This document proposes a way to classify them, based on two characteristics: the relations between the different entities (vehicles, users or operators in an AD setup, infrastructure, etc.) addressed during data collection, and the information which typically captures the entities' different aspects (measures). This classification system is as close as possible, but also complementary, to the definitions in FESTA (FESTA, 2021), which essentially relate to data collection and analysis. This system emphasizes the typical structure of a CCAM dataset and contains the following main categories: context data, acquired and derived data, and aggregated data (see Figure 5). The sub-categories are described further in the following sections. Each category may contain either objective data (which is normally quantitative data), subjective data (which can be either qualitative or quantitative data), or a mix of both.

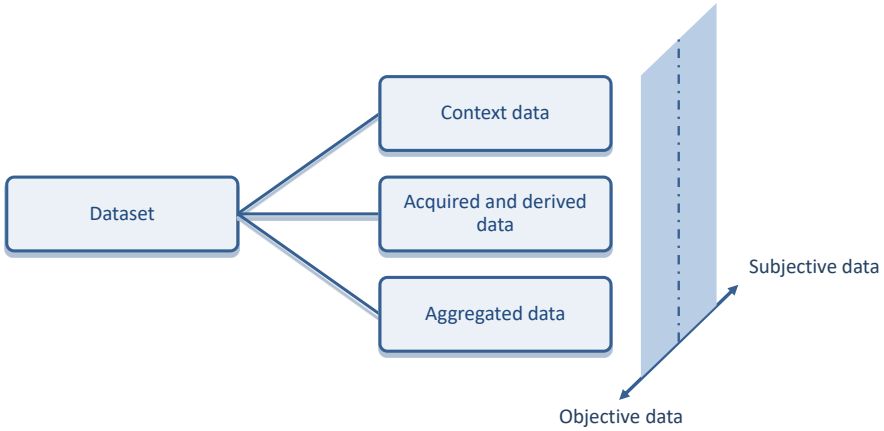

Figure 5: Dataset categories

Objective data are collected through direct physical measurement, without any influence from the experimenter or the participant's subjective impression. They are collected using sensors, which can be pre-existing or installed on purpose, and data acquisition systems, which can be installed inside vehicles or on the roadside.

Subjective data are provided by the participants or observers, based on their impressions, feelings, memories or opinions – collected (for example) by questionnaires, travel diaries (usually quantitative data) or interviews and focus groups (qualitative data).

This categorization will be used as a basis for recommendations regarding what should be recorded in a study, and how the corresponding metadata should be created.

In each sub-chapter of 4.2.x, different sub-categories of data are described in a tree-structure. The parent category is indicated to understand the relation between the different sub-categories.

4.2.1 Context data

Context data correspond to all information which doesn't change during the study but helps explain the observations or document their values. They may be directly collected, generated for the purpose of the experiment, already exist, or retrieved from external data sources.

They contain, for instance, background information – such as infrastructure characteristics (e.g., *map data*) and vehicle /driver characteristics and roles during automated driving, including questionnaire results.

Questionnaires collect qualitative and quantitative data reported by each individual participant. They typically cover basic data, such as age, gender, and general attitudes about driving. They can also cover more specific aspects, such as personality traits (e.g., sensation-seeking, introverted). Quantitative data is obtained by means of closed questions (e.g., multiple choice, scales) whereas qualitative data is obtained when specific questions are open for rich text information, often of a more interpretive nature.

There is a grey zone where some elements may be considered 'contextual' data, such as participants' characteristics or weather, traffic and driving conditions, and thus part of the dataset, whereas in other datasets this may be considered metadata.

4.2.2 Acquired or derived data

Acquired data are all data collected during the study for the sole purpose of the analysis.

Derived data are obtained by transforming raw data into more directly usable data through, for instance: data fusion, filtering, classification, and reduction. They typically contain derived measures (such as, for instance, time-headway, which derives from both longitudinal speed and headway), and performance indicators (PI), referring to time- and location-based segments such as particular events.

In most cases, transforming acquired measures into derived measures during pre-processing or processing doesn't change their nature, established that no information loss occurred. For instance, an acceleration low-pass-filtered to remove noise doesn't cease to be a vehicle-dynamics measure; the depression of a pedal converted to a discrete *pressed/not-pressed* state doesn't cease to be a driver-action measure. As a result, in most cases, the subclasses presented below apply to both acquired and derived data.

However, in some cases, several kinds of measures are combined to form new, more interpretable measures, which can't be categorized simply. For instance, speed and acceleration from several vehicles can be combined to form a time-to-collision variable.

This category includes both objective data, in the form of measures from sensors (referred to as sensor data in FESTA), and subjective data, collected from either the participants (referred to as self-reported measures in FESTA) or analysts. Subjective data can be as varied as time-history data, subjective classification of time segments, or rich-text information

from travel diaries, interviews, and focus group discussions. Questionnaires can also be seen as acquired data when collected periodically during the project (compared to the static questionnaire data described in 4.2.1).

In this section a structure of the sub-classes is described below and presented in Figure 6.

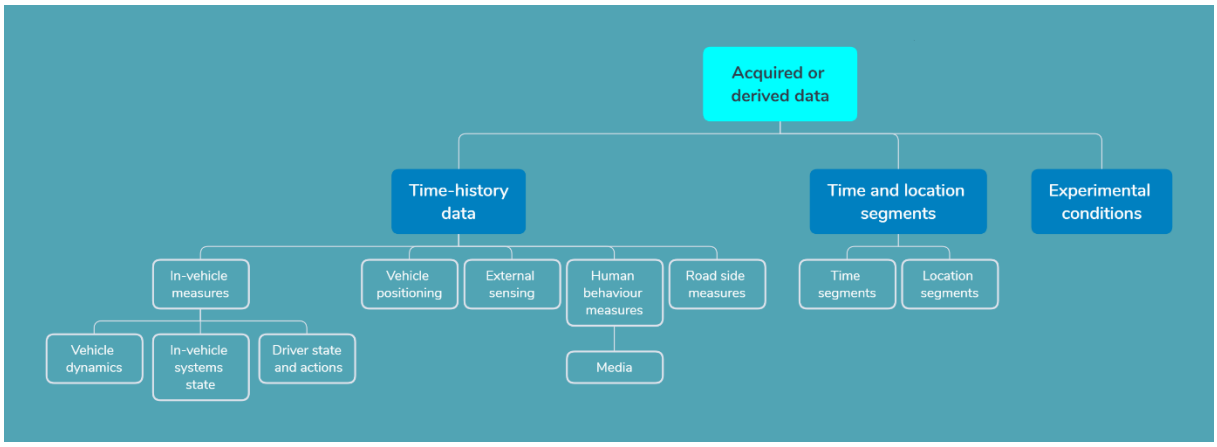

Figure 6: Subclasses of Acquired or derived data

Time-history data (parent: Acquired or derived data)

Time-history data describe the history of a measurement over time. Time-history data can be collected with a specific measurement frequency, or when triggered by an event, typically a value change.

Time-history data may consist of the variation over time of single physical values (e.g., speed), a collection of physical values (e.g., 3-axis acceleration) or more complex media, such as sound or video.

Time-history data can either be collected from the vehicle perspective, by means of (for instance) an instrumented vehicle, smart device application or travel diary, or from the infrastructure perspective, by means of roadside measurements. They can be historical or real-time observations, or by measurements done in a mobile phone app from persons inside or outside of a vehicle.

Time-history data consist of both direct measures, i.e., raw data measured over time, or derived measures, after any kind of transformation (such as resampling, offset correction, filtering, and removal of incorrect values) has been performed.

In-vehicle measures (Time-history data)

Instrumenting vehicles enables the collection of vast amounts of data, using either original sensors (tapping their communication networks, such as CAN) or additional sensors. Applications on smart devices (i.e., smartphones) can also collect important information in the following categories, and the data they collect can basically be treated the same way as the data from instrumented vehicles.

Vehicle dynamics (In-vehicle measures)

Vehicle-dynamics measurements describe the motion of the vehicle. Typical measurements are longitudinal speed, longitudinal and lateral acceleration, yaw rate and slip angle.

In-vehicle systems state (In-vehicle measures)

The state of in-vehicle systems can be accessed by connecting to the embedded controllers. The data category comprises continuous measures, like engine speed, or categorical values, like ADAS and active safety systems activation, and automation level.

It is important to document the system state when in baseline mode. As new vehicles have numerous active systems installed, there is a need to know the settings also for baseline data. The AD-level (SAE J3016, 2018) is important both for baseline and treatment data when comparing an autonomous function in level 4 in treatment and in level 1 for the baseline data.

Driver state and actions (In-vehicle measures)

In addition to variables describing driver actions which command the vehicle, like steering wheel angle, pedal activation or HMI button press, variables characterizing the physical and emotional state of the driver can also be measured. For instance, cameras and computer vision can measure driver position, detect engagement in a secondary task or detect eyelid closure (which highly correlates with alertness). In fully autonomous vehicles (e.g., a shuttle) the driver might be a remote operator having the responsibility to act in the case of problems that the AD-system cannot solve. The actions are of course important to record, in the vehicle but also in the remote operator system.

Vehicle positioning (Time-history data)

The geographical location of a vehicle is most frequently determined with global navigation satellite systems (GNSS) and the aforementioned advanced sensors. It can also be determined by information from the cell phone network, surrounding Wi-Fi networks, or a combination of these and GNSS.

External sensing (Time-history data)

A precise understanding of the environment can be obtained by advanced sensors like radars, LIDARs, cameras, and computer vision, or by simpler sensors (e.g., optical or temperature).

It is imperative to detect different objects around the (ego vehicle), including their position, heading, speed and size. Depending on the need of a project, the output of some of these sensors could be stored for assessing the ground truth. Raw high-resolution video or point-clouds can consume vast amount of storage.

Some signals could be logged from the in-vehicle data network. For instance, luminosity (indicating the presence of rain), characteristics and dynamics of the infrastructure (e.g., lane width, road curvature) and surrounding objects (e.g., type, relative distance, and speed) can all be measured from within a vehicle.

Human behaviour measures (Time-history data)

The actions of drivers can be measured from the vehicle perspective but also there is with the introduction of AD function also a need to view the driver as a passenger and also to understand the whereabouts of other passengers. In some cases, you can collect objective data from sensors in the vehicle or devices attached to a mobile app (e.g., a device measuring heart rate and transferring this to a mobile app).

Complementary to sensors and instrumentation, some continuous measures can also be built through the perception of analysts or annotators using video data. Eye glance and driver state (e.g., drowsy, impaired, angry) can be evaluated manually by analysing video from driver-face-oriented cameras. This is even more important in the context of autonomous vehicles, where rich data from video could automatically be converted into individual measures of interest (e.g., hands on steering wheel, head and body pose).

Media (Human behaviour measures)

Media data are usually video, but in some data-collection projects audio is recorded. The image or video data are used to give a ground truth of the persons inside the vehicle, to understand their behaviour and response to external factors.

Roadside measures (Time-history data)

Roadside measures comprise vehicle counting, speed measurement and positioning – using radar, LIDAR or simpler rangefinders, video-based counting, inductive loops or pressure hose.

In the case of ITS systems, they may also contain more complex information remotely transferred from vehicles to or from roadside units. These messages are given more attention in an autonomous driving scenario where vehicles could communicate with each other to warn or inform. One challenge is to have a common time reference on timestamps both when the message was transmitted and received at the sender and receiver.

Media data (typically video – for instance in traffic conflict observations) are also often collected from beside or above the roadside. Roadside measures are evolving rapidly, with data being collected by drones or open-data services, for example.

Experimental conditions (Acquired or derived data)

Experimental conditions are the external factors which may have an impact on participants' behaviour. They may be directly collected during the experiment or integrated from external sources. Typical examples are traffic density and weather conditions. Controlled factors, such as the ability to use a system, also need to be included in the dataset, depending on which phase of the experimental plan a participant is currently participating in.

Time and location segments (Acquired or derived data)

For the purpose of the analysis, it can be relevant to analyse the data aggregated for a delimited period in time or space (such as journeys, certain events as defined in FESTA or e.g. road segments). These data segments are defined by a combination of specific conditions and characterised by specific attributes, some which are automatically computed, and some which are manually annotated from video. The attributes mostly consist of situational variables and/or PI, depending on the studied phenomena and its expected contributing factors; they can also consist of links to other segments or contextual data. For instance, each trip might link to a specific driver and vehicle, each of which have their own characteristics. Finally, the segments might serve as a container for time-history data: a trip can contain the history of the vehicle speed and an event may contain successive eye-glance values, manually coded by an annotator. As a result, the segments contain a large amount of initial data, which is structured, reduced, and summarized into more manageable tables, suitable for data analysis.

The creation of the segments can either be automated (i.e., they are created in response to a specific value or threshold of one or more variables), manual (when a specific event is observed on a video), or a combination of both (e.g., automatic detection of candidate events, accepted or rejected in video annotation). In the same way, attribute values can either be automatically computed (i.e., the mean or maximum value of a measure during a time segment) or manually annotated, typically from video. In the latter case, standardized annotation schemas are used to enrich data with information available from video recording. Annotation variables are thus a subjective assessment of the situation by an analyst or annotator. They can be quantitative, using single or multiple choices (i.e., present/not present or level of rain); they can consist of specific time stamps, for example when the driver is first aware of a hazard; or they can be qualitative narratives, which describe a specific event or situation.

Finally, subjective, participant-reported data can be collected as certain kinds of segments, such as self-declared events, or they can populate some segment attributes (e.g., travel diaries that contribute some characteristics to trips).

Time segments (Time and location segments)

Time segments are the most common type of segments, collected and/or generated during data reduction. They correspond to a time period when some specific conditions are met. Depending on the kind of conditions which define them, their typical duration, and the researcher's own vocabulary, they are identified as trips (a vehicle is started, driven for a period of time by a driver, then stopped), events (typically a short period of time with very particular characteristics), situations or chunks (division of the complete dataset into segments of comparable size according to a combination of situational variables, characterized with PI).

Locations (Time and location segments)

While time segments take the perspective of a driver in a vehicle during a trip, locations take the perspective of a place, where multiple trips might pass through. Roadside observations will typically generate locations, and typical location attributes are vehicle counts or speed measurements, which can also be associated with the infrastructure attributes. Furthermore, using geographical information systems (GIS), data from in-vehicle collection can also be projected over a geographical reference system to characterize, for instance, one or several participating drivers' behaviour at a specific location such as an intersection.

Just like trips, locations can be divided into smaller segments. This could be parts of a road stretch, especially when a fixed trajectory is followed. The segments can be used individually or be linked in a chain.

4.2.3 Streaming data

In the context of ADAS or CCAM testing, some data may be transmitted continuously and in real-time over a network of the internet, from one source node into a recipient node (peer-to-peer), or even from multiple sources into several recipients (broadcasting).

Streaming data refers to real-time data continuously transferred from a data source and processed as soon as it arrives its destination. A streaming data architecture focuses on technologies that allow to process data in motion, such us with extract-transform-load (ETL) batch processing paradigms.

