Ref. Ares(2025)6964385 - 28/08/2025

ROADVIEW

Robust Automated Driving in Extreme Weather

Project No. 101069576

Deliverable 4.10

Report on data privacy and storage

WP4 – Data logging, filtering, annotation, and quality assurance

Authors	Ian Marsh (RISE) and Stefan Parzefall (AVL)
Lead participant	RISE
Delivery date	28 August 2025
Dissemination level	Public
Type	Report

Version 1.0

Co-funded by the European Union

Innovate UK

Co-funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Climate, Infrastructure and Environment Executive Agency (CINEA). Neither the European Union nor the granting authority can be held responsible for them. Project grant No. 101069576.

UK participants in this project are co-funded by Innovate UK under contract No.10045139.
Swiss participants in this project are co-funded by the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract No. 22.00123.

Revision history

Author(s)	Description	Date
Ian Marsh (RISE), Stefan Parzefall (AVL)	Initial document created	28/02/2025
Yuri Poledna (THI)	Review	19/08/2025
Heikki Hyyti (FGI)	Review	19/08/2025
Ian Marsh (RISE), Stefan Parzefall (AVL)	Final version	26/08/2025
Johannes Ripperger (accelCH)	Final checks and formatting	28/08/2025

Contents

List of Figures

List of Tables

Partner short names

HH	Hogskolan Halmstad
THI	Technische Hochschule Ingolstadt
RISE	RISE Research Institutes of Sweden AB
FGI	Maanmittauslaitos – Finnish Geospatial Research Institute
accelCH	accelopment Schweiz AG
WMG	The University of Warwick
AVL	AVL Software and Functions GmbH

Abbreviations

AWS	Amazon Web Services
D	Deliverable
DGT	Dynamic Ground Truth
EOSC	The European Open Science Cloud
EU	European Union
FAIR	Findable, Accessible, Interoperable, and Reusable
GDPR	General Data Protection Regulation
GNSS	Global Navigation Satellite System
TARA	Threat and Risk Analysis
WP	Work Package

Executive summary

This deliverable reports on the data storage and privacy issues with respect to data management in the ROADVIEW project. It is a report on the data upload, data security, and storage, including post-project preservation and sharing. For one data provider within ROADVIEW, AVL (of three large contributions), we have investigated how the data transfer from vehicle to cloud is performed. A Threat and Risk Analysis has been done and forms the first, and most important part of this deliverable. A second portion deals with storage we used, quantities, speeds and the often overlooked, post project storage and sharing.

Objectives

Simply, to manage data between data gatherers and data consumers, safely, securely and within budget. Specifically, Task 4.6 and D4.10 is to report on the privacy of the data on the vehicles and movement to the RISE ICE data server (called ICE). Ease of use regarding the data during the project, no data leakage and a Threat and Risk Analysis (TARA) should be performed. That is any loss of data if parts, or the whole system is comprised. Counter measures are discussed should data breaches occur. AVL has proposed a data flow, which is outlined below. This is laid out in the Grant Agreement.

Methodology and implementation

The main methodology is to study the flow of data in the project. Figure 1 depicts the proposed workflow as per the description of Task 4.6 and how it was implemented within the scope of the project. Due to multiple data logging partners, each of which using their individual sensor suite and data logging system, following the proposed workflow in some steps has not proven to be feasible. The variety of data types in different format and length makes an automated data validity check impossible. Instead, each partner did the post-processing (marked in blue) of data locally, only uploading already validated data to the RISE server and therefore making the intermediate storage server as well as the validity check obsolete (steps marked in red).

