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

SYNERGIES

D3.1 Methodology for data quality assessment

Document Type	R – Document, Report
Document Number	D3.1
Primary Author(s)	Fadi Alakkad | University of Warwick
Document Version / Status	Final – V 1.0
Distribution Level	PU (public)

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

Funded by the European Union

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 the European Climate, Infrastructure and Environment Executive Agency (CINEA). Neither the European Union nor the granting authority can be held responsible for them.

CONTRIBUTORS

Name	Organization	Name	Organization
Fadi Alakkad	University of Warwick	Thanh Bui	RISE
Siddartha Khastgir	University of Warwick	Fredrik Warg	RISE
Burcu Kolbay	Mosaic Factor	Marc EL ZEENNY	IRT SystemX
Sara Garcia	Mosaic Factor	Gildas THIOLON	CEESAR
Theresa Weber	VUFO	Lars Klitzke	DLR
Lucas Dulewicz	VUFO		

FORMAL REVIEWERS

Name	Organization	Date
Emmanuel Arnoux	Renault	2025-01-24

DOCUMENT HISTORY

Revision	Date	Author | Organization	Description
0.1	2025-01-09	All	First draft version
0.2	2025-01-24	All	Incorporated peer review comments
1.0	2025-01-30	Fadi Alakkad | University of Warwick	Final version ready for submission

DOCUMENTATION OF SUBMISSION DELAY (OPTIONAL)

Planned date	Actual date
2024-11-30	2025-01-31

Justification	
Partner(s) causing the delay:	Task 2.3 team (delayed provision of technical quality descriptors)
Justification:	The extension is necessary because the development of the methodology for D3.1 (Methodology for Data Quality Assessment) relies on input from Task 2.3. However, Task 2.3 will not be able to provide the required technical quality descriptors in time for the original deadline. As a result, Task 3.1 must proceed without these descriptors, requiring additional time to ensure that the methodology remains adaptable to various metrics.
Effects on follow-up activities and other tasks:	The delay may affect subsequent tasks dependent on D3.1, requiring adjustments in the project timeline. However, the additional time will ensure a more comprehensive and adaptable methodology.

Specific mitigation actions to catch up:	
Statement deliverable owner:	
Statement task leader:	
Statement WP leader:	

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

EXECUTIVE SUMMARY

This document outlines the methodology developed under Task 3.1 for assessing data quality and evaluating the potential of (quasi-)stationary sensor units and moving vehicle datasets in high-quality scenario creation. This work plays a critical role in addressing key challenges in the Connected, Cooperative, and Automated Mobility (CCAM) domain, including the lack of interoperable scenario databases, prolonged and costly development cycles, and regulatory ambiguities. The methodology systematically analyses the quality and relevance of data from diverse sources, including both collected data and synthetically generated data.

This deliverable is a cornerstone of the SYNERGIES project, which aims to create a unified European platform to support the development, training, virtual testing, and validation of CCAM systems. Building on the Safety Assurance Framework from the HEADSTART and SUNRISE projects, SYNERGIES delivers an interoperable, federated scenario database that integrate data from prominent sources

The primary objective of this deliverable is to establish robust methodologies for evaluating the quality and relevance of collected and synthetically generated data for scenario generation. Specifically, it aims to:

1. Conduct a state-of-the-art review to consolidate knowledge from existing methodologies, tools, and metrics used in assessing data quality for CCAM applications. This review highlights key insights, gaps, and best practices, providing a foundation for developing a new, comprehensive approach to scenario generation.
2. Define and implement two complementary methodologies for data quality assessment:
    a. **High-Level Qualification Methodology**: A screening process focused on essential quality criteria such as relevance, completeness, and accessibility to quickly evaluate datasets for basic usability.
    b. **Detailed Level Qualification Methodology**: A more granular, post-screening approach to assess datasets across key quality dimensions, including accuracy, completeness, consistency, timeliness, and coverage.
3. Develop quality metrics and thresholds for the methodologies, enabling datasets to be rigorously evaluated for technical relevance and regulatory compliance.
4. Support the integration of diverse traffic data sources, including roadside units, drones, and vehicle data, into scenario extraction pipelines,
5. Contribute to SYNERGIES' overarching goal of creating a European Scenario Dataspace, fostering data sharing, standardization, and scalability in CCAM development.

Main Results Achieved:

1. **Comprehensive Literature Review**: A synthesis of state-of-the-art methodologies, tools, and best practices, providing a foundation for developing scalable and standardized data quality assessment approaches.
2. **Development of Quality Metrics and Thresholds**: Introduction of criteria for evaluating dataset attributes, including relevance, coverage, completeness, accessibility, licensing, metadata quality, and usability.
3. **Two Complementary Methodologies** Implementation of high-level and detailed methodologies to ensure robust data evaluation, supporting both initial screening and granular analysis.