Streaming data may be non-persistent, i.e., volatile, or generated, consumed, and destroyed without entering into any serialization or persistent process. However, it can also be made persistent, if captured and stored into storage systems by the processing node, and as such is subject to consideration under "Acquired *or derived data*".

Functions and use cases using short range communication protocols, are especially susceptible to generate streaming data. It may be produced sparsely (e.g., at vehicles, or other RSU), and streamed through communication and networking architectures into recipient nodes at the edge or cloud. Examples of such data streams include (but are not limited to):

- V2V, V2X messages (e.g., ISO CAM, DENM, CPM messages)
- sensor data (e.g., video streaming, point cloud, vehicle data)
- traffic data (e.g., road occupant statistics).

Video data streaming might need special treatment as it often demands higher bandwidth and lower latency networking to ensure images are delivered in real-time for its consumption at each specific use case. The main technical difference with other data flows is that video needs to flow continuously compared to other data types that might be buffered and processed in batch processing steps that collect data trunks to be processed as a group at some future time. Consequently, it is important to distinguish between streaming data use cases that require video (or equivalent data, such as point clouds), versus lightweight data streaming such as V2X messages or other metadata.

Therefore, streamed data needs to be taken into consideration in the context of CCAM project, as it may play a key role during the computation of KPIs, aggregation of information from multiple sources, or to provide provenance and traceability mechanisms.

Processing of data streams implies the node can produce updated responses observing only recent data, without the need to have the entire history of received data, which may be discarded once consumed. Such behaviour might be needed for a number of reasons: excessive data volume, high-frequency, or due to the presence of potentially sensitive or private information (e.g., in federated learning frameworks).

Depending on the nature of the data (e.g., light-weight messages, heavy raw sensor data, etc.), treatment and format of streaming data may be different, and the requirements of format might differ as well.

In general, data streaming implies sending data in small packets, rather than in large blocks, which allows for faster and more efficient data transmission, robustness against errors, etc. Data streaming requires a protocol stack that defines how to interpret binary data, de-codification methods, add necessary redundancy and headers at application level, and handle other technical aspects of data transmission. Some examples of technologies often used at transmission level include TCP (Transmission Control Protocol) and UDP (User Datagram Protocol). TCP is a reliable protocol that ensures that data is transmitted accurately and in the correct order. It uses a three-way handshake to establish a connection between the sender and receiver and includes error-checking and flow control mechanisms. UDP, on the other hand, is a faster protocol that does not include the same level of reliability

as TCP. It is often used for applications that require fast data transmission, such as online gaming or streaming video. Other protocols are specifically designed for controlling the delivery of real-time data, such as audio or video, over a network. For instance, RTSP (Real-time Streaming Protocol) works by establishing a connection between a client and a server, and then sending requests and responses between the two to control the streaming session. RTSP is often used with other protocols, such as RTP (Real-time Transport Protocol) and RTCP (Real-time Transport Control Protocol) to cover the entire set of requirements of effective real-time data transmission.

There are several other real-time protocols that are used in applications. One example is ITS-G5, which is a communication stack specifically designed for Intelligent Transportation Systems (ITS) applications. ITS-G5 includes a variety of protocols, including IEEE 802.11p for wireless communication. WebRTC (Web Real-Time Communication) is another example of a real-time data streaming protocol. It is an open-source protocol that enables real-time communication between web browsers and other devices, such as smartphones and tablets. WebRTC uses several technologies, including audio and video codecs, NAT (Network Address Translation) traversal techniques, and SRTP (Secure Real-Time Transport Protocol) encryption, to facilitate real-time communication over the internet.

4.2.4 Aggregated data

Using relations between segments, reduced data (e.g., segment attributes) are typically aggregated into smaller, more usable tables, suitable for data analysis or data interpretation. For instance, driver characteristics can be grouped together with attributes from one type of situation, to evaluate the impact of drivers' characteristics on their behaviour in that situation. Reducing the resolution or down-sampling the frequency of time series data can omit sensitive and confidential details about the system while still capturing key patterns and trends of the original data.

The data resulting from aggregating different kinds of reduced data together are called aggregated data. Although they are generally linked to a specific research question, the aggregated data may be re-used with different statistical methods, or re-aggregated with other data, to quickly answer new questions without the need to go back to harder-to-use, raw data.

As they don't contain instantaneous values, aggregated data don't allow potentially problematic re-use, such as pinpointing illegal behaviour from one specific driver or benchmarking a driving assistance system without authorization from its supplier. As a result, aggregated data are generally easier to share than other categories of data.

4.3 Metadata

Metadata refers to information that describes and provides context for a dataset, and which can take on different forms depending on the type of information being conveyed. Metadata is a central element of digital curation, facilitating findability, accessibility, interoperability and reusability of data. Metadata can take different forms, such as a structured metadata file packaged with data or structured metadata in a data catalogue.

In this document, metadata includes the concept of documentation, which is sometimes defined as a separate entity. Here, documentation requires a structured approach and

therefore is included in the metadata concept. This is a common approach in several fields, such as Geoscience.

Metadata can be categorized for ease of use:

- Descriptive metadata, describing the content of a dataset, is perhaps the most useful type for finding and selecting relevant data, as well as for analysis.
- In contrast, structural metadata are the prerequisite that helps the analyst understand the structure of the dataset, by describing 'data about the containers of data' (Roebuck K., 2012).
- Administrative metadata are collected for the effective operation and management of data storage.
- Finally, the study documentation provides an overall description of how the study was performed.

Relevant metadata can be created according to predefined metadata standards, which among other things describe the structure and content of metadata through a metadata schema – a structured collection of relevant predefined metadata elements, relationships and terminologies. Using an existing schema allows for easier transfer of relevant information, since fields and terminology can be predefined, and semantic issues minimized. Metadata standards can be generic or specialized and can build on, or be mapped to, one another to make communication across standards easier.

Although this document does not describe a fully-fledged metadata standard, a set of metadata elements is suggested for use in data creation and sharing activities. In the future, this profile can be schematized mapped to existing metadata standards to increase metadata interoperability.

4.3.1 Descriptive metadata

Descriptive metadata includes information needed to understand the contents of a dataset well enough to make an informed decision on whether to look closer at it. The purpose is to describe the dataset and build trust in it – by providing not only the characteristics of each measure or component, but also information about how the data were generated and collected.

Descriptive metadata shall preferably be available close to the actual data to facilitate analysis. The descriptive metadata need to define the dataset and include detailed descriptions of measures, PI, time and location segments and their associated values. In addition, external data sources, subjective data from self-reported measures and situational data from video coding must be described in detail. Not only must the output of the data be described, but how the data were generated and processed is equally important; this is where one can build trust in the dataset. The more thoroughly the origin of a measure is described, the greater the trust. The proposed structure of descriptive metadata follows the data categories in 4.2.

Context data description

The level of detail when describing contextual data can vary. Information about drivers and vehicles is often obvious from the name of the variable (e.g., gender and age for participants,

and model, brand, and year for vehicles). Other information, such as questionnaire data acquired from participants, might need a more in-depth description (e.g., a definition of the self-assessed sensation-seeking measure).

As databases within this domain often consist of a variety of different external data sources, it is very important to document them all to get a full picture of the data. The external data sources can include static contextual data from map databases or dynamic data from weather services and traffic management services. In these cases, a more in-depth description is needed where it is important to describe the origin of the data, the methods used to match the different datasets (e.g., a description of the map-matching algorithm), and each output variable.

Some additional data might be merged with the acquired data (e.g., map attributes or weather codes). These data are described in their respective sections below.

Acquired or derived data description

A description of every measure in a dataset is mandatory, making the data re-usable for future analysis. The origin of the data and the processing steps performed are equally important for drawing correct conclusions in the analysis.

It is important to include definitions of time and location segments in descriptive metadata, as the definitions vary between different datasets depending on the purpose of selected segments. The segments need to be defined (i.e., how the segment start and stop times are calculated), and so do the associated attributes (e.g., summaries, situational variables, and PI). The different types of time and location segments are often important products of the dataset, providing easy-to-use references to the actual data.

This section also includes a suggestion for describing PI and summaries – data which are often attached to time or location segments but may also be used independently of them.

Direct or derived measures in time-history data description

The description of direct measures is often beyond the project's control and needs to be requested from the supplier of the equipment generating the data. If the data are acquired from the CAN bus of a vehicle, the OEM can supply information which describes the data. Understanding the origin and full history of direct-measure data is important, but often overlooked. To get access to this information, the use and restrictions of direct-measure metadata should be included in the contracts and NDAs with the suppliers. The origin of the measure should at a minimum include where the data were generated (e.g., sensors, ECU) and acquired (e.g., CAN or other equipment/channels), the frequency, the units, the range, the resolution, whether they were derived from other data and error codes.

When direct measures are being processed into derived measures, it is important to document all the data processing steps. Derived measures are often processed several times, and the final product might consist of more than one measure. The need for a detailed description is crucial for creating trust for data re-use.

The output of the data processing must be documented and include information on data precision, unit and sample rate. This metadata must also include information about how the data were processed (e.g., synchronization policies, re-sampling filters, harmonization rules). In an ideal scenario, an analyst performing an analysis can quickly understand not only the

meaning of the measure, but also its origin and history, and use this information to interpret the results.

Proper naming conventions for all data containers can go a long way towards helping interpret data's origin and understanding how it can be used. Tags describing the data type and origin can, for instance, be used. However, naming conventions are always a trade-off between comprehensiveness and legibility, and although necessary, are not sufficient for the proper documentation of a dataset.

Preferably all information in Table 5 should be included for each major data-processing step. As an example, interpolation filters must be documented in detail, so that the analyst can understand whether the measure can be used for a specific research question. Additionally, the tolerance for missing data (e.g., the number of frames or seconds) and how these values are stored should also be described in the metadata, because the values are often managed differently in different data formats (e.g., *NaN* in MATLAB, but *NULL* in Java and relational databases). Describing the measure in detail avoids misinterpretation.

Table 5: Metadata attributes for time-history data measures

Element	Instruction/example
Data precision	What is the data precision of the measure? The terminology is derived from database technologies where the precision is the maximum allowed number of digits (either the maximum length for the data type or the specified length). If not specified, the data type will define the maximum allowed precision. When measuring the signal this is the *resolution*. This information, as well as the precision and accuracy of the measurement, should be provided in the *origin* section below.
Unit	What is the unit of the measure (e.g., m/s, RPM or if an enumeration)?
Sample rate	What is the current frequency of the measure (e.g., speed resampled at 10 Hz or 1 Hz)?
Filter	Which filters were applied (e.g., low-pass, interpolation or outlier filters)? This could also include the maximum allowed time during loss of signal data for the filter to be applied. The value can be very different depending on the measure (e.g. interpolation might be implemented on the speed signal unless the next available sample is less than two seconds later).
Origin	How was the measure generated and from what data source? This includes information about precision, accuracy and resolution of the measurement. For instance, it is important to know if the speed measures originated from CAN at 20 Hz or GPS at 1 Hz. It is also important to know how precise and accurate the measurement was done, as well as the resolution of measuring device and the logger system

	translating the signal. This could also refer to another described measure.
Type	Is the measure an integer, float, string or picture file?
Range	What is the expected range (minimum and maximum values) of the measure?
Error codes	Which values trigger error codes? What is a null value? It is also important to describe how the errors are managed.
Quality	Are there any quality measures related to this measure and how are they defined? The quality could be set on a per-trip, per-measure or even per-sample level (e.g., for GNSS data: HDOP, number of satellites).
Offset	Is there a known offset of the measure? The information is related to the actual measurement and data logger. If an offset is known this should be included in the metadata of the measure.
Enumeration specification	Can enumerations be translated into readable values (e.g., 1 means ‘left’ and 2 means ‘right’ for the turn indicator)?
Availability	Can the measure be shared? What are the conditions to access it?

Time segment data description

Calculated time segments or triggered events represent singularities over time, which may be as short as a single time instance, or longer based on a specific set of criteria. The definitions of time segments differ among datasets; the more common ones are trips, legs and events. This variation makes it even more important to describe the purpose and how the segments were designed, including their origins. It is also important to understand the conditions that define the start and stop of a time segment.

Events are often described by type, which explains why an event was triggered or threshold met. To understand the event properly, event type descriptions must include references to the measures and method used to calculate the event, as well as threshold values.

Different segments can have different associated PI, summaries or attributes, and these should also be described: for example, a trip record might include the duration, distance travelled, average speed, number of times passing intersections, or just the number of samples. Time segments should include the attributes in Table 6.

Table 6: Metadata attributes for time segments

Element	Instruction/example
Type	What is the purpose of the trigger (e.g., a hard braking event, swerving at high speeds, overtaking or entering an intersection)?
Definition	What is the definition of the time interval? How are the time series grouped? The output could be a single point, fixed or variable in time.
Origin	Which measures were used to create the entity? What was the overall principle of the data computation that generated the entity?
Unit	What is the unit of any output value (defined by type)?
Enumeration specification	Description of enumeration values.
Attribute, PI or summary specification	Time segments might have associated data that need description. It could be attributes, such as driver ID or duration. It could also be computed data, such as PI or summaries (e.g., distance travelled, number of intersections passed, average speed or the number of times a button was pressed). The definition of all PI and summaries associated with the object are described later in this chapter.
Availability	Can the segment be shared? What are the conditions for accessing it?

Location data description

In many studies the vehicle is not the main entity; rather it simply provides values for locations. Locations must be defined, usually by position or a set of positions. This could be an intersection, a sharp bend, the specific position of a roadside unit or a stretch of road (anything from a city street to a European highway). The definition is of great importance because of this great variance. As with time segments, the value of the locations is not only the encapsulation of time or position, but also the determination of associated attributes and the output of computations. The metadata attributes of location segments are presented in Table 7.

Table 7: Metadata attributes of locations

Element	Instruction/example
Type	What is the purpose of the location segment?

Definition	What is the definition of the location, in terms of position, scenario or equipment? Can locations be grouped or arranged in a hierarchy?
Attribute, PI or summary specification	Location segments might have associated data that need description. It could be attributes, such as number of exits at a roundabout. It could also be computed data, such as PI or summaries (e.g., number of vehicles passing or average speed). The definition of all PI and summaries associated with the object are described later in this chapter.

PI and summaries definitions

PIs are used to measure the performance of one or more measures, and are often associated with a specific analysis project, although some might be re-used for other purposes. Each implementation of a PI should therefore be described precisely; see metadata attributes in Table 8.

PIs as summary tables are pre-computed data, used to make the analysis more efficient. The summaries are stored as attributes, often with time or location segments as a base; the summaries could, for example, describe the mean speed of a trip or the number of passes through an intersection. Summaries are convenient in data reduction. They are especially useful in a larger dataset for excluding data not needed for the analysis.