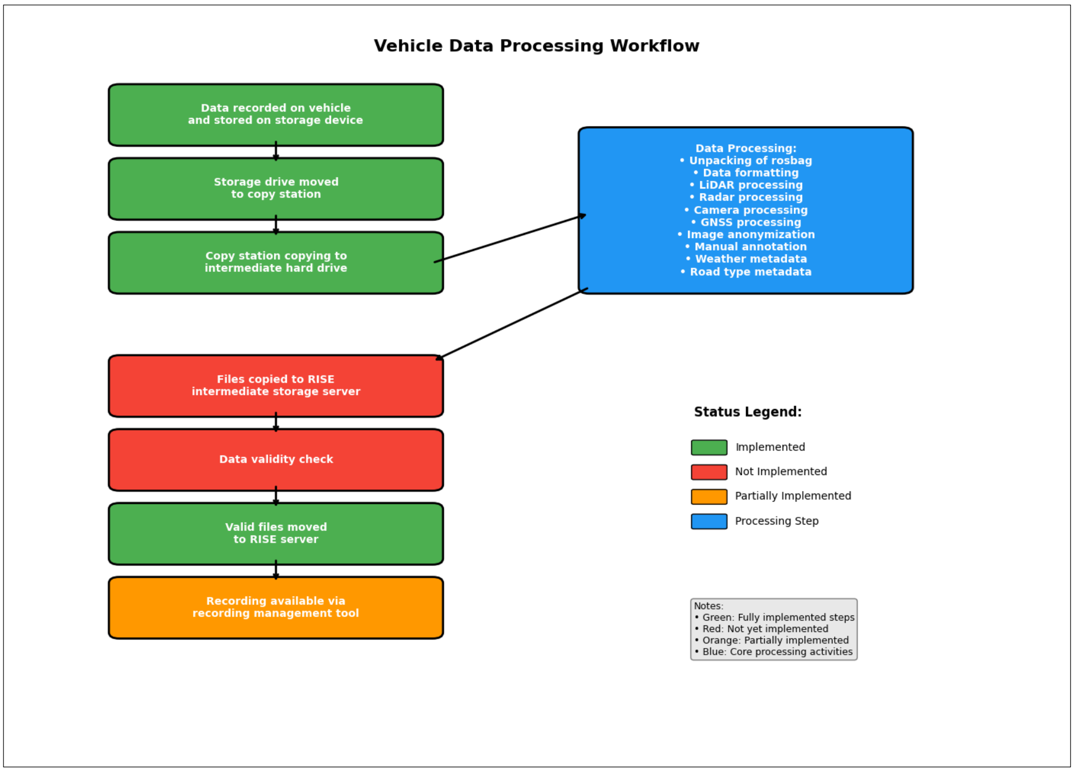


Figure 1. Flow diagram to visualize the vehicle data processing workflow with respect to the description of Task 4.6.

1 Data Privacy (AVL)

To satisfy the purposes of the ROADVIEW project, especially within the framework of WP4, there has been considerable real-world data logging activity. While most of the data such as lidar, radar or GNSS does not easily allow to draw conclusions on natural persons that might have been part of the driving scenarios, the camera data is still a risk with regards to data privacy. In this chapter we will discuss data logging in the context of privacy with a special focus on General Data Protection Regulation (GDPR) compliance, followed by a Threat and Risk Analysis (TARA) to assess the damaging scenarios for the project.

1.1 GDPR Compliance Considerations

The processing of personal data within the European Union is governed by the General Data Protection Regulation (GDPR, Regulation (EU) 2016/679 [1]), which establishes a uniform legal framework for data privacy and protection. Any project involving the collection, storage, or dissemination of potentially identifiable information - such as video recordings, vehicle telemetry, or sensor data from public environments - must adhere to GDPR principles including lawfulness, fairness, transparency, purpose limitation, data minimization, accuracy, storage limitation, integrity, and confidentiality (Art. 5 GDPR).

In the context of this project, particular attention must be paid to the handling of camera data, as such recordings may capture identifiable individuals (e.g., pedestrians, drivers) or vehicle license plates in public spaces. To ensure compliance, the consortium implements a multi-step anonymization process and performs a TARA as required under Article 35 GDPR for processing operations that are likely to result in high risks to the rights and freedoms of natural persons.

Furthermore, technical and organizational safeguards, including access control and structured data handling workflows, are applied in accordance with Article 32 GDPR ("Security of processing"). Any data intended for long-term publication is subjected to rigorous anonymization and verification procedures to ensure that no personal data can be reconstructed or re-identified. The consortium also ensures that the principles of privacy by design and by default (Art. 25) are embedded throughout the system architecture and data pipeline.

1.2 Threat and Risk Analysis

1.2.1 Concept

A Threat and Risk Analysis (TARA) is a structured methodology used to identify, evaluate, and mitigate potential security threats and vulnerabilities that could compromise the confidentiality, integrity, or availability of systems or data. TARA plays a central role in the design and operation of systems that process sensitive or safety-critical information, especially in domains such as automotive engineering, cybersecurity, and data protection. The goal of TARA is to systematically assess how an individual might exploit system weaknesses, what consequences such an attack might have, and which countermeasures can effectively reduce the associated risks to an acceptable level.

The TARA process is typically composed of several key steps: (1) asset identification, (2) threat modeling, (3) vulnerability assessment, (4) risk evaluation (often using a risk matrix) and (5) definition of mitigation strategies. Risks are commonly rated by estimating both their likelihood and their impact, allowing for prioritization. In EU-funded projects involving personal or safety-relevant data, TARA is especially useful for ensuring alignment with data protection regulations (e.g., GDPR), cybersecurity requirements (e.g., ISO/SAE 21434), and functional safety standards (e.g., ISO 26262).

The findings serve as the foundation for implementing technical and organizational safeguards, such as encryption, access control or anonymization.

1.2.2 Application

First, a summary of the data collection and processing is provided to identify what kind of data is collected and the purpose of the recorded data. Afterwards we'll assess the related threats and risks before finally the mitigation measures are presented.

1.2.3 Overview of Data Collection and Processing Workflow

Within the scope of the project, real world data is collected using the AVL Dynamic Ground Truth Box, which captures high-resolution lidar and camera data in real-world traffic scenarios. These recordings may include other vehicles, license plates, and pedestrians in public spaces, which constitutes personal data under the GDPR.

The data is:

- Collected on-board during test drives.
- Unpacked and anonymized using the open-source tool anonymizer by understand.ai.
- Uploaded to a secure server hosted by RISE.
- Used within the consortium to train and validate AI models.
- Published as a closed dataset for future research and benchmarking purposes.