4. **Dataset Assessment**: Evaluation of fourteen datasets using the high-level methodology to demonstrate the effectiveness of the high-level methodology,
5. **Support for SYNERGIES Platform**: Ensuring that datasets integrated into the European Scenario Dataspace meet stringent quality and interoperability standards, fostering innovation in CCAM systems.

1 INTRODUCTION

The SYNERGIES project aims to address critical challenges in the development and deployment of CCAM systems. These challenges include the lack of interoperable scenario databases, time-intensive development cycles, and regulatory ambiguities. By leveraging the Safety Assurance Framework from previous initiatives such as HEADSTART and SUNRISE, SYNERGIES is creating a European Scenario Dataspace to enable seamless data sharing, standardized processes, and accelerated development cycles, while ensuring safety and regulatory compliance.

1.1 Overview

Deliverable D3.1 focuses on developing a comprehensive Methodology for Data Quality Assessment, a cornerstone for creating high-quality scenarios essential to the validation, testing, and deployment of CCAM systems. This methodology supports SYNERGIES' broader objectives by establishing robust processes for evaluating both collected and synthetically generated datasets. It enables stakeholders to ensure data quality and relevance while addressing interoperability and scalability requirements for CCAM scenario generation.

This deliverable introduces two levels of data quality assessment:

1. **High-Level Qualification Methodology:** A rapid screening tool designed to assess datasets against essential quality criteria such as relevance, completeness, and accessibility.
2. **Detailed Level Qualification Methodology:** A deeper evaluation process focusing on key criteria as accuracy, consistency, timeliness, and coverage, offering a comprehensive analysis of dataset reliability and utility.

To build these methodologies, the deliverable integrates insights from a comprehensive state-of-the-art review, consolidating knowledge from existing tools, methodologies, and metrics. Also, in constructing these methodologies, the deliverable builds on several key input documents:

- The high-level requirements developed in D2.1 and Task 2.1, which outline the foundational data needs for scenario generation.
- The data and format requirements outlined in WP5 - Task 5.1 Scenario Methodology, which ensure that datasets adhere to standardized structures and formats for interoperability.
- D2.2 (Storylines definition and technical and interoperability requirements), originating from WP2, T2.2 (User Storylines and Technical Requirements), which defines user storylines based on stakeholder needs and establishes technical requirements, provides essential guidelines for scenario generation and data usage.
- T2.3 (Interoperability requirements) and T2.4 (Data traceability, trustworthiness, and cyber-security requirements), originating from WP2, help ensure data consistency, interoperability, security, and compliance with privacy and regulatory standards. T2.3 focuses on standardized labels, metadata formats, hierarchical classifications, and open interfaces for seamless data exchange, while T2.4 establishes traceability, trustworthiness, cybersecurity measures, and GDPR compliance for safe data usage.

The methodologies developed in D3.1 address critical needs within the CCAM domain, such as ensuring that datasets are complete, accurate, and fit for purpose. Furthermore, they support

SYNERGIES' broader ambition of creating a unified and scalable Scenario Dataspace, enabling stakeholders to efficiently develop, validate, and deploy automated mobility systems in a safe and reliable manner.

1.2 Relation with other tasks

In the context of the SYNERGIES project, Task 3.1 serves as a foundational element by establishing a robust methodology for assessing data quality and relevance. This methodology encompasses quality criteria, metrics, and thresholds, which are essential for evaluating both existing and newly generated data. The interconnections between Task 3.1 and other tasks within the project highlight a cohesive and iterative approach to achieving the objectives of the SYNERGIES platform.

As seen in Figure 1: Description of the relation between task T3.1 and the other tasks T2.3, T3.2, T3.3, T3.4 and T3.5., the following subsections detail the relationships between Task 3.1 and other key tasks (2.3, 3.2, 3.3, 3.4, and 3.5), underscoring the collaborative and interdependent nature of their activities. Each subsection explores the specific ways in which Task 3.1 influences and integrates with these tasks to ensure the development of high-quality, interoperable datasets and methodologies critical for advanced traffic and scenario modelling. This alignment ensures that the project outcomes are comprehensive, accurate, and reflective of real-world conditions and challenges.