Table 8: Metadata attributes of PI or summaries

Element	Instruction/example
Purpose	What is the purpose of the PI or summary?
Definition	Details about how the PI or summary was calculated and the denominator (e.g., per time interval, per distance or location),
Origin	Which measures were used to create the entity? What was the overall principle of the data computation generating the entity?
Unit	What is the unit of the output value?
Variability	What is the variability of the PI or summary?
Bias	Is there a known bias of the PI or summary?
Data precision	Details on the data type and the resolution of the output value.
Availability	Can the attribute be shared? What are the conditions for accessing it?

Description of video annotation codebook

Documenting the video annotation codebook is important for helping the person coding the data to understand the instructions, but also for defining enumerations (for incident severity there are the conditions that define a crash, near-crash, increased risk or normal driving). It is also important to document the process of coding the data, whether inter-rater reliability testing was conducted, and other important aspects of the persons coding the data; typically, this information is part of the project study design. For each measure (as part of the video annotation codebook) the recommendation is documented in Table 9.

Often the reduced data are coupled to time or location segments. Because it is important to know why those segments where selected for video coding, the reference must be documented.

Table 9: Metadata attributes of video annotation code book measures

Element	Instruction/example
Description	What is the purpose of the measure?
Instructions	In what way was this measure described to the person coding the data?
Type	What type of input is expected (single or multiple choice: e.g., present/not present or level of rain, continuous, free text or voice)?
Options	What are the possible alternatives (often coded as enumerations)? How reliable are the data expected to be?

Description of self-reported measures

Other subjective data include travel diaries, interviews, and documentation from focus groups. These data are often in rich text format and the data description should cover why, when and how the data were collected. Questionnaires acquired during the data collection period should also be described in this section. These data are very similar to video annotations and could be described by answering the questions in Table 10.

Table 10: Metadata attributes of self-reported data

Element	Instruction/example
Description	What is the purpose of the self-reported measure?
Instructions	In what way has this measure been described to the participants?
Type	What type of data is expected (single or multiple values, continuous, free text or voice)?
Options	Descriptions of possible alternatives (often coded as enumerations) and how non-answers should be handled.

Streaming data description

Streaming data is very dependent on source, purpose, and protocols and standards used, but a general recommendation provided below, which can be adapted depending on the context.

Table 11: Metadata attributes of streaming data

Element	Instruction/example
Description	What is the purpose of the streaming?
Type	What type and data format (standard) is expected?
Options	Descriptions of possible alternatives (often coded as enumerations) and how non-answers should be handled.

Aggregated data description

The shape of aggregated data can vary to such a degree that it is difficult to propose a structured format. Depending on the level of aggregation, the data could be described as time history measures or time segments. Also, in many cases the aggregated data are shared with the promise that the underlying data will not be revealed; the algorithms are not described in depth (to eliminate the risk of making raw data information available by means of reverse engineering), and only a high-level description is allowed. The trust in this data will be reduced and it is up to the recipient to judge if it is good enough for re-use. An appropriate set of metadata questions is proposed in Table 12.

Table 12: Metadata attributes of aggregated data

Element	Instruction/example
Description	What is the purpose of the aggregated data?
Definition	Which algorithms were applied to the underlying measures?
Origin	Which underlying measures were used to calculate the aggregated data?
Unit	What is the unit of the output value?
Variability	What is the variability of the data?
Bias	Is there a known bias of the data?

Data precision	Details on the data type and the resolution of the output value.

4.3.2 Structural metadata

In a typical study, different parts of the dataset will use different storage technology, such as file systems, SQL and Not-Only SQL databases.

Structural metadata are used to describe how the data are structured in relation to other data. Data are organized into a system (e.g., a database and/or file system), a structure or database schema and a data content format. The aim of structural metadata is to facilitate the initial phase of data re-use by providing the necessary documentation about how the data is organized. The description should include the file system, the file structure and how to interpret the contents of a data container. All components of the dataset need to be described.

Since data may be stored for a very long time, it also becomes important to describe and preserve tools that can read the data. This issue is highlighted when it comes to data archives. Even only five years after a project has ended, the knowledge about specific tools might have been lost and the cost of building up the competence again might exceed the data's value. It is therefore recommended that the tools, platform and prerequisites be described – in even more depth if using a non-standard data container, file format or file structure.

File system/Database

At the lowest level the file system format, or encapsulation, must be known. This information gets especially important as the years go by, as tools and formats slowly depreciating and are replaced by newer technologies.

Popular formats include NTFS (for Windows), EXT4 and XFS for Linux, or FAT32 (supported on many platforms). However, the demands and scale of the dataset might require less common file systems. Examples are ZFS (Unix) and ReFS (Windows), which offer superior reliability for large volumes. Some file systems also contain metadata for each file, such as the 'forks' in HFS. For large projects requiring scalability and distribution of calculations over many servers, data may also be stored on a distributed file system such as HDFS.

If data are stored or archived in a relational database (e.g., Oracle, MySQL) or a Not-Only SQL database (e.g., Cassandra or MongoDB), it is important to know the type and version, to facilitate data import to an identical system or conversion to a different product.

Files themselves can also be encapsulated in archives (with or without compression and/or encryption) or in binary objects in databases.

File structure/database design

The file structure should be described. As an example, it could be described as Vehicle/Year/Month/Trip.

Files might not always be accessed with a traditional file system; if not, it is also important to describe how to access them. Examples include Content-Addressable Storage (CAS). The analyst accesses the content, without knowing its location, using a key.

It is recommended that the schema be documented graphically to indicate the relations between the different tables, a task usually easily accomplished using data management software. This principle should be applied whether data are stored in a relational database data or an alternative (i.e., in a file system or Not-Only SQL environment)

Data container

The data container describes the format of a file. This could be avi for a media file, csv for a text file or mat-file for data used by MATLAB. With a non-standard format it is important to describe it in detail, including file content structure, header length, data type and indices. It is also good practice to include information about tools that can interpret the data format of the container.

Content

The content description should include how the data are organized in a file or object. Thus, codec and indices could be provided for an avi file, the description of a row for a csv file and the object design for a mat-file. It is recommended that the data descriptions be kept in a readable format; XML is recommended, since most tools/programming languages have built-in methods for reading xml files. A description of the file contents gets even more important if a non-standard format is used. Similarly, when different data types are mixed in the same file (e.g., video and CAN data) it is vital to have a precise description of the content. The content description of a database includes detailed information about the tables, such as columns and their respective data types, indexes, triggers, sequences and views.

4.3.3 Administrative metadata

Administrative metadata are collected for the effective operation and management of data storage and catalogues. This administrative information, covering various topics, is stored along with the datasets. From a data re-use perspective, the key role of administrative metadata is to cover access conditions, rights, ownership and constraints. Generally, administrative metadata can include (Puglia, S., Reed, J., and Rhodes E., 2004):

- version number
- archiving date
- information about rights, reproduction, and other access requirements
- archiving policy
- digital asset management logs
- documentation of processes
- billing information
- contractual agreements
- end of life of the data

The method for storing administrative metadata depends on the specific archive or repository. Many of the items above need to be stored at least as supplementary

documentation, according to repository/catalogue guidelines, if not directly as attributes of a dataset. The administrative metadata also have a role in data protection: defining processes, personal data management, access rights and keeping track of (for example) periodic backups.

For online data catalogues covering FOT/NDS/CCAM, information about a contact person/organisation and licensing options or required agreements must be included, so potential analysts know how to gain access to a dataset. Another required administrative feature of a catalogue is usage logs of information queries and retrieved data, to be able to summarize the level of interest for different datasets.

Assigning persistent identifiers for datasets is necessary for references and citations. Some persistent identifiers like the Digital Object Identifier (DOI) also support dataset version management. Each time there's a change in the data, a new DOI can be assigned, depending on the repository's versioning policy, and a log of changes collected. As an alternative, a fixed dataset reference could be used in publications when the dataset has been used. The SHRP2 Insight portal[7], for instance, regulates a dataset reference in the terms of use.

4.3.4 Test study design and operations execution documentation

The study design and experimental procedures must be documented well enough so that persons and partners who did not take part in executing the test can perform analyses. The main purpose of this documentation is to describe, in free form, the purpose of the data collection, the experimental procedures and the important details of the actual execution – including a description of the test site, which must be known before the data are interpreted. As a result, this documentation should contain not only initial plans, but also the final details of the study. More information is available in the FESTA Handbook. The document should give an overview of the following (at least):

- purpose of the field tests or data collection
- research questions
- sample selection criteria and overall description of recruitment
- possible grouping of participants (e.g., test groups 1 and 2 and a reference group); description of the groups
- overall description of equipment used, functions, HMI, additional driver support in the vehicle (navigators, etc.) and vehicle fleet – preferably with links to videos demonstrating usage
- description of the test site (if it was within a specified perimeter), including maps and photos
- date and timing of different phases of the study
- description of scenarios/test runs/study phases (e.g. baseline vs treatment phase), if relevant – with photos of key locations and views from participants' perspectives

[7] https://insight.shrp2nds.us/terms

- test plan and execution, describing (for example) what the participants were asked to do, how and when the briefing was given, what questionnaires were administered or what interviews were given
- in the case of a FOT, how the participants were introduced to the system
- how contact was maintained during the study
- special events and changes that may affect data analysis (e.g., roadwork, strikes, economical changes, special weather)
- summary information of the project and cooperation partners, duration, budget etc.

4.3.5 Metadata in data catalogues

In online data catalogues, metadata plays a critical role in enabling users to discover and access data. A well-designed metadata schema can facilitate the search and discovery of datasets by providing standardized, consistent, and structured information about the data.

Fragmented metadata practices hinder data sharing and reuse of data. Different organizations use different metadata formats, which can result in inconsistencies and incompatibilities between datasets. Therefore, there is a need for a consistent and widely-accepted metadata format(s) across transport and CCAM tests, which will improve the sharing and reuse of data. The metadata format(s) should be flexible enough to adapt to evolving technology, allowing the addition of new fields as the technology advances.

Field-specific metadata can provide a deeper level of integration with existing data services and generic catalogues. The standardization of metadata improves discoverability and interoperability between datasets, enabling users to make meaningful connections across different fields.

For FOTs/NDSs, FOT-Net data implemented a data catalogue structure (Innamaa S. & Koskinen, S., 2017). To search effectively for CCAM datasets in an online catalogue, it is essential to document specific values. These include:

- vehicle types (passenger cars, industrial vehicles, public transport, other)
- tested system (automated driving, integrated driver support, aftermarket, other)
- number of vehicles and test subjects
- data log contents (naturalistic driving, fixed routes, raw sensor data, processed surrounding objects, accurate positioning)
- public anonymous sample data available
- logs contain driving in/during (urban, rural, hilly, snowy, heavy rainfall, fog, night, traffic jam).

In addition to these searchable keywords and values, it is important to include a summary of the dataset in the online data catalogue. This should include the title of the dataset, the DOI and URL where the dataset can be accessed, a short public description, the test start and end dates, the country where the test was conducted, and the main coordinate. The online data catalogue should also include basic administrative metadata such as the publisher and

contact persons, access requirements, and documentation language. Finally, key structural metadata such as the log file format should also be included to provide users with a more complete understanding of the dataset.

A metadata documentation template is provided as an Annex I of this document. By following this template, test leaders can ensure that their CCAM datasets are easier to re-use, and that valuable research data is not lost due to incomplete metadata practices.

5 Data-protection recommendations

Data protection is key in creating trust between a Data Provider, Data Owner(s) and Data Consumers. The Data Provider is responsible towards the Data Owner(s) to ensure that data are being handled according to agreements or contracts as well as the legal context in the country where the data is managed. Subsequently, if the Data Provider knows that the Data Consumer has good, proven procedures in place to keep control of who is using the data, and that the persons working with the data have knowledge of the legislation surrounding the handling of personal and IPR data, they will be more willing to allow access to or share data.

This chapter applies whenever the data are shared between two (or more) organisations. There are many different scenarios where data can be shared, and the organizations must discuss the following questions beforehand:

- Which categories of data are being handled and exchanged?
- What risks are considered when exchanging or handling the data?
- How are the data going to be accessed between the organisations?
- What is the purpose of exchanging data and are there limitations in usage?
- What physical security requirements must be in place?
- Which logical (as in software and IT-infrastructure) security requirements must be in place?
- Which organizationa<l measures (procedures and routines) must be in place?
- When must data be erased?
- Which laws, policies, agreements and licences apply to the handling and exchanging of the data?

When data are collected and used within the same organisation there might be greater control of how the data is handled, but this chapter could still be applicable.

This chapter discusses the different demands imposed on data protection by different categories of data. The scope of data protection includes unauthorized access, data theft, data loss and the proper documentation of the implementation. The chapter includes a suggestion for data-protection requirements to facilitate the setup of the necessary data-protection framework, for a D*ata Provider* and a D*ata User*. This concept is extended with the principles of federated data access.

5.1 Stakeholders

There can be many stakeholders involved when two or more organisations agree to share data. In International Data Spaces terminology, the Data Owner is the organisation that owns the right to define Usage Contracts, Usage Policies and Payment Models (incl. third party usage). In other contexts, such as the Gaia-X Data Exchange Services specification (Gaia-X, 2022), this role is referred to as Data Licensor. A Data Creator creates data, e.g., from one or many sensors in a vehicle. The Data Provider makes data technically available for data exchange. In many cases the Data Creator, Provider and Owner are the very same organisation. The Data User is the organization that has the legal right to use the data as by the Usage Policies defined by a Data Owner. The Data Consumer receives data from, or access data at, the Data Provider.

In the terminology there is also a pre-defined role of a Service Provider. The Service Provider receives data from one (or many) Data Provider(s) (or other Service Provider(s)) and distributes the result to a Data Consumer.

It is important to state that a single organisation can act in one, many, or even all, of the roles.

A participant in a study is defined as a Data Subject per GDPR, the person who generates or is monitored while the data being collected. This person is protected by legal rights concerning the usage of the data.

Data Licensor/Data Owner

A natural or legal participant who own usage rights for data (personal or non-personal). In the context of IDSA, the Data Owner may attach usage restriction information to their data before it is transferred to a Data Consumer. To use the data, the Data Consumer must fully accept the Data Owner's usage policy. This ensures that the Data Owner maintains data sovereignty and control over their data.

Data Provider

The Data Provider must implement appropriate data-protection means to ensure responsibility and liability, as stated in Usage Policies with Data Owner and the Data Subject. Since an organization can obtain many roles, this can lead to chains where organizations are acting as Data Consumers and Data Providers. It is important for all parties in the chain to have a clear picture of the data flow, comply with the data-protection requirements and thoroughly understand the rights and Data Usage Policies and privacy laws in the country where the data are being managed.

Data Consumer

A data consumer might be allowed to download data from, or operate within, a data provider. An organization can establish a physical or logical *Site* where the requirements stated by the Data Provider are implemented. The data consumer within this Site must accept and follow the data-protection principles. In many cases an organisation acting as a Data Provider also acts as a (internal) Data Consumer, although it might be practical to keep the distinction between the two, especially in large organisations when managing personal and/or confidential data.