The key risk area concerns the camera data, which inherently poses privacy risks if not properly anonymized. The following Threat and Risk Analysis identifies and evaluates those risks and outlines mitigation strategies.

1.2.4 Identified Threats

Table 1 outlines the identified potential threats that may arise from the data collection, post-processing and transfer from the data logging partners to the consortium.

Table 1. Threat analysis and description of threats.

ID	Threat	Description
T1	Incomplete or failed anonymization	Faces, license plates, or other identifying elements remain visible due to algorithmic or configuration failure.
T2	Unauthorized data access during transfer	Data is intercepted, accessed, or modified during transmission between systems.
T3	Unauthorized access to stored data	Improper access controls or compromised credentials allow unauthorized access to raw or processed data.
T4	Incorrect configuration of anonymization tool	Misuse or misconfiguration of the anonymizer leads to insufficient de-identification.
T5	Re-identification through metadata	GPS coordinates, timestamps, or contextual data enable inference of identity or behavior.
T6	Accidental disclosure of raw data	Non-anonymized or intermediate files are stored or shared by mistake.
T7	Publication of insufficiently anonymized data	Final dataset contains traces of personal information despite anonymization procedures.

1.2.5 Risk assessment

Risk assessment in the context of the complete ROADVIEW project is already part of WP1 with its own deliverables (D1.2, D1.3 and D1.4) and one of the already identified foreseen risks in WP4 (Risk Nr. F8) has been the compromise in data privacy since the very beginning of the project. Therefore, the consortium has been aware of this issue ever since the start of the project and has been working to find counter measures at the same time. Nevertheless, in the context of this deliverable and the performed TARA, a more detailed risk assessment with focus on data privacy has been performed and summarized in Table 2.

Table 2. Risk assessment, potential impact and risk level analysis.

ID	Risk	Potential Impact	Likelihood	Risk Level
R1	GDPR violation due to personal data exposure	Legal sanctions, reputational damage, potential project funding risk	Low - Medium	Medium
R2	Data leakage during transmission or storage	Loss of confidentiality, data integrity issues	Low	Low
R3	Re-identification through auxiliary data	Violation of data minimization principle	Low - Medium	Medium

1.2.6 Mitigation Measures

Finally, potential mitigation measures with respect to each of the previously identified threats will be presented in Table 3 to ensure avoidance of the assessed risks and their respective impacts.

Table 3. Potential mitigation measures per threat ID.

Threat ID	Mitigation
T1	Implement automated quality assurance checks post-anonymization, including manual visual inspection on sampled frames to confirm successful de-identification.
T2	Employ end-to-end encryption for data transmission.
T3	Enforce strict role-based access control and two-factor authentication. Regularly review access rights.
T4	Establish standard operating procedures and configuration presets for the anonymization tool. Ensure staff is trained in tool usage.
T5	Remove or obfuscate sensitive metadata (e.g., GPS data, timestamps) before public release. Apply data minimization principles.
T6	Use automated scripts to clean temporary or non-anonymized files after each processing step.
T7	Conduct a final compliance audit prior to dataset release. Create a GDPR checklist. Thorough anonymization checks.

2 Data Acquisition and Transfer Process

The data acquisition and transfer workflow implemented in this project adheres to the process proposed in the description of Task 4.6, with necessary extensions to accommodate data preprocessing requirements. Data collection is performed using the AVL Dynamic Ground Truth (DGT) Box, which records high-resolution sensor data, primarily lidar and camera streams, during real-world vehicle operation. The raw data is written directly to an in-vehicle storage device, specifically a B-Plus BRICK Logger equipped with 12 TB of local storage capacity.

After completion of a recording session, the storage drive is physically transferred to a copy station, where the data is offloaded for further processing. While in the original workflow immediate upload to an intermediate RISE server was foreseen, this step has been replaced with a local data processing stage, as pre-upload transformation is required to meet format and privacy specifications.

During this processing phase, the recorded data, initially stored in ROS1 bag format (rosbag), is unpacked and the camera images are anonymized using the open-source tool anonymizer by understand.ai. This step is essential to ensure compliance with the GDPR by masking or removing any personally identifiable information captured in public traffic environments. Further post-processing of additional sensor data (e.g., lidar, GNSS) is performed but documented separately in a dedicated deliverable (D4.4) which will be submitted at the same time as this deliverable.

Following anonymization, a data validity check is conducted based on the threat scenarios identified in the Threat and Risk Analysis. This includes a manual review of sample images to confirm that anonymization was successful and that no residual personal data remains.

Once validated, the processed and anonymized dataset is uploaded to the RISE ICE server, where a dedicated simple storage service (S3) bucket is created for each recording. Data upload is performed via Amazon Web Services (AWS) command-line tools, using secure and access-controlled channels [4].

While no fully integrated recording management tool is currently in place, a structured naming convention of available buckets provides registered users with access to individual recordings.