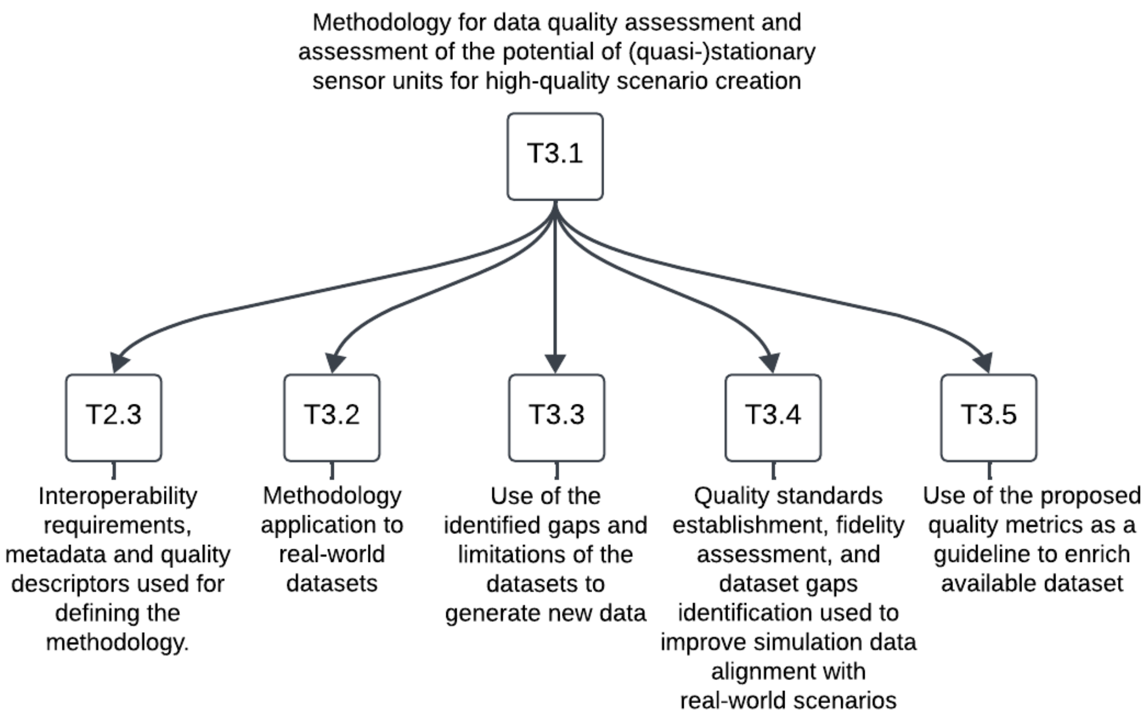

Figure 1: Description of the relation between task T3.1 and the other tasks T2.3, T3.2, T3.3, T3.4 and T3.5.

1.2.1 Relation to Task 2.3

Task 2.3 focuses on data consistency and interoperability requirements, including standardized labels, format, and data model.

Task 3.1 and task 2.3 are interlinked as following:

- Interoperability requirements: The methodology developed in T3.1 shall consider the interoperability requirements, ensuring key information exist to support the data model hierarchy.
- Metadata: Quality assessment methodology developed in T3.1 shall consider assessment of metadata key information in high-level qualifications.
- Data quality metrics: Task 3.1 proposes metrics per each quality metrics provided by T2.3, together with thresholds (via statistical research and domain expertise).

Note:

At the time of writing this deliverable, T2.3 has not yet finalized the quality descriptors. As a result, it has been decided that T3.1 will proceed with D3.1 without incorporating the quality descriptors from T2.3. Instead, the D3.1 methodology will be developed using specific metrics defined through state-of-the-art research, the expertise of project partners, and discussions with other tasks, such as T2.3, T3.2, and WP5 tasks.

1.2.2 Relation to Task 3.2

Task 3.2 focuses on analysing the current state-of-the-art regarding available data and scenario coverage. This task involves several key activities:

1. **Surveying Existing Data Sources**: Task 3.2 conducts a thorough review of open data sources, accident databases, and data from other national and European Union (EU) projects. Key projects include Hi-Drive, L3Pilot, UDRIVE, and other Cooperative, Connected, and Automated Mobility (CCAM) initiatives. While an initial survey of potential data sources was performed during the project development phase, this analysis will be updated at the beginning of the project to incorporate any newly available data sources.

2. **Data Qualification**: The methodology developed in Task 3.1 will be applied to assess the quality and relevance of the identified data sources. This process will also include the creation of metadata for the datasets under consideration, ensuring their suitability for scenario generation.

3. **Focus on Accident Data**: Special attention will be given to the collection of accident data. Such data is critical for studying challenging and potentially dangerous scenarios. By understanding these scenarios, the project aims to replicate them for analysis and ultimately devise strategies to prevent similar incidents in the future.