5.2 Data classification

The level of data protection required depends on the harm the data could do if revealed and the legal requirements. If the dataset consists of personal or confidential commercial data, it is mandated by law that action is taken to ensure data protection, regardless of the size of the dataset. Confidential commercial data is usually accompanied by agreements stating the conditions for access and use, whereas the use of personal data is regulated by law and the agreement with the participant (via consent). This document classifies data into *personal*, *special categories of personal data*, *confidential*, and *licensed* data.

Personal data

GDPR (GDPR REGULATION (EU) 2016/679) reformed the usage of personal data in Europe when the regulation came into force on May 25th, 2018. The GDPR strengthens the

rights of individuals and sets a common legal framework for all European Union countries. Any organisation that handles or processes personal data in the European Union must ensure that personal data are managed according to the law. GDPR states in Art. 3 that the law applies also to processing of personal data monitoring of person behaviour taking place within European Union, regardless of the processing being done within or outside of European Union. Any organisation planning to share personal data to third countries outside of European Union must pay great attention to what can be shared and how.

Even though GDPR harmonizes the regulations in a European Union context, there will still be differences in implementation between the US, Australia and Asian countries. For example, in the US, 'personal data' are known as 'personal identifiable information' (PII) and 'specific categories of personal data' are known as 'sensitive personal information' (SPI or SPII). The definitions are not identical to the ones being stated in Europe, and it is therefore advised to take any necessary actions to ensure that data are managed according to the laws of the country where the data are located.

The term *personal data* is defined in GDRP Art. 4:

'personal data' means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person

There are also special categories of personal data that requires additional consideration defined in GDPR Art. 9:

Processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation shall be prohibited.

The suggested data-protection requirements in this chapter aim to guide the Data Providers and Data Consumers in setting up a data-protection concept that meets the regulations and respects the will of the participants as stated in the consent form.

Confidential commercial and licensed data

Confidential commercial data is information which an organisation has taken steps to protect from disclosure, because disclosure might help a competitor. The sensitivity of confidential commercial data usually dictates the data-protection requirements stated in the data-sharing agreements. When contracts for providing the data are being signed, it is advisable for both parties to discuss, and agree on, the level of protection level that will be suitable.

Some data might be less sensitive and are here defined as *commercial data*. There is no exact definition but is in the control of the Data Owner to classify.

Any confidential commercial data should be shared in accordance with a *license* which describes the rights for handling specific signals. These signals can be organized in different sub-categories (from the perspective of where data was generated) where each might need a specific data protection implementation. These sub-categories could include 1) data generated by a driver or passenger, 2) data available on ODB2, 3) data from specific

systems indicating state, instructions/recommendations, or actuations, 4) data from a third-party system or sensor provider, and 5) external data.

The questions that could be asked are:

- Who / what is generating the data (a vehicle, a person, a sensor, post-process)?
- Are any data merged / processed with other datasets (e.g., measurements from a GPS are merged with the content in a map database)?
- Are there any contracts/licenses that restrict the usage?
- Can post-processing the data reduce the sensitivity (e.g., aggregating data)?

Several considerations affect the usage and protection level of the data. It could be relevant to divide the different categories logically to make it more obvious which data is being in use. This should be seen in relation to the effort needed for an analyst in using the data efficiently. It is important to have both technical and organisational measures in place to comply with the data protection requirements agreed upon in the respective contracts.

There could be mix of the different categories within the same dataset. Depending on the classification multiple data protection requirements may apply.

5.3 Privacy preservation, anonymization and feature extraction

The term *personal data* relates not only to data used in the actual analysis, but also to any other pieces of information connected to the dataset that could somehow identify a person (e.g., any references to a person in a file on your local computer or a printed document stored in a safe with contact information to a participant). The HIPAA Privacy Rule (The US Code of Federal Regulations including the HIPAA Privacy Rule at 45CFR Parts 160 and 164, 2012) lists 18 elements as direct identifiers, including the following data types commonly used when performing a study: names, zip codes, all elements of dates (except year), telephone numbers, mail addresses, social security numbers, account numbers, vehicle identifiers and serial numbers (including license plate numbers, full face photographic images and any comparable images). GDPR is not as explicit, however defining the term personal data as: "*any information relating to an identified or identifiable natural person*".

De-identified data are obfuscated data, making a person's identity less obvious and minimizing the risk of unintended disclosure (Nelson G., 2015, GDPR), whereas *anonymised data* are data that cannot be traced back to an individual by any means (Nelson G., 2015). A Data Provider must strike a balance between identifiable and de-identifiable data, using different approaches to obfuscate the participant's identity by implementing a variety of features or algorithms. Possible methods include record suppression, randomization, pseudo-identification, and masking and sub-sampling (Nelson G., 2015). It is important to underline that even if data is de-identifiable, it is still personal data according to GDPR.

To adhere to GDPR it is important for the organisations to act according to Art. 25 on data protection by design and by default. This means that any organisation managing personal data must implement technical and organisational measures from the earliest stage of data processing. This could mean that any data transferred must be encrypted, safely and securely stored, and that any direct identifiers (e.g., driver name or vehicle registration number) would be replaced in a pseudonymization step (e.g., using driver id or vehicle id). In addition, the European Union Agency for Network and Information Security (ENISA) propose

different strategies for “Privacy by design in the era of big data” (D' Acquisto et. al. 2015), described in Table 13.

Table 13: Privacy by-design strategies

Privacy by-design strategy	Description
Minimize	The amount of personal data should be restricted to the minimal amount possible (data minimization).
Hide	Personal data and their interrelations should be hidden from plain view.
Separate	Personal data should be processed in a distributed fashion, in separate compartments whenever possible.
Aggregate	Personal data should be processed at the highest level of aggregation and with the least possible detail in which it is useful.
Inform	Data subjects should be adequately informed whenever processed (transparency).
Control	Participants should be provided agency over the processing of their personal data.
Enforce	A privacy policy compatible with legal requirements should be in place and enforced.
Demonstrate	Data controllers must be able to demonstrate compliance with privacy policy into force and any applicable legal requirements.

The actors must take the two concepts of data protection by design and by default, and privacy by design, and balance them to the research needs. It is important to document the decisions, implementation, and have the organizational means to monitor the compliance with the required data protection and privacy measures.

Having a completely anonymised dataset could mean that the usefulness and value for analysis is reduced. In any case, legal and ethical restrictions on how long one is allowed to keep a personal dataset may force its deletion.

For rich media (such as video or images), feature extraction is a method to preserving privacy, still allowing valuable datasets to be used for analysis. Feature extraction could be used to translate rich media data into measures, thus removing the identifiable elements. Efficient feature extraction could solve two major issues when having rich data collected in a project and current datasets could be shared, and features could be extracted from data before purged.

The first decision to make is which features should be extracted from the data; if the extraction is being performed prior to data deletion, Data Owners, Providers and Consumers

must collaborate on this difficult task. Finally, the project must decide if it has the extensive computational resources required to extract features from a large dataset.

The main benefit of feature extracting is the possibility of enhancing existing datasets with new attributes or measures, previously only available from costly manual video coding processes.

GPS traces are also considered personal data, albeit indirect, as they can potentially reveal where people live and work and even their children´s schools. Similarly, no detailed travel diaries covering long periods of time can be made public if they contain addresses, even though a person making a single trip in the diary could be anyone living or working at those addresses. There are many approaches being explored to ensure personal integrity, e.g., k-anonymity and differential privacy (D' Acquisto et. al. 2015). The trade-off here, between anonymization and maintaining usefulness of the data for research, is difficult.

5.4 Data access methods

The method of data access is important for the requirements on the actors involved. The principles are described in section 2.4.

The principles can be translated to three different data access methods: 1) public or conditional download where data is transferred using any secure file transfer protocol or hard drives to the data consumer, 2) remotely or exclusively accessed at the premises of the data provider, or 3) the data consumer accessing a connector controlled by the data provider, the hybrid version of the previous three in a data space context. Each method has its own implications; usually, the data categorization has the greatest impact on selecting which method to use.

5.5 Organisational measures

Any actor shall be allowed to access, manage, or share, personal or confidential data before the data-protection implementation is documented. The actor must fulfil legal requirements and document how to meet requirements. It could also be valuable to get an independent review of the implementation by an external party. Table 14 describes the steps of implementing the organisational measures:

Table 14: Organisational measures for data protection

Measure	Description
Data Supervisor	Appoint an individual as *Data Supervisor*. The Data Supervisor is responsible for mapping, implementing, documenting, and following the requirements for data-protection.
Data controller and Data Protection Officer	If personal data are included, the organisation (in a European context) handling the data assumes the responsibilities of being a *Data Controller* and must appoint a *Data Protection Officer*.
Legal	Ensure legal compliance with current legislation.

Documentation	Compile data-protection documentation describing the implementation.
Ethics	Determining whether the intended data usage requires approval from a national ethics committee.

Ensure that anyone using or handling data has relevant contracts signed, including *non-disclosure agreements*.

5.6 Data extraction

Data extraction is the process by which results from analysis are being extracted from a research infrastructure to be (publicly) used in journals, project reports, or for presentations. Depending on the scope for publication, some instances also require the data to be posted aside the paper. The reason for this is the concept of reproducible results. This is challenging in the eyes of both GDPR and IP related matters. It is important for any actor to be aware of this beforehand.

Data extraction must be executed in-line with any existing agreements (which typically are agreements with Data owner(s) or Study participant(s)).

Published information shall be de-identified as well as checked and cleared for IP related matters beforehand. The Data Provider is responsible for defining the data extraction process and the Data Consumer must be aware of the principles, rules, and routines, specific for the dataset.

Both Data Supervisors should be aware of any data extraction request (unless their responsibility has been delegated) and decision. A certificate of the decision could be attached with the actual information extracted. It is recommended for the actors to archive the approved data extraction requests.

A data extraction request could cover following elements: 1) intended use of the extracted data; 2) list of data types; 3) description of the data; 4) total size of the data; and 5) list of files or folders to extract.

5.7 Data protection at a Data Provider

It is imperative that any organisation hosting IP classified or personal data document its data management processes. The term *Data Provider* include any infrastructure where data is stored in the long term and can mean a data repository or a data space connector. Depending on the classification of the data, different precautions might have to be taken. If the data include personal identifiable data or confidential data, stronger requirements need to be formulated. The data handling needs to be documented and there are frameworks that could be considered (if not already established) to ensure that the necessary processes are documented and traceable. For example, ISO 9001:2008 for Quality management systems, ISO/IEC 27001:2013 for Information security management, ITIL (IT Infrastructure Library), or the UK initiative Cyber Essentials could be used. Additionally, similar (although not formally acknowledged) quality assurance procedures might also be suitable; the most important consideration is that the organisation reflects on data security and access – and implements routines that ensure data protection.

It is important that third-party organisations (e.g., a cloud-based data-hosting company or a third-party organisation managing parts of the IT infrastructure) comply with the requirements GDPR addresses this in Art. 28 and Art. 29, and that these are documented.

It is also stated in GDPR Art. 35, that any organisation managing personal data must make an impact assessment on the risk in case of a data breach.

The Data Provider must address data protection measures and document the data-protection implementation. An overview of topics to address is described in Table 15

Table 15: Data Provider data-protection documentation

Topic	Description
Overview	• Presenting the scope for data hosting, handling, and processing. • Defining the start and end date (if applicable) for data hosting. • Providing a description of the organisational structure. • Providing an overview of personnel who will have access to data.
Legal	• Analysing the responsibilities in the context of data protection and privacy issues, including GDPR and national legal compliance. What legal issues must be handled, and how will this be done? • Consider formulating a Data Protection Impact Assessment (DPIA) if applicable. • Describing relevant contracts/agreements and the impact on data usage, publication, and further data sharing/exchange.
Status, implementation and assessment	• Providing status of the described implementation; is it planned or already implemented? Provide time plan with technical details where applicable. • Providing disaster recovery plan with risk assessments. • Providing incident response plan for data security breaches with risk assessments. • Providing relevant internal routines/guidelines, as well as training for personnel. • Describing how data is protected from unauthorized (physical and logical) access. • Describing how data can be securely transmitted from Data Provider to Data Consumer. • Describing how data is protected from accidental deletion (see appendix A.3). • Describing the principles for how data access is granted

	and which agreements that need to be in place. • Describing the usage of the Data Management Plan as a mean to plan for data preservation and handling after a project has ended.

5.8 Data protection at a Data Consumer

The Data Consumer must address and document the data-protection implementation. Before a Data consumer gets access to data, the Data Provider can require to be presented the implementation of the data protection measures stated in Table 16, as an important step in creating trust between the actors.

Table 16: Data protection documentation for data consumer

Topic	Description
Overview	• Presenting the scope for data usage, handling, and processing, including plans for disseminating the results. • Defining the start and end date (if applicable) for data usage. • Providing a description of the organisational structure (relevant to the usage of data). • Providing an overview of personnel who will have access to data.
Legal	• Analysing the responsibilities in the context of data protection and privacy issues, including GDPR and national legal compliance; what legal issues must be handled, and how will this be done? • Describing relevant contracts/agreements and the impact on data usage, publication, and further data sharing/exchange.
Status, implementation and assessment	• Providing status of the described implementation; is it planned or already implemented? Provide time plan with technical details where applicable. • Providing a detailed description of the infrastructure used (for the purpose of analysis, but also intermediate resources used for e.g., downloading, storing, or processing data). • Providing incident response plan for data security breaches with risk assessments. • Providing relevant internal routines/guidelines, as well as training for personnel. • Describing how data is protected from unauthorized (physical and logical) access.

	• Describing the principles for how data access is granted and the data extraction process.

5.9 Cybersecurity in FDS

The establishment of data spaces includes higher level of awareness related to cyber security by being online resources. The DSF from 2018 described a scenario where cyber security issues were left out, leaving it to the Data Provider to handle this issue. The proposed idea was to have a completely separated network handling these resources to reduce the number of attack surfaces and block intruders. This was a simplification, and many larger organizations deal with these challenges in their internal networks every day, having global presence and thus a large, distributed networks.

Leaning on firewalls and trusted IP networks might be impossible in FDS. These services will most likely be put on the Internet which open for different types of attack surfaces:

- DoS
- Certificate spoofing
- Data leakage or theft
- Malware (to get to other systems).

The trust mechanisms for authentication and authorization require a strict implementation to ensure that only authorized users and systems can access the data. This can include two-factor authentication, secure certificate handling and monitoring, password policies, and other access controls.

Data should be encrypted both at rest and in transit to protect it from unauthorized access or interception. Encryption algorithms and protocols should be carefully selected and configured to provide a high level of security.

Data Providers should only share the minimum amount of data necessary to achieve its goals. This can help reduce the risk of data breaches or unauthorized access to sensitive data. Access to shared data should be limited based on the principle of least privilege. This means that users should only have access to the data they need to perform their job functions.