3 Data Storage (RISE)

3.1 A project perspective

The storage server in ROADVIEW is central to holding all data that needs to be *shared*. We use a storage facility, or data lake in the North of Sweden (see below, Section 3.3). Data is uploaded from different partners, in different formats (see below). Using a programmatic interface, it is also possible to 'mount' the file system, or even via an API key to access individual files. Some applications simply read data, process them and write outputs on local files systems. The storage (S3) is writeable, and we allow users to write files, including temporary outputs.

3.2 Data formats

3.2.1 ROS bags

ROS bags are a file format used in the Robot Operating System (ROS), a flexible and open-source framework for developing robotics software. ROS bags provide a way to record, store, and play back data generated by ROS nodes during the operation of a robot or a simulation. This data can include sensor readings, messages exchanged between nodes, and other types of information generated within the ROS environment.

ROS bags use the .bag file extension and are designed to be an efficient and flexible format for storing large amounts of time-stamped data. They are especially useful for:

- Logging data: ROS bags can be used to record sensor data and other information during robot operation, which can be later analysed, visualised, or processed for various purposes such as debugging, testing, or performance evaluation.
- Simulation and testing: Recorded data in ROS bags can be used to replay specific scenarios or situations, enabling developers to test algorithms or tune parameters in a controlled and repeatable environment.
- Sharing data: ROS bags provide a standardised format for exchanging data between different researchers, institutions, or projects. This makes it easier to collaborate, reproduce experiments, and compare the performance of different approaches.

To work with ROS bags, you can use the rosbag command-line tool that comes with ROS. Some of the common functions in CLI format are below in `Consolas` format.

- `rosbag` **`record`**: Records data from specified topics into a new .bag file.
- `rosbag` **`play`**: Plays back the data stored in a .bag file, effectively reproducing the recorded messages and their associated timestamps.
- `rosbag` **`info`**: Displays metadata and summary information about a .bag file.
- `rosbag` **`filter`**: Extracts specific messages from a .bag file based on user-defined criteria.

In addition to the command-line tool, there are various tools and libraries available for visualising and processing data from ROS bags, such as RViz (a 3D visualisation tool for ROS) and the Python API provided by the rosbag package.

3.3 The RISE ICE Cloud

One of the key features of the ICE data centre is its use of innovative cooling solutions. The facility takes advantage of the cold climate in Luleå, Sweden to cool the servers with outside air, significantly reducing the energy required for cooling. This makes the data centre more environmentally friendly and cost-effective, as it minimises energy consumption and operational costs.

The ICE data centre at RISE in Luleå serves as a testbed and research platform for various projects related to sustainable computing, including:

- Energy-efficient data storage and processing
- Renewable energy integration and smart grid solutions

- Edge computing and distributed data centres
- Cloud computing and resource optimisation
- Data centre security and resilience

By providing a state-of-the-art facility for the research and development of green data storage technologies, RISE in Luleå contributes to the global efforts towards more sustainable and energy-efficient IT infrastructure.

RISE offers several storage solutions, one is S3-Ceph buckets. We use Multi-Factor Authentication (MFA) for GUI access to the storage. Command Line Interface (CLI) is authenticated using API keys, similarly when programming directly on the storage. RISE ICE has upgraded the Ceph storage cluster with NVMe and has moved Ceph RBD based persistent storage volumes from hard disks to NVMe which makes it > 50x faster. The CephFS (rook-ceph-fs) volumes are on hard disks, however due to the large data amounts stored.

Note also, although the data processing is available on the RISE ICE server, we chose to process our data on each partners on-prem hardware. Mostly due to cost, but also to familiarity of locally stored data and latency back and from Northern Sweden.

3.4 A users' view

The RISE ICE server has been the main repository for sharing data within the project. Figure 2 shows all RISE ICE's services.