Relationship Between Task 3.1 and Task 3.2:

Task 3.1 and Task 3.2 are closely interlinked. Task 3.1 focuses on developing a comprehensive methodology for analysing the quality and relevance of both available and newly generated data. This methodology defines quality metrics, and thresholds to assess data comprehensively. It also identifies gaps and limitations in existing datasets to improve their accuracy, completeness, and appropriateness for scenario generation.

Task 3.2 builds on this foundation by applying the methodology from Task 3.1 to real-world datasets. It evaluates open data sources, accident databases, and data from national and EU projects to determine their relevance and quality for scenario generation. By leveraging the framework established in Task 3.1, Task 3.2 ensures that the data analysis and qualification processes are systematic and robust.

1.2.3 Relation to Task 3.3

Task 3.3 focuses on generating new data to complement the datasets analysed in Task 3.2. The main activities within this task include:

1. **Data Collection from Equipped Vehicles**: Task 3.3 involves gathering data from vehicles equipped with advanced sensors such as cameras, radars, and lidars. These vehicles are owned by members of the project consortium.

2. **Utilization of Stationary Sensors and Drones**: The task incorporates data collection using (quasi-)stationary sensor units and drones. These tools will capture traffic data across different geographic areas and road topologies.

3. **Focus on Urban and Rural Roads**: Data collection will prioritize urban and rural environments. A variety of conditions, including diverse weather scenarios, will be covered to ensure comprehensive representation in the datasets.

4. **Targeted Data Campaigns**: Based on the quality assessment framework established in Task 3.1, any gaps identified in the datasets will prompt targeted data collection campaigns to fill these deficiencies.

5. **Feedback Integration**: Feedback from other work packages (WP5, WP6, WP7, and WP8) will be used to ensure that the data generated aligns with the broader needs and objectives of the SYNERGIES platform.

Relationship Between Task 3.1 and Task 3.3:

Task 3.1 and Task 3.3 within the SYNERGIES project are intrinsically connected as part of a systematic approach to ensuring high-quality and relevant data for scenario generation. Task 3.1 is dedicated to developing a robust methodology for analysing the quality and relevance of both existing and newly generated data. This includes defining quality metrics, and thresholds to evaluate datasets comprehensively. The outputs of Task 3.1 serve as a critical foundation for Task 3.3 by identifying gaps and limitations in available datasets, thereby highlighting specific areas where new data is required. Task 3.3 builds upon the framework established in Task 3.1 by generating new data to address these identified gaps. Data collection in Task 3.3 utilizes advanced technologies such as equipped vehicles with sensors, quasi-stationary sensor units. This task focuses on urban and rural environments under diverse conditions, ensuring a wide range of scenarios are represented. The quality metrics and thresholds from Task 3.1 are applied to evaluate the new data, ensuring it meets the project's rigorous standards for scenario generation. The relationship between these tasks is dynamic and iterative. Insights from Task 3.1 guide the data collection efforts in Task 3.3, while the feedback from the data generated in Task 3.3 helps refine the methodology and metrics of Task 3.1. Together, these tasks ensure that the SYNERGIES project produces high-quality datasets that are comprehensive, accurate, and tailored to the needs of advanced traffic and scenario modelling.

1.2.4 Relation to Task 3.4

Task 3.4 focuses on data generation using advanced simulation tools such as CARLA, SUMO, and AIMSUN, aiming to produce high-quality datasets of complex traffic scenarios, including accidents, edge cases, and challenging road situations. Task 3.1 and Task 3.4 are closely interconnected through an iterative approach to data quality assurance and enhancement. The methodology established in Task 3.1, which includes quality metrics, and thresholds, will be directly applied to assess and refine the datasets generated in Task 3.4. This ensures the data meets the required standards for accuracy, completeness, and appropriateness for scenario generation.

The outputs of Task 3.1 guide Task 3.4 by:

- Defining quality requirements for datasets generated via simulation tools, ensuring alignment with the interoperability and metadata standards outlined in Task 2.3.
- Providing a framework for assessing the fidelity of simulated data against real-world data to identify discrepancies and improve simulation accuracy.
- Highlighting specific gaps in existing datasets that simulation tools in Task 3.4 can target, such as underrepresented traffic scenarios or edge cases.

1.2.5 Relation to Task 3.5

The Task T3.5 focuses on using generative AI methods as additional tools for synthetic data generation. They key contributions of the task are:

- Identification and development of generative AI methods using both, real world data from T3.2/T3.3 and simulated data from T3.4 as inputs and generate synthetic data that enrich the existing data.
- Providing best practices for applying AI in CCAM applications, while supporting scenario generation in WP5.