Data Providers should implement continuous monitoring and auditing of data access and usage to detect and respond to any security incidents or breaches.

Both Data Providers and Consumers should have a well-defined incident response plan in place to respond to any security incidents or data breaches. The plan should include procedures for identifying, containing, and mitigating the impact of a security incident.

5.10 References to accident databases

Accident data are a special type of data, related and connected to CCAM/FOT/NDS data as a collection of special situations which usually form a very small (but highly interesting in the safety context) subset of data, and are widely used globally. They are discussed here as a case study.

There are several projects world-wide that collect and protect accident data for scientific analysis. The context of these projects is differing and very mixed. Partners range from governmental institutions to universities and companies. With this great variety of users, there is a need for effective data protection. Interestingly, the actors even within one accident data project could be located in several countries and form different types of legal organisations.

Moreover, accident data projects are long-term, so the process of anonymization is crucial for their survival. There is always a chance that persons involved in an accident may ask for data related to their case. Safety-related data, especially accident data (which imply legal aspects), need more care than non-safety related data (e.g., data for driver behaviour analysis) when collected in a scientific context and thus are a good testbed for data protection. The level of anonymization is largely independent from the level of sharing the data, it can even be accident data collected and stored only by one OEM. It is a matter of the legal requirements that must be applied at the location of the Data Provider.

When data are anonymised, the link between a dataset and a specific person, accident or geographic location is cut. Then, the data can be used for a scientific purpose, but you cannot use it anymore in the context of legal affairs. Anonymization is crucial, as those who are responsible for data protection would stall the project without it.

Technically, accident data is protected by the Data Provider, who removes any details which can be directly connected to a single accident, or a person involved in that accident, before entering the data into the database. In particular, participants' identities, exact geographic locations and exact dates are removed. Usually, pictures are also included in the data, necessitating a more complex process of anonymization: e.g., faces and company logos (e.g., printed on vehicles) must be blurred to make them unreadable, which yet cannot be done fully automatically, and manual intervention is required.

An interesting challenge arises when there is a need to link third party data to the already anonymised accident data supplied by the Data Provider. It would be useful, for example, to know the equipped safety features of a car involved in an accident, to analyse their effectiveness. However, direct access to the equipment information for a single vehicle (other than standard equipment, which can be determined by make, model and year) require the vehicle identification number (VIN) of the vehicle, which is not usually available to the Data Provider.

One solution is to provide a list of VINs, without any accident data information, to the third party. But as these VINs identify vehicles known to have been involved in accidents, this solution is not compliant with common data-protection requirements. In fact, to date this problem has only been solved in a closed environment (like an OEM). However, hosting data in a closed environment also needs to honour the legal restrictions which are valid at the Data Providers location. This differs between legal systems and the type of personal data stored, e.g., names and other details must be removed from medical data and faces on pictures must be blurred. In this example of information linkage, it must be considered that the VIN only points to the owner of a vehicle, not directly to the persons involved in the accident. This example shows how important data protection is, and how seriously it is handled in current scientific accident databases. This situation is not necessarily restricted to accident data and should be considered in other domains, too.

In some legal and political constellations, an increased level of data protection has to be practiced. Such constellations can occur in mixed environments, when public and private institutions run a joint project. The Data Provider has to meet certain additional requirements: for example, it has to be a server at a university, and the anonymised data are transferred over secured lines to the Data Consumer.

There is some variance in data-protection requirements around the globe. For example, in the US, accident data collected by the government are made public and can be downloaded from websites. Access is regulated by the US Federal Research Public Access Act (FRPAA) and the US Fair Access to Science and Technology Research Act (FASTR). The main reasoning behind public access is the social benefit from publicly funded research to all taxpayers which on the other hand is opposed to the protection of each individual's data in the case of accident data. There is no other country with similar regulations worldwide. When data is published, it is highly important to remove/hide personal information. It should be noted that the anonymization level of US accident data is about the same as that of non-public databases in Europe, including blurred pictures and cut-off vehicle identification numbers.

In practice, data protection has proven for decades to be feasible when dealing with accident data in a scientific context.

In a CCAM context there are different legislation and requirements on reporting accidents. In the US, many states investigate any accident for AV prototype vehicles. Some countries do as well in Europe. However, the protocol for collecting the information is not standardized, but there are initiatives looking into this.

6 Training on data protection related to personal data and IPR

All personnel handling data need to undergo training in data protection if the data is personal or conditioned based on intellectual property rights (IPR). Persons or organisations collecting and managing personal information must handle the data according to GDPR, protect it from misuse, and respect the rights of the data owners.

Protection of intellectual property rights is another important aspect when working with datasets, including video, especially when research partnerships include industrial partners. The data could reveal algorithms of certain systems if re-engineered, and therefore need to be protected.

Training on personal privacy issues and IPR needs to accompany the general training on the data security measures put in place to protect the data. The level of training should be adjusted to the content of the specific dataset to be protected.

6.1 Set-up and content of the training

Who and when?

To ensure protection of personal data and IPR, training procedures must be in place and provided prior to any data access. Training material and procedures can be created by the organisation providing the training or possibly bought from the data provider's Support Services. Training must be given to analysts, video annotators, those responsible for the database, visiting researchers and all other staff handling, analysing or looking at personal or IPR data. Even persons to whom data are shown (during a demonstration, for example) must be informed about relevant data protection and IPR issues beforehand.

What?

The training needs to cover the following topics (the level of detail can be adapted to target audience's needs):

Table 17: Data protection topics to address in training

Topic	Description
Description of the data with special focus on personal data and IPR	• What are personal data, in general and in this specific context? • What are intellectual property rights, in general and in this specific context? • What data are collected with (for example) video, questionnaires or GPS tracks? • Information about data ownership and access rights for partners/third parties.
Data-handling requirements originating from national and other applicable laws,	Personal data must be: • processed fairly and lawfully, • obtained for specified and lawful purposes,

regulations, and rules.	• adequate, relevant and not excessive, • accurate and up-to-date, • not kept any longer than necessary, • processed in accordance with the participants' rights and acceptance, • securely kept, and • not transferred to any other country without adequate protection in situ.
Explanation of the consent form content, especially the specific active consents related to data sharing (voluntariness, comprehension and disclosure):	• How should the study participant be informed about data collection, purpose, handling, storage, and access – including re-use after the project ends? • What is included/excluded in the participant's consent? (For example, participants give their consent to collect videos for analysis purposes and to video of them being shown in conference presentations).
Data-handling procedures	• Practical rules and procedures for data access (rooms and workspaces with limited access, personalized keys, password protection) • Data structure • How the data are anonymised, pseudo-identified and/or encrypted • How the data are accessed, in order to (for example) perform analysis • The contact persons for different procedures including the data protection responsible • Whom to inform in case of deviations. • Information about publication rights.

How?

It is recommended that a personal training session be organised to answer questions and make sure that all staff members know their responsibilities. Online courses might be helpful to provide additional valuable information, but they are not considered sufficient on their own as they cannot cover local implementation of the security precautions. A basic understanding of three principles essential to the ethical conduct of research with humans: respect, beneficence and justice.

6.2 How to document?

Documentation of all training is recommended, most conveniently recorded on the analyst's information sheet, which the participant needs to sign. Although analysts might have an NDA in their certificate of employment, the process of signing the document enhances the protective level of the data.

It is recommended that the following records be kept:

- persons who have undergone training
- training procedures
- process descriptions
- contact persons for different procedures including the data protection responsible.

7 Support and research services

Support and research services are essential to data management, preservation and re-use. Depending on the knowledge and responsibilities of the persons re-using a dataset, either support services alone are provided, or research services are also required.

Support services comprise all activities in which support is being provided for successful data re-use. Support can be provided in various forms, starting with supplying information and ending with assistance with data analysis methods and procedures.

Research services comprise all activities where research work is carried out for the client, ranging from advice on specific research questions in different research stages to a more complete research endeavour providing a detailed analysis of specific research questions. The services are more targeted to the latter.

Analysis tools are an integral part of support and research services. The efforts and costs are to be included in the business model for the re-use of the data. These are discussed in chapter 8.

7.1 Support services

Support starts as early as the application stage, with discussions on the suitability of the data to answer the specific research questions at hand. Support services target the researcher's ability to perform analysis and re-use existing data. The services are divided into different stages depending on the degree and impact of the support. These stages are:

- information and data provision
- supporting tools
- assistance with dedicated research needs
- data-protection and analysis facilities
- long-term preservation and data-sharing services.

Information and data provision

The first stage of support is to make researchers aware of available datasets and tools for data handling. This information is usually provided in online data catalogues. Furthermore, discussions may be necessary to answer questions about data usability (based on feedback from initial data analysis or from already performed data re-use) and which procedures have been established and proved to be successful. Metadata and other detailed background information on the data collection and initial study design can provide a better understanding of the dataset and improve data handling. Additional services, such as basic data aggregation and data extraction and transfer, could also be provided.

Supporting tools

Tools are an integral part of the support services. These tools consist of viewing and annotation tools, scripts to extract useful datasets from a database and licensed SW – and can also include entire frameworks for retrieving, processing, and uploading data back into a database. However, it is important that the analysts are free to choose what tools to use without being constrained by factors other than the raw data formats and data descriptions (for example, by complex frameworks with graphical interfaces). It is, as mentioned in

chapter 4, important that raw data can be read in a clearly described format directly from the data storage source (e.g., database or file storage), regardless of what analysis tools are used in the project. Note that appropriate access restrictions should always apply. Allowing analysts to choose their tools is important, since different analysts have different ways of analysing data. Support services should impose as few constraints as possible on what processes analysts can use to analyse the data (within the data-protection framework). Examples of different ways to structure data and metadata are given in chapter 4. In addition, data and metadata formats will have to be able to support different analysis processes and needs, to be accepted and used by as large a community as possible. It is also important that the dependency on third-party software for access is kept to a minimum.

Support may consist of providing dedicated tools for specific tasks (if available) and setup and basic maintenance of the analysis tools. Due to the complexity of data analysis, the setup of these tools requires a profound understanding of the datasets. Further developments of the tools fall under the stage Assistance with dedicated research needs of the support services.

Assistance with dedicated research needs

Assistance, the most advanced stage of support services, can take the form of dedicated advice on analysis methods and the custom modification of tools. In a strict sense, analysis methods are not applied (this would be part of research services, see 7.2); instead, this service selects, provides, and adjusts analysis methods.

Data protection and analysis facilities

The following support services can also be provided:

- analyst training
- support relating to privacy issues
- data-protection measures
- secure facilities for analysis work

The researcher could be given training in security and privacy matters, thus gaining a deeper understanding of the sensitivity of the data. Training in using analysis tools could also be included (see chapter 6).

Support for new research projects on confidentiality and privacy issues is a common role for data warehouses.

Advice and support could be given on the need for data-protection measures.

Certain data warehouses offer secure sites/rooms for analysis. In these cases, the data may not be transferred, but must be analysed on-site to fulfil security requirements.

Long-term preservation and data-sharing services

Support services for long-term data preservation, access and reuse feed into the next cycle of data re-use, lengthening the lifespan of data into future projects.

For data to retain its value it needs to be preserved during or at the end of the data-collection/research project during which it is created. Long-term preservation requires a dedicated infrastructure and human resources, as well as planning and preparatory work by

the data creators. This ensures that data preserved this way are properly formatted and documented, and that their future management is planned for.

Another aspect is access and findability. For data to have future value it needs not only to be properly preserved, but also to be found and accessed. Users need to be able to able to search for relevant datasets and evaluate them with the help of metadata and documentation. There also needs to be information about how to obtain access to the data.

The above functions can be fulfilled in different ways – by a centralized solution such as a data repository which preserves and makes data findable and accessible, or by a decentralized solution such as a dataspace, a data ecosystem where trusted partners host and share their own data according to agreed-upon data storage, access, sharing and interoperability standards.

For the framework to be adhered to, some data expertise and resources need to be dedicated to preparing and managing data. Three commonly used roles within research data support are Data Steward, Data Curator, and Data Manager. These roles can overlap one another and be fulfilled in different ways. There are also other data-related roles which either overlap or are equivalent to one or more of the above roles: data librarian, data custodian, RDM (Research Data Management) specialist, research engineer, etc. The three main roles are described below:

Data steward: a governance and compliance role responsible for ensuring that data processes and usage is in line with organizational policies and regulations, certifications, legal and ethical requirements, and IT-security requirements, all with the aim of producing FAIR data. Data Stewards can be responsible for overseeing data documentation, curation, and structure across an organization, with FAIR and long-term preservation as the outcome. Reproducibility and reusability are key concepts here. A Data steward can be generally available to an organization or embedded within a project or infrastructure.

Data manager: a more operational role responsible for overseeing the lifecycle of data within an organization or project. Data managers work with data storage, retrieval, processing, and possibly analysis of data. They may be responsible for designing and implementing data management systems and processes, and for ensuring that data is accessible and usable by those who need it.

Data curator: ensures that data is accurate, relevant, and useful, and may be responsible for acquiring data from various sources, such as databases, websites, and other data repositories, and for ensuring that the data is properly formatted and organized. Data curators also work to ensure that data has proper metadata and documentation to aid findability and reusability, as well as maintaining data, metadata, and documentation over time.

7.2 Research Services

Research services have a role beyond the initial start-up provided by the support services. In this case, the data provider takes part in the actual research to be performed, if required by the analyst. If the analyst comes from another discipline and/or is unfamiliar with the type of data and therefore would like to have it aggregated to a more suitable format, the research services can assist. A deep understanding of the research questions is necessary to aggregate the raw data in the best way without losing relevant information. The work

performed by the research services can extend as far as performing the complete analysis, answering specific research questions.

The three levels of research services are:

- research advice on methodology
- research involvement/research support
- complete research performance

These levels are not necessarily distinct but can overlap each other.

Research advice

On this level, advice is provided on data analysis. The advice, based on experience from data collection or previously performed analysis of the dataset, focuses on the best practice to answer the actual research questions and the related hypothesis. That is, the advice does not deal with how to solve a problem (using tools, data handling, data protection and/or data processing), but focuses on what methods should be used to get to the desired results.

Examples of research advice are:

- determine whether a dataset can be re-used
- review the scientific approach/method for re-using data
- review whether a dataset is appropriate for the research questions, hypothesis and indicators.

Research involvement/research support

The second stage of research services is an active involvement in the research to be performed in terms of:

- support in the identification/selection of data for re-use
- development of specific tools for:
    - data handling
    - data analysis and evaluation
- performing parts of the analysis, such as:
    - formulating research questions based on research content
    - formulating a hypothesis based on research questions
    - deriving indicators from hypothesis
    - applying data analysis based on indicators
    - statistical analysis of data

Complete research performance

Finally, the highest level of research services consists of the data provider, or a third party, performing all the research. In this case, complete work packages for research on the datasets are taken over by the research service provider. Work packages can consist of work in one or more of the following fields:

- selection and provision of data
- selection and/or development of specific analysis tools
- complete analysis
- scientific reporting

8 Financial models

Efficient management of FOT/NDS/CCAM datasets is the key for successful re-use. If data sharing is not economically feasible for data owners and potential data re-users, re-use of data does not take place and the benefits of data sharing aren't achieved. Thus, in order for data sharing to gain popularity, financial models are needed that cover data management costs.