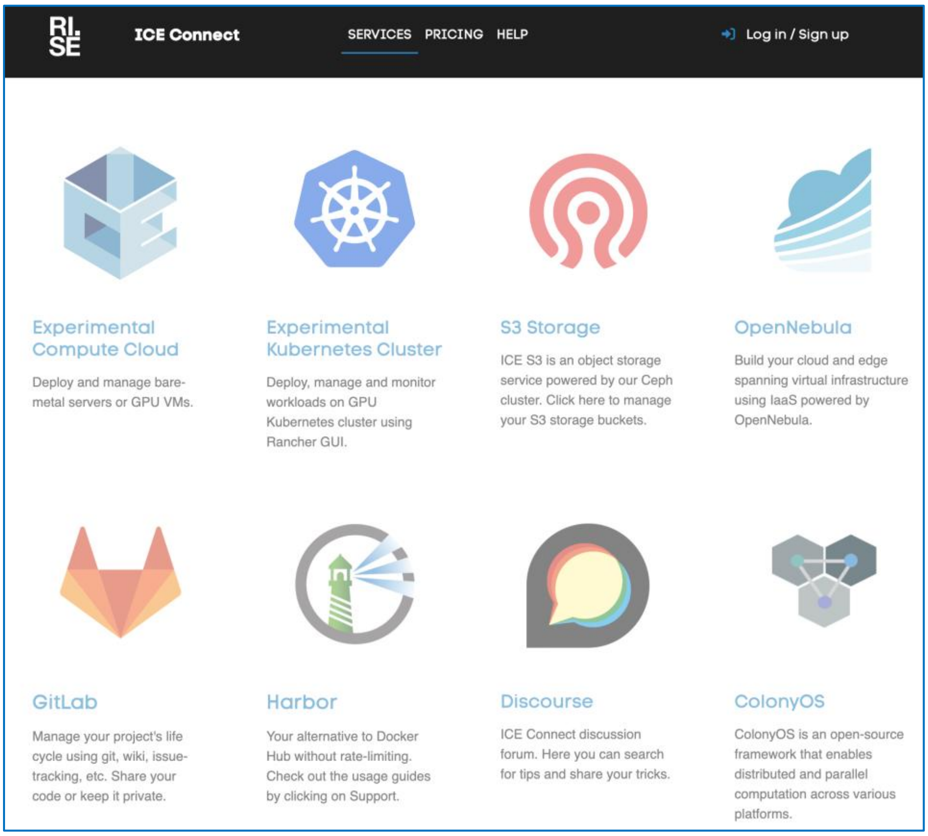


Figure 2. Dashboard view when logging into the RISE server.

In ROADVIEW, we exclusively use the S3 Storage, see Figure 3 below.

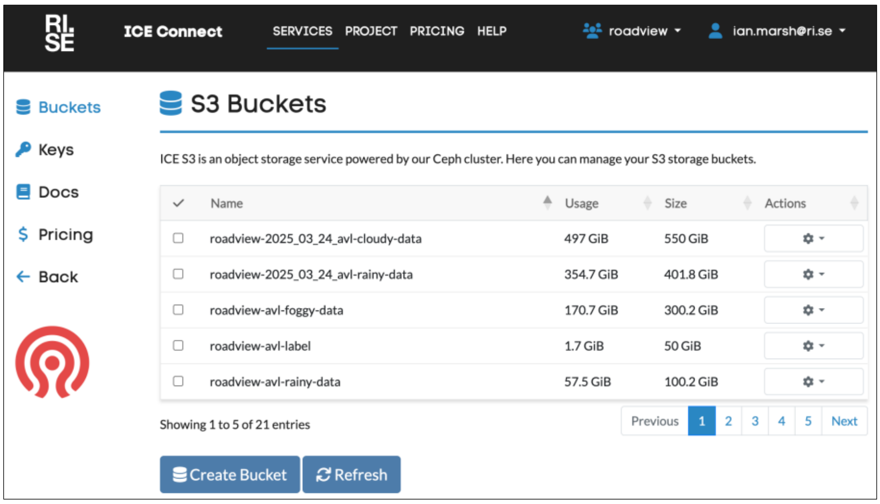


Figure 3. S3 storage view for ROADVIEW.

3.5 The S3 storage format

S3, or Amazon Simple Storage Service, is a scalable cloud storage service provided by Amazon Web Services (AWS). It does not have a specific "S3 data format." Rather, it allows you to store and retrieve various types of files, objects, or data in a highly available and durable manner. The files and objects stored in Amazon S3 can be in any format, such as text, images, videos, or binary data. We use S3 to store datasets for collaboration.

Amazon S3 organises data in a hierarchical structure with the following components:

- Buckets: These are top-level containers that store your data objects. Each bucket has a unique name within the S3 service, and you can create as many buckets as needed.
- Objects: Objects are the individual data files or items stored in S3 buckets. Each object consists of a key (a unique identifier within a bucket), data (the actual content of the object), and metadata (additional information about the object).

Amazon S3 supports various features such as versioning, lifecycle policies, access control, and data transfer acceleration. You can interact with Amazon S3 through the AWS Management Console, AWS CLI, SDKs, or RESTful APIs. They are suitable for ROADVIEW due to speed, relatively simplicity and the ability to add Access Control Lists (ACLs) to Buckets. Our S3 can be accessed via a GUI or simply a CLI command. Both are authenticated and secure.

3.6 How fast is the RISE data centre?

In terms of data upload an initial speed test upload to the ice data centre showed that 650MB was uploaded in about 9 secs. 4.3 GB worth of data took about 140 seconds. The dataset from FGI is about 42 Gig and after processing, 60 Gig. Note a merged 30 second video requires about 2 Gig. It has been processed via anonymisation.