The Task 3.1 will provide an overview of evaluation categories and metrics to assess datasets with real-world data and synthetically generated. The Task 3.5 uses the proposed quality metrics as a guideline to enrich available dataset as well as generating data to fill the gaps between the data availability and data requirements.

2 METHODOLOGY FOR DATA QUALITY ASSESSMENT

2.1 Introduction

This chapter outlines the methodology developed for assessing the quality of data used in CCAM scenario generation. High-quality data is critical for ensuring the reliability, accuracy, and interoperability of the scenarios required for testing, validating, and deploying CCAM systems. The methodology described here aims to address the challenges of evaluating diverse datasets by offering a structured, scalable, and robust approach to data quality assessment.

To develop this methodology, A comprehensive literature review (Section 2.2) has been conducted to explore existing data quality assessment practices, metrics, and methodologies. Insights from the literature, combined with project requirements gathered from earlier tasks and deliverables T2.1 (D2.1), T2.2, T2.3, T2.4 (D2.2), T5.1.

The chapter is structured as follows:

Section 2.2: Literature Review – Synthesizes findings from state-of-the-art research on data quality assessment, highlighting key methodologies, metrics, and gaps.

Section 2.3: Requirements Collection – Summarizes the data, format, and interoperability requirements derived from earlier tasks in WP2 and WP5.

Section 2.4: Approach to Qualify Datasets – Presents the approach adopted to assess the quality of datasets withing SYNERGIES project. In this section, two complementary methodologies were defined: High-Level and Detailed Level qualification methodologies. Additionally, this section discusses the application of the High-Level Qualification Methodology to a selection of 14 datasets, demonstrating its performance in identifying datasets suitable for further detailed evaluation. This practical application highlights the robustness and scalability of the methodology, offering insights into its effectiveness and potential areas for refinement.

2.3 Literature review

Data quality is a fundamental concern in various fields, particularly when data is used for critical decision-making, system functionality, or collaborative processing. Ensuring that data meets high standards of quality is important to achieving reliable, accurate, and trustworthy outcomes. This literature review explores the various dimensions of data quality, its significance across different domains, and the methodologies employed to assess and manage it.

2.3.1 Significance of Data Quality in Automated Driving Systems

The publication of [2] primarily focuses on addressing data quality challenges and requirements within safety-critical systems like ADAS. The study introduces a Candidate Framework for Data Quality Assessment and Maintenance (CaFDaQAM), which offers a structured approach for assessing and managing data quality.

Key points:

1. Importance of Data Quality:
    - High data quality is crucial for the safe and effective operation of data-driven systems like ADAS

   - Poor data quality can lead to system failures and even fatal accidents in automated driving environments

2. Data Quality Challenges:
   - Study identifies multiple data quality challenges, which include:
     - Data availability (e.g., data delay, data drop, incomplete data)
     - Data management (e.g., manual data collection, imbalanced datasets)
     - Data source issues (e.g., new data types, defective sensors)
     - Data structure challenges (e.g., fragmented or unstructured data)
     - Data trust (e.g., incorrect labelling, lack of reliable data)
3. Data Quality Attributes:
   - The framework outlines various data quality attributes such as accuracy, timeliness, completeness, and consistency, and provides metrics to assess them
4. Framework Components:
   - Data Quality Workflow: A step-by-step process for identifying, assessing, and managing data quality issues
   - List of Data Quality Challenges: A categorized template for identifying and recording data quality challenges
   - List of Data Quality Attributes: A collection of attributes that define good data quality, including metrics to assess these attributes
   - Solution Candidates: Strategies and tools to address data quality challenges, such as automated data labelling or data corroboration techniques
5. Application to Automated Driving:
   - The framework is applied to ADAS, where data quality is critical for ensuring that deep learning models operate safely and effectively
   - The study highlights how training data, validation data, and runtime data quality are crucial for the system's reliability

The framework provides a comprehensive method to systematically manage data quality, which is especially important for systems that rely heavily on data, such as those in automated driving. The study offers solutions to improve data quality and ensure that safety-critical systems can function as intended.

2.3.2 Key Dimensions of Data Quality

The work of [1] discusses the principles and methods for assessing data quality within organizations. It explores both subjective and objective dimensions of data quality and provides a framework for developing usable metrics and describes different aspects for data quality as the following : .