Organisations supported by public funding are facing new requirements to plan long-term data preservation and management. Future work on financial models will have to take into account both the changing conditions of public funding to promote data sharing.

This chapter discusses options for organisations carrying out field trials to fund the sharing and upkeep of datasets after the project.

8.1 Data management costs

In terms of cost items, data management has many things in common with open data efforts and large-scale user tests in various scientific disciplines. Documentation and user support have heightened roles, though, as the datasets are generally in non-standard form and have their origins in studies with specific goals. In addition, strict requirements to uphold user privacy and product IPR may require secure facilities and processes, raising the management costs of such datasets higher than those of fully open datasets.

Table 18 lists the items requiring funding in data management within this domain. The items are related to data management after a project – or more generally, after the data collection has ended.

Clearly, storing a massive dataset and organising proper backups to avoid losing data incurs costs. Data may also have to be anonymised to enable wider sharing. When sharing a dataset, licences/agreements usually need to be completed, as well as financial arrangements. Further, to justify the benefits of data sharing to funding organisations, it is important to collect information on the use of the data. As a result of such requirements, the list of data management cost items can grow long. However, that does not necessarily mean that data sharing causes a huge burden on organisations. Effective processes, support and tools provided internally or externally by professionals, can reduce the stress on participants in single projects. Basic preparations to share data should become part of good scientific practise.

Considering the general costs of data management, it is unlikely that all test data can be stored for future science. A selection process is foreseen that would concentrate the efforts and funding on promising and valuable datasets. This selection could be carried out by those who fund the costs of data sharing and supported by the experts who collected the data for the original project. The selection could be based on the following criteria:

- potential for re-use, from both scientific and business perspectives
- efforts needed to store the dataset
- quality and amount of data.

Table 18 presents cost items and tasks involved in data sharing after data collection has ended. It is assumed that tasks enabling data sharing, such as concluding legal agreements,

metadata documentation and data quality checking, have already been performed in the original project that collected the data. Some of the cost items in Table 18 are optional, such as advertising datasets or participation in international harmonisation/standardisation efforts of data collection and data catalogues. However, such tasks are common in professional data management services and can also be foreseen in data-sharing activities that have achieved an established status.

Table 18: Data-sharing costs

Cost item	Comments	Timing of cost
Data selection, enhancement of documentation (metadata), creation of entries in relevant data catalogues	Finalisation and structuring of data. As a pre-requisite for sharing, the datasets need to be comprehensively documented.	When project/data collection ends
Anonymizing data	The level of anonymization and related efforts depend on how widely the data will be shared.	Before data is shared
Management & coordination personnel costs	Basic management of e-infrastructure, including user support, data catalogue operations and updates, data import to archives, backups, compilation of usage statistics, license management, agreements and finances	Continuous
IT operations	Database servers, storage, protection, licenses and IT personnel costs	Continuous
Cyber security	Protect the system from cyber threats	Continuous
Analysis or data handling facilities	Physically secure workspace	Continuous
Analysis support services	Expert support at different levels	When data is shared and during analysis efforts
Promotion and advertisement	Informing potential data re-users and data-sharing funders Optional: Direct funding of further analysis projects, to ensure good use of valuable datasets Optional: Direct advertisement of datasets for potential research	When project ends/Continuous

	projects and those planning new projects, beyond common catalogues	
Optional: Standardisation and collaboration regarding dataset formats	Taking part in national and international collaborations regarding dataset formats	Continuous

8.2 Financial models

This chapter suggests financial models for data sharing, starting mainly from the point of view of the organisation that has collected the dataset. As the main funding for transport-related research today comes from direct governmental grants, this is also likely to be the case for the re-use of data. Future funding might be directed toward established data-sharing and e-infrastructure activities. In fact, the first two financial models in this chapter (A and B), are based on such activities.

Project-based funding is one of the current methods for financing data sharing. The models C–E consider the pros and cons of directly including data sharing and re-use in the project activities. In the models F–H, the costs fall mainly on the end user (e.g., through membership fees or licenses). Several funding sources might be required to keep data available and provide services for third parties. Therefore, the financial models can also be complementary.

A) Organisations' core activity

Digital preservation becomes a part of organisations' core activities. This model is motivated by conditions set by public grant agreements. A part of the grant for the original projects that collected the data will be directed toward central data preservation activity inside the organisation. This would cover data management and sharing for a certain period after the project is finished. The data availability for third parties should be based on reasonable conditions and costs.

A selection process may be required to decide which data will be stored, the way they will be stored and for how long. The operation of a repository can also be outsourced. However, when a dataset containing personal data is stored by a third party, it needs to be strongly encrypted to avoid misuse and liability problems (see chapter 6).

Table 19: Model A (example: social sciences universities, possibly larger datasets)

Pros	Cons
• Data would be considered IPR, valuable datasets would not be lost • Dedicated professionals would enhance the quality of the data provision procedures and analysis tools	• A burden for small organisations not prepared for such requirements • No existing selection process for funding

B) e-Infrastructures

Public funding is directed to data infrastructures, serving multiple organisations and disciplines. Centralised data management could offer professional data management services, general harmonisation and possibly greater cost-effectiveness when compared to distributed approaches. The roles of public infrastructure would cover certain tasks, but project-specific funding would still be needed when data-users or data owners request additional services.

Table 20: Model B (example: Supercomputing infrastructures and their services to universities)

Pros	Cons
• Professional data management services • Data and processing services are free (i.e., for academic re-use)	• The operators of the data infrastructure will have limited knowledge and means to provide detailed support for analysts, other than existing documentation • Dataset confidentiality sets limitations for storage by third party services • No selection process for funding exists • Valuable datasets from smaller projects might not be considered

C) Archiving included in project budget

Project budget allows for dataset finalisation and archiving in commercial services. In this model, the project budget allows for final cleaning, documentation and fees for archiving in selected data storages for a fixed period (e.g., 10 years). The project creates entries in relevant data catalogues.

Table 21: Model C (example: Research team storing its data – or making them open-source)

Pros	Cons
• The commercial service could get part of their funding from advertising, even enabling free storage	• Who answers questions regarding the dataset after a few years have passed? • Is the documentation of the required quality? • No existing selection process for funding

D) Project extension

The project is awarded a continuation to maintain its data. Model D is like model C, except the dataset is archived by the project partners. For notable projects, separate grants for operation (including data storage, promotion, calls for analysis proposals, etc.) would be awarded based on a review board decision, under specific conditions.

Table 22: Model D (example: Large research projects apply for extensions)

Pros	Cons

• Targeted promotion activities for datasets can also include funding for analysis activities and effective monitoring of results	• No selection process for funding exists • Valuable datasets from smaller projects might not be considered

E) New project funding

New projects finance maintenance or revival of a dataset. In a chain of projects, the benefit of using past data is obvious, encouraging efforts to be put into maintaining and exploiting the old dataset. Depending on the follow-up activity, the data owner might also be motivated to share data with third parties, to extend analyses for mutual or customer benefit (e.g., offering material for thesis work, benefiting the customer who originally funded the data collection).

If a data request from outside of the organisation meets the business interests of the data owner, it is welcome. Otherwise, it fails to motivate the efforts needed for data sharing.

Table 23: Model E (example: Various research projects analysing and benefiting from previous dataset)

Pros	Cons
• No changes to current funding methods (additions are needed in call texts to promote existing datasets) • When data is re-used by those who collected the data in a previous project, the re-use is very efficient	• Plain project-based funding may not be sufficient to keep datasets available and it should be seen instead as a complementary funding source • Project owners have difficulties estimating the costs required to access a dataset, when they are making an initial project plan/offer

F) Established network

A network of organisations with participation fees arranges data management jointly. Organisations within the same discipline form networks that share and promote data. Datasets are collected, documented and catalogued using agreed-on/standardised methods. The networks are likely to be formed for handling continuous operational data which meet their business interests. There could be various levels of memberships and fees.

Table 24: Model F (example: Accident data collection and sharing)

Pros	Cons
• Business aspects can be applied on fees, high-quality harmonised data • Could include freemium services, where the basic information is available for free but advanced services have a cost • Facilitates cooperation in research	• Only certain disciplines seem to reach this status

G) Analysis services

An organisation with several valuable datasets uses them to create business, offering both data and related services. This model can enable the original group that carried out a study to get further funding for their work. The model is for organisations with a prominent role in a discipline.

Table 25: Model G (example: Notable data owners / Data providers)

Pros	Cons
• Continuous research quite possibly results in high-quality results.	• Small organisations and partnership projects have difficulties setting up this sort of business and their data easily get lost. • Even valuable datasets become old and lose value for organisations purchasing analysis services.

H) Data integrators

Companies acquire and market datasets along with transport and other related datasets. In this model, a data integrator markets particularly useful datasets (among others, such as those containing real-time traffic data) licensed from original sources. Customers are offered a single access point for data, so they don't have to go through negotiations with several parties, facilitating (for example) the development of mobile applications. For a dataset to be shared without fees for the re-users, the maintenance would have to be financed through the organisation that contributed the dataset – or supporting business operations.

Table 26: Model H (example: Road operators putting together information services)

Pros	Cons
• Easy licensing of various high-quality information resources.	• Data integrators may have little interest in non-commercial academic work.

I) Data space

A number of organizations create a common space for sharing CCAM related data. The stakeholders agree upon common principles for data exchange, formats, description and taxonomy. The overriding principle is that within this data space, the control of data is in the hands of the data provider, and that access to data is conditioned by an agreement between the data provider and user. There is also a third category, "service provider", which has access to data by a data provider and adds value to the data (or integrates multiple sources). Still, bi-lateral agreements are necessary between all stakeholders where the service provider will act as both user and provider. It is possible to handle monetary compensation for data access and exchange, but as of 2023 this is not yet implemented in any framework; therefore, this must be handled outside of the data space. Also, automated contracts could facilitate access but are still not available.

Table 27: Model I (example: using data space technologies for data exchange)

Pros	Cons
• The data provider is in full control over who has access to which data.	• The data user might not have insight in the raw data which can lead to misinterpreted conclusions.

8.3 Distribution of costs

Depending on the financial model and the activities set up for data sharing, the costs are shared differently among the project that collects the data, the organisation(s) owning the data and, finally, the re-users. Table 28 considers the funders for data management and re-use in the different financial models presented previously. The costs are divided into three classes:

1) dataset finalisation, costs that often come at the end of a project
2) continuous costs coming from management and upkeep of data
3) costs when data are shared, e.g., selection of data and user support.

Additionally, the table considers the cost for the re-user in each financial model. Those costs that are potentially funded by an external party or the organisation's non-project funding (i.e., part of the organisation that is not involved in sharing or re-using the data) are highlighted.

Table 28: Funding source and re-use costs in different financial models

Financial model	Funder			Costs for re-user
	Dataset finalisation	**Continuous costs**	**Costs for provider**	
A. Organisation's core activity	Project/ Organisation's selection process	Organisation (with public funding)	Organisation/ Re-user	Non-profit price
B. eInfrastructures	Project	Publicly funded e-infra, where organisation may have a role	Publicly funded e-infra, additional services have a price	Free (basic services)
C. Archiving included in project budget	Project	Project or e.g., data storage service supported by advertising	Both project and re-user	Free or non-profit price
D. Project extension	Project	Project	Project	Free or non-profit price, even calls for analysis
E. New project funding	Re-user	Re-user	Re-user	Commercial price
F. Established network	Project	Re-users via participation fees	Re-user	Different levels of memberships and fees
G. Analysis services	Project	Organisation	Re-user	Commercial price
H. Data integrators	Project	Integrator	Re-user	Commercial

				price
I. Data space	Data provider / bilateral agreements	Participation fees	Data provider and user agreement	Data provider and user agreement

9 Data Governance procedures

The partners should agree on an application procedure for re-use of data early on in the project, so that all project partners (and any possible third parties) know the conditions for additional research using the specific dataset. This will provide the necessary information so that new research applications to utilize the data can already take the data application time and potential costs for re-using the data into consideration during the proposal phase, before the application is sent to the targeted call. The feasibility of disseminating this information in the proposal phase should be investigated. It should be noted that these procedures are often much more time- and resource-consuming than expected.

9.1 Application procedure for data access and usage

The application procedure shall address the following items (at least):

- **Application submission:** information regarding where and how to send in the application, as well as a contact person for questions regarding the application.
- **Information needed for evaluation:** the application procedure should specify the information that is needed to evaluate the application. This may include the definition of:
    - Approval authority: the identification of person/organisation approving an application
    - Response times and conditions: to be taken into account in the approval decision
    - Mandatory training: if there are any requirements for mandatory training in data protection and privacy issues
    - Non-disclosure Agreements (NDA): requirement for signing an NDA
    - Data-access procedure: requirements for data protection, including possible certification of data-protection implementation
    - Conditions for access and use of the data, including potential costs for data storage, access, support and research services
    - Acknowledgements on publications, reports and presentations
    - Documentation of data applications and the related approval decision(s).
- **Application form**: containing the necessary information about the intended use of data for the application.

9.2 Application form for data access & usage

The suggested information to be provided by the applicant for a decision within the set response time:

- **Applicant details**:
    - organisation(s) applying
    - contact person(s) for each organisation

    - project partners applying (when applicable) – list of partners that want data access for project analysis.

- **Project proposal**: applications may be required to submit a detailed or summarized project proposal that outlines the objectives, methodology, expected outcomes, and potential impact. Also including:
    - What are the expected research results?
    - What research questions are the data expected to answer?
    - What are the expected results?
- **Budget and funding**: applicants may need to provide evidence on approved budget and funding bodies behind the research project to outlines the expected costs of the proposed research, as well as information on exploitation plans and return of investment.
- **Ethics approval**: if the research project involves participation of humans or animals, or any other ethical need to be considered, then the application form should provide evidence of ethics approval from an external review or a compliance report on a provided ethics committee or guidelines.
- **Data Management Plan**: a DMP is a document that describes how data will be collected, processed, analysed, stored, shared, and preserved during the project. Including the DMP in the application form confirms the applicant's commitment on transparency and responsible data management. Expected content:
    - which dataset (if many available)
    - How is the data to be accessed?
    - specific data requested (time-series data, video, GPS, questionnaires, etc.)
    - list of persons to get access, and the related access time period.
- **Dissemination**: information on the intended publication of the data:
    - How will the results be disseminated?
    - What data, graphs, etc. are intended to be published?
- **Responsible use**: need for training in data protection and privacy issues:
    - Have the researchers had previous training? If so, what kind?
    - Is training related to data protection required, or only in data analysis setup at the data provider?
- **Need for support and research services**:
    - Level of knowledge of the concerned analysis tools? Using self-supplied tools or needing training on provided tools?
    - Other support needs for the analysis, such as extracting datasets, etc.?
    - Should the research facilities do part or all of the analyses/research?