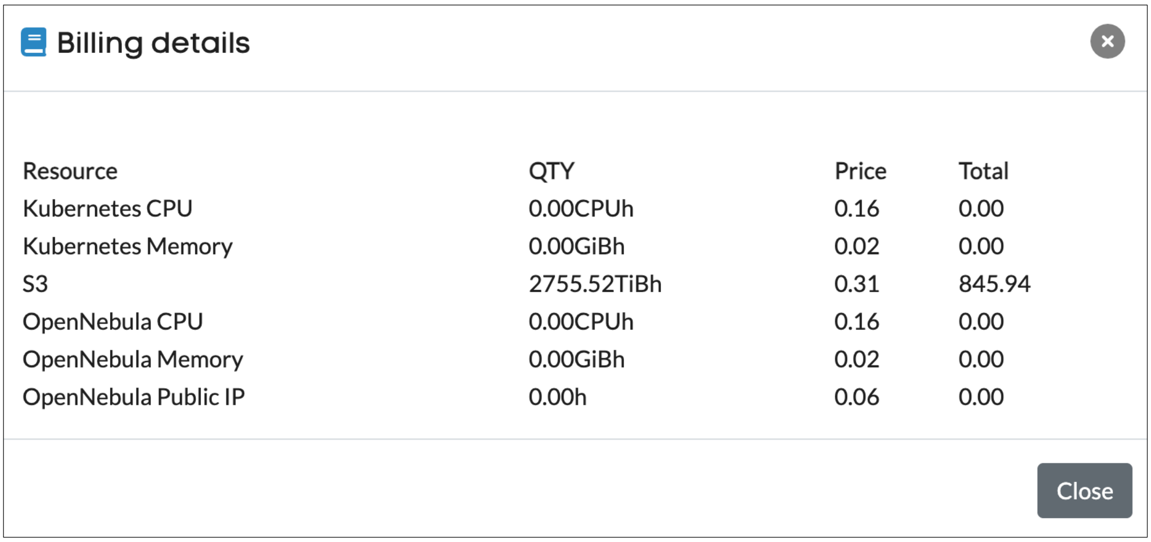
No radar data is included in the FGI dataset. The S3 data format, developed by Amazon, is used in the centre, albeit an open-source implementation thereof called Ceph. After some testing, Stockholm to ice.ri.se (S3 buckets) 650 MB in 9 seconds. RISE added Non-Volatile Memory Express (NVMe) which sped up access.

3.7 Costs and calculations

The idea of the calculation in Table 4 is to estimate the number of Terabytes available for the costs allocated.

Table 4. Storage units, costs, and formats: SEK to Euro conversion as of August 2025, (~11 crowns to 1 Euro). More at ice.ri.se (under-menu "pricing"). A total of 15k Euro is available to ROADVIEW, September 2022 to August 2026. Note, no computation is included in this cost, it is 'solely' storage.

Unit	Calculation	Result	Notes
TB	0.31 SEK per TiB	0.478	S3 costs. @ 2025 costs
Project time left	1.0 year * 1860	1860 hours	1 year remaining
Data costs in SEK	1860 hours * 0.31	576 SEK	Cost per TB for time left
Convert to Euro	576	52 Euro	Cost
Project allocation	15k Euro	**288TB.**	Dividing 15k Euro / 52 Euro


Billing details

Resource	QTY	Price	Total
Kubernetes CPU	0.00CPUh	0.16	0.00
Kubernetes Memory	0.00GiBh	0.02	0.00
S3	2755.52TiBh	0.31	845.94
OpenNebula CPU	0.00CPUh	0.16	0.00
OpenNebula Memory	0.00GiBh	0.02	0.00
OpenNebula Public IP	0.00h	0.06	0.00

Close

Figure 4. Costs in TiBh and consumption.

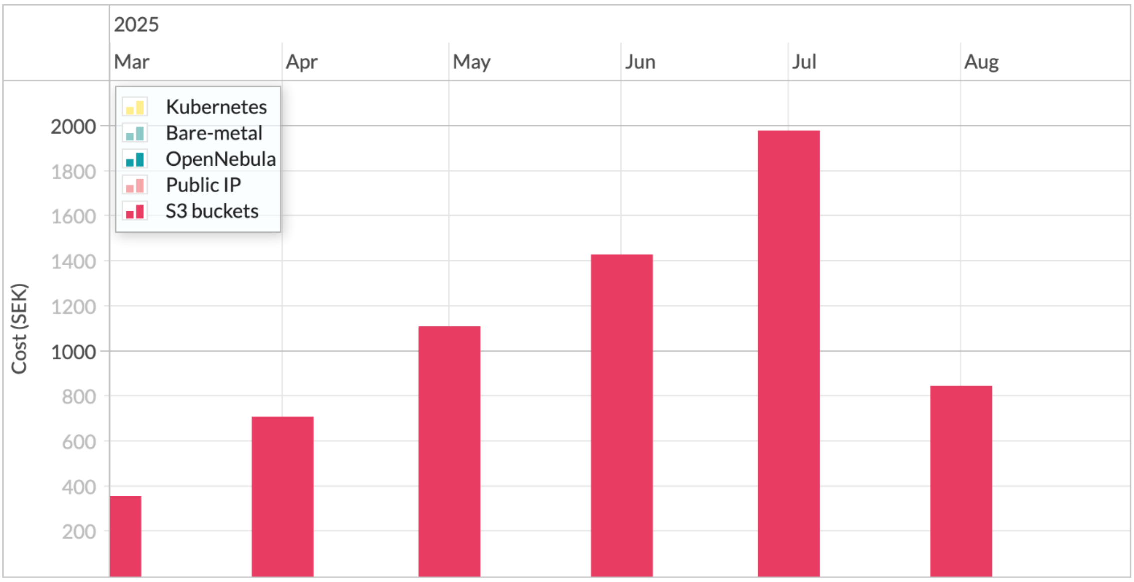


Figure 5. Example storage *only* costs for ROADVIEW (March-August 2025).