1. **Data Quality as a Multi-Dimensional Concept:**
   - Data quality involves both subjective perceptions (from data stakeholders like collectors, custodians and consumers) and objective measurements

- Subjective aspect reflects how stakeholders view the data's quality, often captured through surveys
- Objective measures can either be task-independent or task-dependent, considering the context and application of data

2. **Functional Forms for Objective Data Quality Metrics:**
   - Simple Ratio: Commonly used to assess errors, completeness and consistency
     - expresses a ratio of desirable outcomes to total outcomes
   - Min or Max Operation: Used to aggregate multiple indicators
     - the min operation takes the weakest indicator's value, while max operation is more liberal
   - Weighted Average: Averages different quality dimensions, assigning weights based on their importance
3. **Subjective vs. Objective Assessments**
   - The article outlines a method of comparing subjective and objective data quality assessments, identifying discrepancies and addressing root causes of poor data quality
   - Companies like Global Consumer Goods (GCG) and Data Product Manufacturing (DPM) have used these assessments to improve data consistency, completeness, and believability
4. **Examples of Data Quality Dimensions:**
   - Free-of-error: Measures the correctness of data
   - Completeness: Can be viewed in terms of schema, column or population completeness
   - Consistency: Refers to uniformity across data sets
   - Timeliness and Accessibility: Focus on the data's relevance and availability for a specific task
5. **Improving Data Quality:**
   - The process involves conducting both subjective and objective assessments, comparing them, and taking corrective actions as needed
   - The goal is to establish high data quality across all dimensions and roles within an organization
6. **Industry Benchmarks**
   - Developing standard data quality metrics across industries could provide companies with comparative performance measures
   - While a single, aggregate data quality measure is often desired by organizations, it comes with challenges due to subjectivity and scale-type differences.

Furthermore, the work of [11] presents a comprehensive overview of 265 autonomous driving datasets, systematically analysing them based on characteristics like sensor modalities, data size, tasks, and environmental conditions. The authors propose an impact score to assess the impact of a dataset on the autonomous driving domain. This score is a combination of three scores, while the data dimension score and the environmental diversity score are relevant for this project's use case. Both scores highlight different aspects of a dataset. The data dimension score allows to assess a dataset based on "dataset size, temporal information, task number, and labelled categories." [2, p. 4] The environment diversity score is used to measure the environmental conditions based on weather, time of day, driving scenario type (e.g. urban or rural) and geometric scope (e.g., the countries or cities of data collection).

On the other hand, A comprehensive analysis by [6] identifies core dimensions and metrics necessary for assessing data quality. The study, which reviews 29 key articles on data quality, highlights several crucial dimensions that are foundational to evaluating the quality of data in various domains. These dimensions include:

Completeness: This dimension evaluates the breadth and coverage of datasets, ensuring that all necessary data is present to provide a full representation of the system.

Accuracy/Precision: It measures the correctness and exactness of data, ensuring that data values reflect real-world truths.

Consistency: This dimension ensures uniform representation of data across different datasets, avoiding discrepancies.

Relevance: This evaluates the applicability and usefulness of the data for its intended purpose.

Timeliness/Currency: It considers whether the data is up to date, highlighting its current relevance and whether it reflects the most recent information.

These dimensions form the backbone of data quality evaluation and directly support the methodology for data quality assessment outlined in [7]. The article presents a structured methodology for systematically evaluating and improving data quality, offering a valuable framework for practical implementation in organizations. The dimensions outlined—accuracy, completeness, consistency, timeliness, and relevancy are further detailed in the context of SMEs, which face unique challenges in managing and assessing data quality due to resource constraints and the complexity of data integration. However, other studies [9] emphasize additional factors critical to specific fields or applications. For instance, a review conducted for health data applications identified several other relevant dimensions, including:

- **Correctness:** the extent to which data are true and unbiased
- **Timeliness:** the extent to which data are promptly processed and up to date
- **Stability:** the extent to which data are comparable among sources and over time
- **Contextualization:** the extent to which data are annotated with acquisition context
- **Representativeness:** the extent to which data are representative of intended use
- **Trustworthiness:** the extent to which data can be trusted based on the owner's reputation
- **Uniqueness:** the extent to which data are not duplicated

2.3.3 Data Quality Frameworks and Methodologies

The importance of ensuring high-quality data in collaborative data processing systems has been widely acknowledged. Various methodologies and frameworks have been developed to address challenges related to standardization, automation, and the unique requirements of specific domains like IoT.

The lack of consensus on data quality criteria (DQC) has long been a critical issue in the field. A systematic literature review conducted by [3] addressed this problem by consolidating and standardizing 30 key DQCs. The study identified and resolved inconsistencies in the naming and definitions of criteria across prior research, presenting a unified framework. This standardization effort facilitates trust-building in decentralized governance systems, compliance with legal regulations, and improved reliability in distributed processing environments. Furthermore, the authors proposed directions for future work, including the development of automated data quality assessment frameworks and addressing challenges related to trust, compliance, and reliability in collaborative ecosystems.