9.3 Data Space Governance

A data space is built on the trust among its stakeholders. To ensure a common understanding of the purpose and use, technical undertakings (for data privacy, protection, and cyber security), a systematic onboarding process is needed. Also, if a partner decides to exit the data space, the implications must be made clear beforehand. The on- and offboarding process is described below.

9.3.1 Onboarding

Onboarding, in this context, refers to the systematic process of integrating new participants into the Federated Data Space (FDS). This process is an important part of FDS governance, as it sets the foundation for the entire data sharing relationship. It involves several steps, each with its own legal implications.

Description

The onboarding process typically begins with an application from the participant (e.g., a company, research centre, or public body) expressing interest in joining the FDS. This is followed by a review of the application by the governing body of the framework to ensure that the participant meets all necessary criteria.

Legal requirements for Onboarding

Whether a participant applies to join the FDS or is invited by the administration entity of the data space, they must comply with all relevant legal requirements. This process can be broken down into several steps:

- Invitation to join: The administration entity may invite a potential participant to join the FDS. This invitation can be sent via an official communication channel, such as an email, outlining the benefits and responsibilities of joining the data space.
- Creating a company profile: Upon accepting the invitation, the participant is directed to a secure login form where they can create a company profile. This profile includes information about the company, such as its name, size, industry, and role within the data space (e.g., data provider, data user, or both).
- Agreeing to Terms and Conditions: Before finalizing their profile, the participant must agree to the terms and conditions of the FDS. These terms outline the legal obligations of each party and include clauses related to data protection, privacy, intellectual property rights, and dispute resolution.
- Review and Approval: Once the company profile is complete and the terms and conditions are agreed upon, the administration entity reviews the application. If all legal requirements are met and the administration entity approves, the participant becomes an official member of the FDS.

By following these steps, participants can ensure they are in full compliance with all legal requirements of the onboarding process.

Data Sharing Agreements and Contracts

A key part of the onboarding process is the negotiation and signing of data sharing agreements and contracts. These documents outline the rights and responsibilities of each party, including how data will be shared, used, and protected.

Compliance with Data Protection and Privacy Laws

In Europe, participants must comply with the General Data Protection Regulation (GDPR), which provides strict guidelines for how personal data can be collected, stored, processed, and shared. Participants must demonstrate their compliance with these regulations as part of the onboarding process.

Technical Requirements for Onboarding

Finally, participants must meet certain technical requirements to ensure that they can effectively participate in the FDS. This may include demonstrating their ability to securely transmit and store data, as well as their ability to implement necessary data protection measures.

In conclusion, the onboarding process is a complex but necessary step in establishing a successful data sharing relationship. By carefully navigating this process, participants can ensure that they are legally compliant and technically prepared to share data in a secure and effective manner.

9.3.2 Offboarding

The offboarding process is equally important as the onboarding process in a FDS. It ensures that the participant's data and rights are protected even after they leave the framework.

Description

Offboarding begins when a participant decides to leave the FDS or when their membership is terminated by the administration entity. The process involves several steps, including notification of departure, data transfer or deletion, and legal closure.

Notification of departure

The participant must formally notify the administration entity of their intention to leave the framework. This can be done through an official communication channel, such as an email or a form on the framework's website.

Data Retention and Deletion

Data retention and deletion are critical aspects of the offboarding process. They involve several challenges and considerations:

- Data transfer: If the data sharing agreement allows for it, the data may be transferred back to the participant. This process must be done securely to protect the data during transit. It's important to ensure that all copies of the data in the framework's systems are accounted for during this process.
- Data deletion: If data transfer is not possible or desired, the data must be deleted. However, simply deleting data from active systems is not enough. The data may still reside in backups or logs, so these must also be purged.

- Proof of deletion: Providing proof of deletion can be challenging. While it's technically possible to generate logs or certificates of data deletion, these do not guarantee that all copies of the data have been deleted. For example, data could still exist in offline backups or could have been copied prior to deletion.

- Legal requirements: Data protection laws, such as the GDPR, have strict requirements for data deletion. For example, under GDPR, individuals have a "right to be forgotten," which means that their data must be permanently deleted upon request. Failure to comply with these laws can result in heavy fines.

- Technical challenges: From a technical perspective, secure data deletion is a complex task. Simply deleting a file does not remove it from a storage device; rather, it marks the space as available for reuse. Until that space is overwritten by new data, the original file can potentially be recovered. Therefore, secure deletion often involves overwriting the data with random information to prevent recovery.

In conclusion, while there are many challenges associated with data retention and deletion during offboarding, they can be addressed through careful planning, robust technical processes, and strict adherence to legal requirements.

Legal closure

The offboarding process also involves legal closure, which includes terminating the data sharing agreement and ensuring that all legal obligations have been met. This may involve a final audit or review.

Technical requirements for Offboarding

Finally, there are technical requirements for offboarding, which may include securely deleting the participant's account and any associated data from the framework's systems.

In conclusion, offboarding is a complex process that requires careful attention to both legal and technical details. By following a clear and thorough offboarding process, both participants and administration entities can ensure that they meet their legal obligations and protect their rights.

10 Conclusions

This document details the elements of a data sharing framework in order to facilitate re-use of the many CCAM/FOT/NDS datasets hosted at different locations globally. This framework can also facilitate data sharing within new projects, as the content of the framework is general and could be used whenever data sharing is performed. The framework introduces the new approach of federated data sharing, applied to a CCAM domain.

The CCAM Data Sharing Framework then consists of the following seven items: project documents (such as the consortium agreement and the consent form), description of data and metadata, data protection, training on data protection, support and research services, financial models for post-project funding and data governance as described in application procedures. All these components need to be considered for efficient data sharing framework.

The report constitutes the essence of the discussions held during the European support action projects FOT-Net 2, FOT-Net Data, CARTRE, ARCADE and FAME time frames. Through the discussions, it has become obvious that the recommendations apply to a wide variety of cases and research areas (applicable but not limited to automotive), including different national contexts. At the end, though, it is always up to the partners of the specific project, national or international, to select the appropriate data-sharing strategy and decide what parts of the data sharing framework are applicable to their project.

The CCAM Data Sharing Framework needs to be discussed and applied by different stakeholders with good knowledge of their national requirements, who collect experiences and make the framework applicable to as many countries as possible. The framework also needs to be updated as new technology and methods provide new possibilities, especially regarding anonymization and feature extraction. Reliable tools for automated feature extraction are key to be able to provide large quantities of essential features from video.

If the suggestions and requirements presented here are taken into consideration, future projects will be well prepared for data sharing during and after the project. Sharing data can be a win-win opportunity for both those who share and those who analyse data.

The Data Sharing Framework has been applied by many European projects and organizations since its release in 2016. The additions of this version will help to set the foundation for a data space to support future projects within connected, cooperative and automated driving.

List of abbreviations

Abbreviation	Full text
AD	Automated Driving
ADAS	Advanced Driver Assistance Systems
ASAM	Association for Standardization of Automation and Measuring Systems
B2B	Business to Business
C2C	Car to Car
C2X	Car to Infrastructure
CA	Consortium Agreement
CAM	Cooperative Awareness Message
CAN	Controller Area Network
CAS	Content-Addressable Storage
CCAM	Connected, Cooperative and Automated Mobility
C-ITS	Cooperative Intelligent Transport System
CPM	Collective Perception Message
DENM	Decentralized Environmental Notification Message
DMP	Data Management Plan
DOI	Digital Object Identifier
DOW	Description Of Work
DPIA	Data Protection Impact Assessment
DSL	Domain Specific Language
ECU	Electronic Control Unit
ENISA	European Union Agency for Network and Information Security
EOSC	European Open Science Cloud
ETL	Extract Transform Load

FAIR	Findable, Accessible, Interoperable, and Reusable
FASTR	US Fair Access to Science and Technology Research Act
FDS	Federated Data Space
FDBS	Federated Database System
FESTA	European handbook on FOT methodology
FOT	Field Operational Test
FRPAA	US Federal Research Public Access Act
GDPR	General Data Protection Regulation
GIS	Geographical Information Systems
GPS	Global Positioning System
GNSS	Global Navigation Satellite System
HDOP	Horizontal Dilution of Precision
HE	Horizon Europe
HIPAA	Health Insurance Portability and Accountability Act
HMI	Human Machine Interface
Hz	Hertz
IPR	Intellectual Property Right
ITIL	IT Infrastructure Library
ITS	Intelligent Transportation System
IVIM	In-Vehicle Information Message
KPI	Key Performance Indicator
LIDAR	Light Detection And Ranging
ML	Machine Learning
NAT	Network Address Translation
NDA	Non-Disclosure Agreement
NDS	Naturalistic Driving Study

OBD	On-Board-Diagnostics
OBU	On-Board Unit
ODD	Operational Design Domain
OEM	Original Equipment Manufacturer
OSI	Open Simulation Interface
PI	Performance Indicators
PII	Personal Identifiable Information
RDM	Research Data Management
RPM	Rounds Per Minute
RSU	Roadside Unit
RTP	Real-time Transport Protocol
RTCP	Real-time Transport Control Protocol
RTSP	Real-time Streaming Protocol
SQL	Structured Query Language
SRTP	Secure Real-Time Transport Protocol
TCP	Transmission Control Protocol
UDP	User Datagram Protocol
V2X	Vehicle to X (infrastructure)
VIN	Vehicle Identification Number
WebRTC	Web Real-Time Communication
XML	Extended Markup Language

List of Tables

List of Figures

List of references

FESTA, 2021. FESTA Handbook Version 8, 2021. Available at: https://www.connectedautomateddriving.eu/wp-content/uploads/2021/09/FESTA-Handbook-Version-8.pdf, accessed on November 27, 2023.

DSF, 2021. Data Sharing Framework v1.1, 2021. Available at: https://www.connectedautomateddriving.eu/wp-content/uploads/2021/09/Data-Sharing-Framework-v1.1-final.pdf, accessed on November 27, 2023.

C-ITS 2023, C-Roads & C2C CC: C-ITS & CCAM –The data layer for sharing experiences in validation, 2023. Available at: https://www.car-2-car.org/fileadmin/documents/General_Documents/C-ROADS_C2C_CC_C-ITS_and_CCAM_Data_paper_V1.0.pdf, accessed on February 4, 2024.

Gu, 2021, NXP, Volkswagen and Partners Continue to Accelerate the V2X Rollout, 2021. Available at: https://www.nxp.com/company/blog/nxp-volkswagen-and-partners-continue-to-accelerate-the-v2x-rollout:BL-THE-V2X-ROLLOUT, accessed on February 4, 2024.

Data strategy, 2020. A European strategy for data, 2020. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020DC0066, accessed on November 27, 2023.

European Directive 95/46/EC Art. 2., 1995.
Available at: http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:31995L0046&from=en , accessed on November 27, 2023.

GAIA-X, 2023. Gaia-X Architecture Document - 23.10 Release, GAIA-X European Association for Data and Cloud, 2023. Available at: https://docs.gaia-x.eu/technical-committee/architecture-document/23.10/, accessed on November 27, 2023.

IDSA, 2022. Reference Architecture Model, International Dataspace Association, 2022. Available at: https://docs.internationaldataspaces.org/ids-knowledgebase/v/ids-ram-4/, accessed on November 27, 2023.

X-Road, 2023. X-Road architecture, 2023. Available at: https://docs.x-road.global/Architecture/arc-g_x-road_arhitecture.html, accessed on November 27, 2023.

GDPR, 2016, REGULATION (EU) 2016/679. Available at: http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=EN, accessed on November 27, 2023.

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al, 2016. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018, 2016. Available at: https://doi.org/10.1038/sdata.2016.18, accessed on November 27, 2023.

Science Europe, 2021. Practical Guide to the International Alignment of Research Data Management - Extended Edition, 2021. Available at: https://doi.org/10.5281/zenodo.4915862, accessed on November 27, 2023.

Data contracts, 2022. Data contracts, 2022. Available at: https://learn.microsoft.com/en-us/azure/cloud-adoption-framework/scenarios/cloud-scale-analytics/architectures/data-contracts, accessed on June 19, 2023.

ISO 20546, 2019. Information technology — Big data — Overview and vocabulary, 2019. Available at: https://www.iso.org/obp/ui/en/#iso:std:iso-iec:20546:ed-1:v1:en, accessed on February 4, 2024.

SAE J3016, 2018, Surface Vehicle Recommended Practice, “Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles”, SAE Standard J3016, Rev. Jun. 2018

Roebuck K., 2012. Metadata Repositories: High-impact Strategies - What You Need to Know: Definitions, Adoptions, Impact, Benefits, Maturity, Vendors. Emereo Publishing.

Puglia, S., Reed, J., & Rhodes E., 2004. Technical Guidelines for Digitizing Archival Materials for Electronic Access. Report for U.S. National Archives and Records Administration (NARA). Available at: https://www.archives.gov/files/preservation/technical/guidelines.pdf, accessed on November 27, 2023.

Innamaa S. & Koskinen, S., 2017: Data Catalogue. FOT-Net Data D4.1. Available at: https://cordis.europa.eu/docs/projects/cnect/3/610453/080/deliverables/001-FOTNetDataD41DataCataloguev3.pdf, accessed on November 27, 2023

Gaia-X, 2022, Gaia-X Data Exchange Service Specification, 2022. Available at: http://docs.gaia-x.eu/technical-committee/data-exchange/22.10/dewg/, accessed on September 8, 2023.

The US Code of Federal Regulations including the HIPAA Privacy Rule at 45CFR Parts 160 and 164, 2012, available at: https://www.hhs.gov/sites/default/files/ocr/privacy/hipaa/understanding/coveredentities/De-identification/hhs_deid_guidance.pdf, accessed on February 4, 2024.

D' Acquisto, G., Domingo-Ferrer, J., Kikiras, P., Torra, V., de Montjoye, Y-A., & Bourka, A., 2015. Privacy by design in big data - An overview of privacy enhancing technologies in the era of big data analytics 1.0. Available at: https://www.enisa.europa.eu/publications/big-data-protection/at_download/fullReport

Nelson, G. s., 2015. Practical Implications of Sharing Data: A Primer on Data Privacy, Anonymization, and De-Identification. In: Proceedings of the SAS Global Forum 2015, Austin. Available at: http://support.sas.com/resources/papers/proceedings15/1884-2015.pdf, accessed on November 27, 2023

Annex I. Metadata documentation template

This template assists test sites in thoroughly documenting their datasets. This documentation aids in evaluating aspects like safety, efficiency, and the environmental impact of new systems. Detailed metadata facilitates data re-use, even by individuals who were not involved in the original data collection.

Where applicable, the fields align with the Dublin Core Metadata Initiative. For more details, visit https://www.dublincore.org/specifications/dublin-core/dcmi-terms/.