S3 Quota

Number of buckets	50
Total Bucket Size	32 TiB

3.8 A European Perspective (FAIR) datasets

The FAIR principles from the EU refer to guidelines for improving the Findability, Accessibility, Interoperability, and Reusability of digital assets—particularly research data. They aim to ensure data can be shared, reused, and integrated across disciplines and platforms, in both human and machine-readable ways. Here's a summary of each principle:

1. **Findable**
    a. Data should be easy to find by both humans and computers.
    b. Use of persistent identifiers (like DOIs) and rich metadata.
2. **Accessible**
    a. Once found, data should be retrievable using standardized protocols.
    b. Metadata should remain accessible even if the data itself is restricted.
3. **Interoperable**
    a. Data should use standard formats and vocabularies.
    b. It should integrate well with other datasets and tools.
4. **Reusable**
    a. Data should be well-described with clear licenses and provenance.
    b. It should be usable in future research beyond its original purpose.

The **EU promotes FAIR principles** to enhance open science, innovation, and collaboration across sectors. They are not about open data necessarily but about **making data as open and useful as possible**. The REHEARSE dataset has been judged FAIR.

3.9 EU laws

3.9.1 AI ACT

The EU AI Act (*Regulation (EU) 2024/1689*) is the world's first comprehensive regulation specifically governing AI. It entered into force on **1 August 2024**, with phased implementation rolling out over the coming months and years. Its mission: to build a **trustworthy, human-centric AI ecosystem** across Europe—balancing safety, rights, innovation, and global leadership. Its core objective is to ensure that data generated by connected products and related services—such as industrial machinery, vehicles, and IoT devices—can be accessed and shared fairly by users, rather than being controlled exclusively by manufacturers or service providers. The regulation sets out obligations for data holders to make such data available under transparent, non-discriminatory, and reasonable terms, while also defining safeguards to protect trade secrets, privacy, and cybersecurity. In addition, it introduces mechanisms for business-to-government data sharing in cases of exceptional need, such as public emergencies, and includes provisions to facilitate switching between cloud and edge service providers. By addressing barriers to data portability and reuse, the Data Act is intended to foster a more competitive and interoperable data economy across Europe, supporting innovation in sectors like mobility, health, and energy.

3.9.2 The Data act

The Data Act (Regulation (EU) 2023/2854), is a regulation designed to create a single market for data by setting clear rules on access, use, and sharing. It focuses on ensuring that users of connected products and services—whether consumers or businesses—can access and reuse the data generated by their devices, while obliging manufacturers and service providers to provide such access under fair and reasonable conditions. The Act also establishes requirements for business-to-business and business-to-government data sharing, including compulsory access in cases of public interest such as emergencies. To strengthen digital sovereignty, it sets interoperability and portability rules for cloud and edge services, making it easier for users to switch providers. Taken together, the Data Act aims to reduce data lock-in, safeguard trade secrets and privacy, and unlock new opportunities for innovation and competition across the European data economy.

3.9.3 Other acts

The **Data Governance Act (DGA)**, effective since September 2023, establishes safe, trusted frameworks for both public and private data sharing, creating common data spaces across strategic sectors

3.10 RISE ICE use – in quotes

"*S3 buckets are using for data sharing. Data are used for developing and testing denoising (de-snowing, de-raining), segmentation and object detection model using Deep Learning, a subfield of Artificial Intelligence. HH published 45 GB data (snowy scenes) and 243GB data (rainy scenes). HH used data shared by FGI to create a de-snowing video by our Unsupervised Deep Learning model and data shared by WMG to compare the performance of de-raining model".*

Halmstad University, Sweden.

"THI use the RISE storage system to disseminate internally and externally datasets collected on the THI test fields (Virtual or Real). The RISE system allows for the quick allocation of data storage with the usage of its UI or AWS tools, which are essential for the upload of large files, such as the ones created during ROADVIEW WP7 and WP3. This is further accentuated by the usage of the public buckets to allow the FAIR usage and dissemination of datasets".

Technical University of Ingolstadt (THI), Germany

"Amazon S3 buckets are used for efficient and scalable data sharing among collaborators. Files can be uploaded and downloaded either manually via the AWS web interface or programmatically using the boto3 Python library, which enables automated data transfer, access control, and integration with data processing pipelines. This flexibility is especially useful when handling large datasets used in deep learning workflows. The shared data supports the development and evaluation of deep learning models for denoising (including de-snowing and de-raining), segmentation, and object detection, which are key tasks in computer vision and Artificial Intelligence."

The Research Institutes of Sweden, RISE, AB.