In the context of real-time data processing, [12] introduced the Data Quality Scoring Operations (DQSOps) framework. This methodology focuses on automating the scoring of data quality in high-volume, streaming workflows typical of DataOps and MLOps pipelines. The framework employs a ground-truth standard-based approach to establish benchmarks for multiple data quality dimensions. Machine learning (ML) predictors are then trained against this ground truth, enabling the automation of quality scoring at scale. Notably, the ML models are designed to continuously evaluate streaming data and recalibrate as new ground-truth benchmarks are established. This approach ensures that the system maintains accuracy and scalability while minimizing manual effort.

IoT systems, characterized by dynamic and multi-source data environments, present unique data quality challenges. To address these, [16] proposed a novel approach that prioritizes data freshness, moving beyond traditional timestamp-based evaluations. The study introduced a four-tuple freshness model that considers the timestamp, value, not-null checks, and acceptable thresholds for IoT data. Additionally, an algorithm was developed to detect and discard unreliable data by analysing multiple timestamps and identifying abnormalities, such as sudden fluctuations in sensor readings. Although limited to tabular IoT data and focused solely on freshness metrics, this framework highlights the need to adapt quality assessment techniques to domain-specific challenges.

To provide a structured methodology for data quality assessment, [17] introduced a framework that integrates prescriptive and descriptive quality metrics. Using international standards such as ISO/IEC 25012 and 25024, the study differentiates between metrics that assess quality post hoc and those that proactively guide data preprocessing and model selection. Descriptive metrics, such as completeness and consistency, serve as diagnostic tools, while prescriptive metrics, such as threshold-based completeness and validity rules, enable actionable recommendations for improving data quality. This dual approach helps reduce repetitive iterations in data handling and supports more efficient decision-making.

Furthermore, other scholars in [7] divides data quality assessment and testing into three key areas:

Individual Tests: These assess each data value independently to verify conditions such as non-null or non-empty values.

Form-Related Tests: These compare data within the same attribute, evaluating characteristics like value distribution and adherence to specified conditions.

Content-Related Tests: These compare data values with external references, such as default values or data from other attributes, to ensure consistency and correctness.

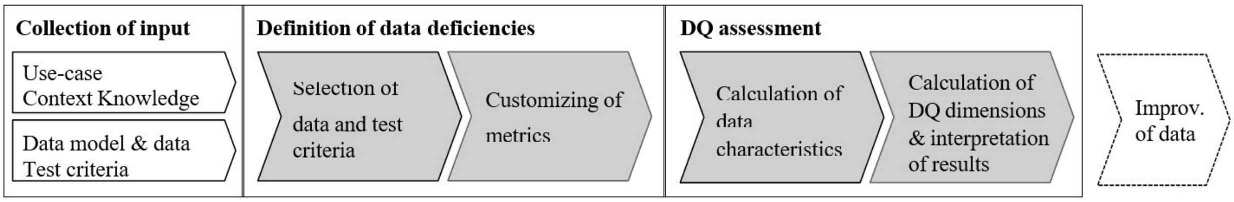

Figure 2: Main steps of DQ assessment methodology

By integrating these testing methods, the article provides a structured approach to assessing data quality, which directly aligns with the objectives of Task 3.1 in the SYNERGIES project. This study provides valuable insights into defining and evaluating data quality dimensions such as accuracy, completeness, consistency, timeliness, and relevancy.

Furthermore, the paper of [15] introduced a framework for assessing and visualising data quality in real-time, focusing on completeness as a critical metric for open government data (OGD). The framework aims to automatically measure the Data Completeness Ratio (DCR), which is the percentage of non-null values in a dataset and visualize it on a real-time Django-based dashboard. The framework emphasizes the automation of real-time quality checks, enabling continuous monitoring of DCR without manual intervention. Additionally, it offers actionable recommendations to improve dataset completeness, such as prioritizing the filling of mandatory fields and addressing missing values. By sorting files based on their DCR values, the framework allows for the prioritization of higher-quality datasets. Despite its innovations, the framework has notable limitations. It is restricted to tabular datasets accessed via APIs, and its scope is limited to the DCR metric; other quality dimensions, such as accuracy, require manual assessment. However, the framework provides opportunities for further enhancements, including the clustering of datasets for cluster-based ranking, automatic labelling of datasets with lower DCR values below a predefined threshold, and creating feedback loops to guide data providers on improving both mandatory and non-mandatory fields. Overall, this framework contributes significantly to the real-time monitoring and visualization of data quality, offering a practical and actionable tool for managing open government data.