1. Dataset Summary

Title: [A name given to the dataset]

Description: [Short public summary for data catalogues]

Persistent identifier, URL: [Unique identifier, hyperlink]

Country: [The country or countries where the data was collected or created]

Location(s): [Main coordinates for map searches]

Test start date: [Date associated with the start of data collection or generation]

Test end date: [Date associated with the end of data collection or generation]

Creator: [The name of the entity primarily responsible for creating the dataset. This can be a person, and organization, or a service.]

Owner: [Organization with ownership rights to the dataset]

Keywords: [Keywords for use in search]

2. Administrative metadata

Dataset version number: [Version]

Date: [Last update]

Publisher: [Person who last updated the dataset and made it available]

Contact person(s): [Contact details regarding general inquiries and requesting access]

Access and sharing conditions: [Summary about access requirements and data sharing conditions]

Language: [Languages used in the dataset and its documentation]

3. Dataset categories for search and select

Number of vehicles:

Number of test subjects:

Public anonymous sample data available: [yes/no]

Vehicle type:

- ❑ Passenger cars (e.g., sedans, SUVs, electric cars)
- ❑ Heavy vehicles (e.g., trucks, lorries)
- ❑ Industrial vehicles (e.g., forklifts, excavators, tractors)
- ❑ Public transport (e.g., buses, automated shuttle buses, trams)
- ❑ Motorcycles & Two-Wheelers (e.g., standard motorcycles, scooters)
- ❑ Other: [please specify]

Tested system(s):

- ❑ Automated driving
- ❑ Integrated driver support system
- ❑ Aftermarket System (e.g., add-on navigation, advanced parking systems)
- ❑ Remote operations & control
- ❑ Connectivity & Telematics
- ❑ Other: [please specify]

Data logs contain:

- ❑ Unrestricted driving in extensive areas (e.g., a city or district, allowing for dynamic route selection based on needs)
- ❑ Driving along multiple predetermined routes
- ❑ Driving on specific test routes
- ❑ Raw data from environmental sensors
- ❑ Processed data on surrounding elements (e.g., detected vehicles, identified lanes)
- ❑ High-precision positioning data (e.g., RTK-GPS or equivalent)

Locations:

- ❑ Dense urban centres (e.g., city downtown)
- ❑ Broader urban areas (e.g., city suburbs)
- ❑ Rural landscapes
- ❑ Motorways or highways
- ❑ Mountainous or hilly regions with notable slopes

Weather Conditions:

- ❑ Heavy rainfall
- ❑ Foggy environments
- ❑ Snow-covered roads

Driving Conditions:

- ❑ Night-time scenarios
- ❑ Traffic congestion

4. Structural metadata

4.1 Log data details

Data storage format: [Database, CSV file, H5 file, XML etc.]

Other file formats used: [Video, Excel etc.]

Data field descriptions: [see example]

ID	Field name	Description	Unit and type	Sample rate	Minimum value	Maximum value	Value, if not available
1	vehicle_speed	Wheel speed sensor	km/h, double precision	10 Hz			

4.2 Manual annotations

Documentation of annotation process and details, e.g. a video annotation codebook in use

4.3 Self-reported data details

Data storage format: [Database, CSV file, H5 file, XML etc.]

Data description document: [Questionnaire template, diary template or similar]

4.4 Processed summary/aggregated data details

Documentation of several indicators and summaries that may have been processed and are provided with the data. For example, the data has been split automatically into trips and for each trip, average speed and maximum speed have been calculated.

4.5 Data quality

Describe how data quality has been ensured and checked, or describe the quality in general.

5. Study design and test execution as separate documentation

This or separate documentation should describe the study from test leader and evaluation perspectives. It is recommended to cover at least the following aspects:

- Summary information of the project and cooperation partners, duration, budget etc.
- Main goals and research questions
- Test site (if it was within a specified perimeter) and routes, including maps and photos
- Overall description of equipment used: vehicle fleet, functions, HMI, additional driver support in the vehicle (navigators, etc.) – preferably with links to videos demonstrating use. Please describe also infrastructure and communication elements.
- Users/operators, description of their selection criteria, recruitment processes and agreement templates
- How user contact was maintained during the study
- Possible grouping of participants (e.g., test groups 1 and 2 and a reference group)
- Test plan and execution. For example, what the participants were asked to do, how and when the briefing was given, when and what questionnaires were administered or what interviews were conducted? How many times some test run or route was repeated?
- Test diary including dates, study phases and commentary (it is possible to use the test diary template, see 5.1). The diary should note special events and changes that

may affect data analysis (e.g., roadwork, strikes, economical changes, special weather).
- Information on how safety and privacy were ensured. Information on required permits to operate, if there were some.

5.1. Test diary example

Test site:

Contact:

Date	Planned hours	Actual hours	Test focus or route	Manual/Automated
1.1.2023	10–14	10–14	Mission 2	Automated
2.1.2023	10–14	–	Mission 2	Automated
3.1.2023	10–14	10–13	Mission 2	Manual
…				

Weather	**Road conditions** (dry/wet/snow/icy)
15 degrees, rainy and foggy	Wet
13 degrees	Dry
–2 degrees, snowfall	Snow
…	

Availability and maintenance	**Notes** (for example, stopped early, broken systems, accidents)
Out of operation 15 minutes due to sensor recalibration	
Out of operation 4 hours	Battery problems, tests cancelled
Cleaning 15 minutes	Stopped 1 hour early due to X

Annex II. Data processing and sharing agreement template within project consortia

For projects with low sensitivity data, this template supports data owners in sharing data within the consortium. Given the non-critical nature of the data, comprehensive agreements may be unnecessary. However, basic GDPR compliance must still be met.

Data Processing Guidelines and Agreement

[Your project] – [Your test site]

Data owner main contact: [Name], [Organization]

Purpose of data collection and processing

Log data has been acquired from [Specific Source, e.g., "automated vehicles"] for the purpose of the project's evaluation. The evaluation covers safety, efficiency, environmental and societal impacts. The data is not to be used to single out individual persons in analyses but evaluation results are to cover averages. Further details of evaluation goals can be found in [test plan report]. (In accordance with GDPR Article 30 on the documentation of processing activities)

Consent & Notification

Explicit consent has been obtained from participants, allowing for the sharing and analysis of the collected data.

Data collection processes have been clearly communicated with relevant stakeholders. Notices were visibly displayed, and necessary personnel were duly informed. (In line with GDPR Article 13 on transparency obligations)

Confidentiality & Risk Assessment

While parts of the dataset may contain personal or confidential data, the nature of this data poses minimal risk even in the event of unauthorized access. Nevertheless, the data is shared as confidential, with stricter guidelines available in the project consortium agreement.

In the event of a data breach, significant harm appears unlikely due to the nature of the data and the testing environment. Despite this, all data is well-protected against unintended leaks. (In relation to Data Protection Impact Assessment, as per GDPR Article 35)

Data Sharing

Data is exclusively accessible to named evaluation experts, shared under the confidentiality clause of the project consortium agreement. Only the data owner possesses rights to distribute the dataset.

Deletion

All data will be deleted no later than a year after project completion. Only certain mutually agreed data will be retained longer, exclusively for specific purposes, like presentation video clips.

Acknowledgment and Agreement

Members of the consortium acknowledge and agree to the terms laid out above:

[Name, Email] _____ Date: _____

[Name, Email] _____ Date: _____

Annex 3. Standardisation

Adoption of standards could significantly streamline data sharing in the CCAM research domains by reducing the complexity of documenting metadata. However, it remains uncertain which specific standards will become widely adopted. There are standards directly or in-directly within CCAM, e.g., traffic information (DATEX II) and C-ITS (IEEE 802.11p, IEEE 802.11bd, ETSI ITS G5).

Large European CCAM project L3Pilot published its log file format as open source (https://www.connectedautomateddriving.eu/data-sharing/l3pilot-common-data-format/) and it has since been utilized, and further developed, by the follow-up project Hi-Drive. Similar project-level definitions have become common for effectively sharing and working with data.

ASAM

This section provides information about standards related to data formats defined in OpenX standards of the ASAM e. V. (Association for Standardization of Automation and Measuring Systems). The family of OpenX standards (simulation branch) within ASAM is devised to cover most of the domains involved in testing and evaluation of ADAS and ADAS, spanning through the entire value chain: sensor data, labelling, scenario generation, high-definition maps, simulation interfaces, etc.

The following table summarizes the scope of each standard project within the simulation branch of ASAM.

Table: ASAM simulation branch

Format	Description
ASAM OpenScenario 1.1	Definition of concrete scenarios in XML format
ASAM OpenScenario 2.0	Extension of definition of scenarios using DSL (Domain Specific Language) and extending description to abstract and concrete levels
ASAM OpenDrive 1.6	Definition of high-definition maps (lane-level roads)
ASAM OpenCRG 1.2	Definition of a file format for the description of road surfaces
ASAM OSI	OSI (Open Simulation Interface) is a specification for interfaces between models and components of a distributed simulation
ASAM OpenLABEL 1.0	Definition of object-level and scenario-level labelling approach and data format
ASAM OpenXOntology 1.0	Common semantic base for all OpenX standards, materialized as a set of ontologies
ASAM OpenODD 1.0	Definition of a data format that can represent an abstract ODD for a vehicle

ASAM Test Specification Study group	Definition of interaction between scenario definition and test execution

Other related standards from ASAM can also be considered relevant for the scope of testing and validation, depending on the function under test.

Table: ASAM formats for testing and validation

Format	Description
ASAM MDF	Measurement Data Format, for the specification of data format as container for sensor data
ASAM XIL API	Specification of interfacing for XiL (X-in-the-loop) frameworks
ASAM ATX	Automotive Test Exchange Format for the specification of standardised XML format to enable exchange of test data between systems
ASAM ODS	Open Data Services for the specification of mechanisms for persistent storage and retrieval of testing data

ASAM actively aligns internal standard development projects with other standardisation bodies and initiatives, aiming to foster common understanding of gaps, needs and interoperability issues. The following figure summarizes the domains where alignments of the OpenX standards with other standards have been identified.

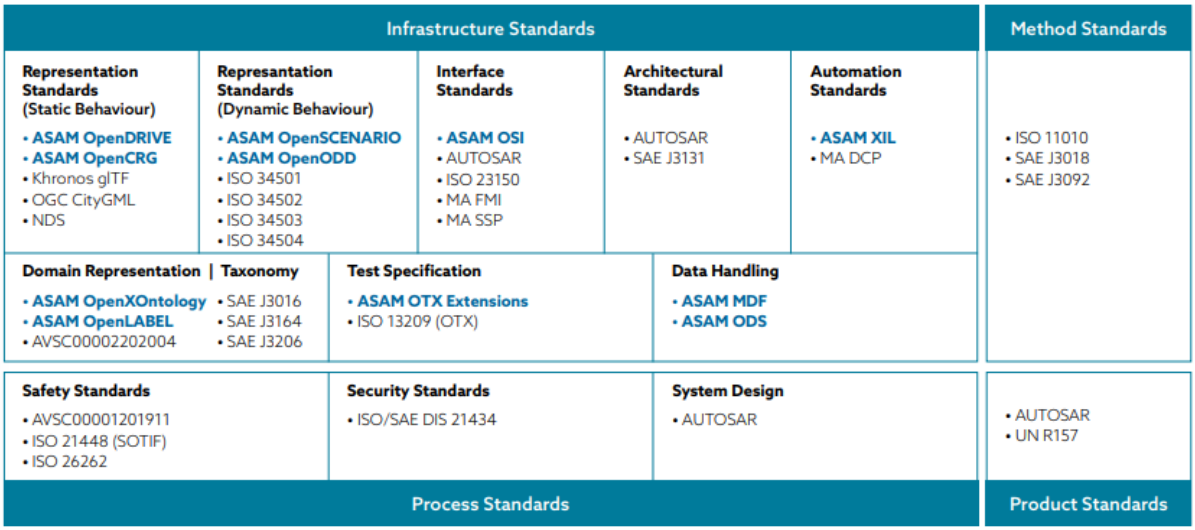

Figure: ASAM standards and their relation to other international standards.

In addition, the OpenX standards can be allocated in the context of the essential components of a simulation environment (see next figure) which are needed to cover the major

requirements for the various tasks that result from the activities of a DevOps cycle (e.g., Plan, construct, build, test, release, deploy, operate, monitor).

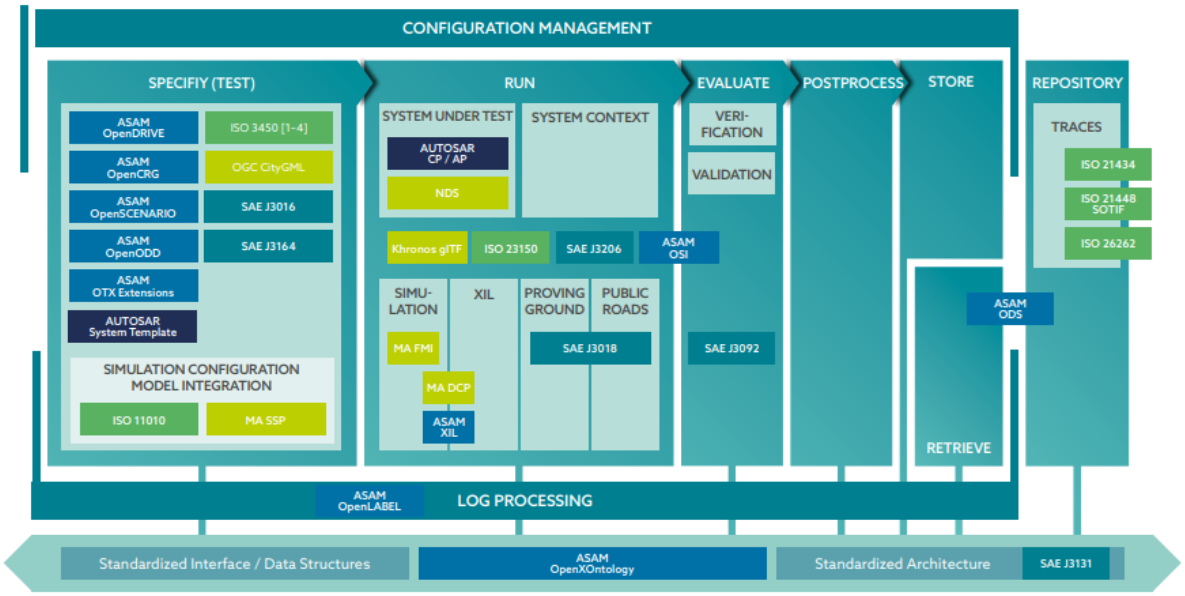

Figure: Reference architecture of simulation environments.

ASAM Open Source Tooling Platform (aka asam-oss, https://github.com/asam-oss) is an open source initiative started from ASAM and encouraged by industrial members to create tools to support the implementation, training and use of the ASAM OpenX standards. In that sense, the asam-oss has been implemented to host and share ASAM compatible tooling.