3.11 Post project storage options

Post project, we need to think about storage, if each partner does not download their own data and make it available (see FAIR). Academic and research projects within the EU, long-term options are:

Zenodo hosted by CERN, open-access

Zenodo is an open-access research data repository developed under the European OpenAIRE program and hosted by CERN, designed to make scientific outputs—including publications, datasets, software, and presentations—freely available and citable. It assigns DOIs (Digital Object Identifiers) to uploads, ensuring persistence, traceability, and scholarly credit. Unlike discipline-specific archives, Zenodo is multidisciplinary, welcoming contributions from all fields of science, as well as citizen science and smaller projects that may lack institutional repositories. By integrating with platforms like GitHub and supporting a wide range of file formats, Zenodo promotes transparency, reproducibility, and long-term preservation of research in line with FAIR data principles (Findable, Accessible, Interoperable, Reusable).

European Open Science Cloud (EOSC) – Long-term research storage

The European Open Science Cloud (EOSC) is a European Commission initiative aimed at creating a federated, trusted environment for long-term storage, sharing, and reuse of research data across disciplines and borders. It provides researchers, innovators, and institutions with access to interoperable services, data repositories, and computing resources, supporting the full data lifecycle from generation to preservation. EOSC is not a single data center but a cloud-like ecosystem that integrates national and institutional infrastructures into a common framework, ensuring compliance with FAIR principles. By offering sustainable storage, advanced analytics, and collaboration tools, EOSC is envisioned as a cornerstone of Europe's digital research infrastructure—strengthening open science, reducing fragmentation, and enabling large-scale, data-intensive research well into the future.

Dryad – Used for scientific data

Dryad is an open-access, nonprofit repository dedicated to sharing and preserving research data, particularly the datasets that support published scientific articles. It provides each dataset with a DOI, ensuring long-term findability, citation, and credit for authors. Unlike more general repositories such as Zenodo, Dryad focuses on curated, well-documented datasets to maximize reusability and compliance with FAIR data principles. It integrates closely with scholarly publishers, allowing authors to deposit data during manuscript submission, and its sustainability model is based on submission fees typically covered by journals, funders, or institutions. By prioritizing transparency, reproducibility, and interoperability, Dryad plays a key role in advancing open science and long-term accessibility of research data.

EUDAT (B2SAFE, B2SHARE) – Data storage

EUDAT is a pan-European data infrastructure initiative that provides services for the storage, sharing, and preservation of research data across scientific disciplines. Two of its flagship services are **B2SAFE** and **B2SHARE**: B2SAFE ensures **reliable, long-term, and policy-driven data replication** across trusted sites, helping researchers safeguard valuable datasets and comply with data management requirements, while B2SHARE offers an **easy-to-use, open-access repository** for researchers to store, publish, and share small- to medium-sized datasets, complete with persistent identifiers for citation. Together, these services support both secure archiving and broad dissemination of data, aligning with FAIR principles and complementing the broader European Open Science Cloud (EOSC) vision of an integrated, sustainable research data ecosystem.

The Fame project

Modern data marketplaces are transforming how data assets are shared, traded, and utilized. Recent European initiatives have made significant strides, particularly in enhancing data monetization, regulatory compliance, and secure data exchange. However, existing centralized marketplaces face challenges that limit broader participation and accessibility. Notable limitations include complex data discovery processes, limited transparency in value-based data monetization, and insufficient integration of trusted, energy-efficient analytics. Addressing these gaps can unlock new data-driven applications in sectors like finance, retail, and smart cities, empowering innovative services that seamlessly integrate financial data. That is where FAME comes in.

Maintain data storage at ICE (RISE)

We have also looked at keeping the data where it is and pay for keeping the data where it is. It is the simplest, by using resources ahead of time to host the data (centrally). However, administratively this is one of the most complex. One issue is estimating the cost of storage ahead of time, storage, electricity storage media. Another issue is the storage lifetime, which is coupled to the popularity of the datasets it host is not trivial to estimate. A simple estimate can be done for 5-10 years.

3.12 Related documents and sites

Related documents include ROADVIEW's Deliverables D4.5, D4.6, The data management plan, the GitHub (code) repository as well as ROADVIEW's license agreement. Initial Data Management Plan, 1.5.

3.13 Lessons learned

Planning early for costs and usage was useful. Writing a "Get Started" guide for novel or new users was also helpful, as people in the automated driving domain may not be familiar with cloud-based services. Regular billing and usage (above) have helped us maintain costs within budget. To be fair, we only used ICE for storage, computation would be a little more difficult, to estimate. However, some services help, Weights and Biases.

4 Conclusions

Data storage and data privacy are very important to the ROADVIEW project. Many, if not all, projects depend on data, however autonomous driving is of primary consequence due to the reliance of the vehicle to learn from the environment in which the vehicle moves. That is the road, its conditions, the surroundings as well as, especially in the case of ROADVIEW, the weather.

We have performed a TARA review of the data accumulation process in the first part of this document.

Privacy and security are key, since the data must be gathered, processed and shared securely. This this not only data breaches, but sensitive data must be kept to a minimum, if any at all. In Europe there are strict laws governing how data should be used. GDPR, for example, states that any personal information, addresses, security numbers, bank accounts must be removed or anonymised. In an autonomous driving project, such as ROADVIEW, sensitive data means number plates from the training (learning) vehicle as well as people's faces.

5 References

[1] European Union, «Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation).,» Official Journal of the European Union, 2016.