Additionally, quantitative metrics for scenario-based coverage were proposed by [14] with the introduction of the “k-Projection coverage” metric. This metric evaluates data coverage by projecting datasets onto k-dimensional hyperplanes based on selected criteria (e.g., weather, illumination, road conditions). The goal is to ensure adequate representation in all regions of the dataset, with weights assigned by domain experts. This methodology helps identify missing combinations in datasets, enabling better coverage and data representativeness for complex scenarios.

2.3.4 Data Quality Challenges and Limitations in Decision-Making

In decision-making, data quality can significantly impact outcomes, especially when uncertainty is high. Traditional methods often assess data quality using fixed metrics, which don't capture its dynamic nature in real-world scenarios. A study by [4] introduces an adaptive classification method, categorizing data quality into four levels: Critical, Below Average, Average, and Above Average. Unlike static assessments, this system adjusts based on past performance and context, providing a more flexible and intuitive way to manage data quality in uncertain environments.

Through a case study in forestry, the research demonstrates the practical application of this method, as seen in Figure 3: Classification method for wood supply. A forest products company uses it to optimize wood supply planning, incorporating predicted harvest volumes with varying levels of data accuracy. By categorizing data quality dynamically, the company adjusts its transportation models to account for uncertainty, resulting in more reliable forecasts and fewer last-minute adjustments.

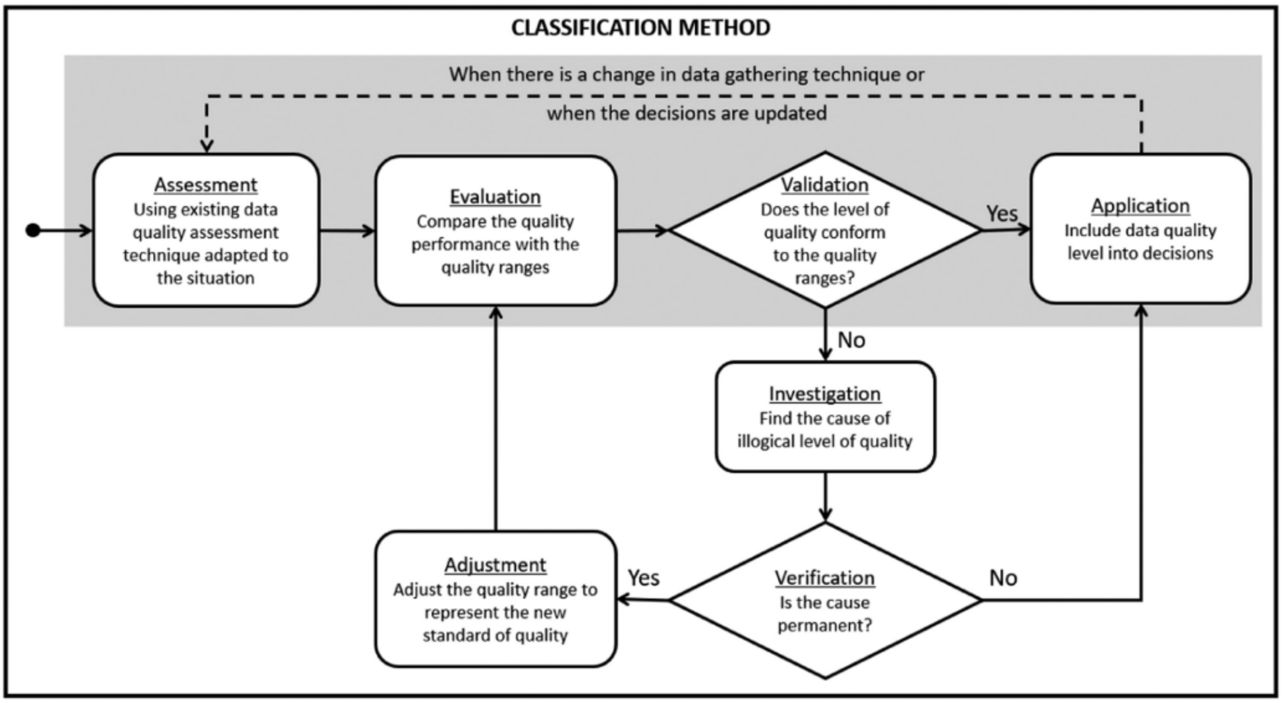

Figure 3: Classification method for wood supply

While this approach offers valuable insights, the study acknowledges limitations in its generalizability beyond the forestry sector, calling for further research across different industries.

At the same time, another paper by [5] highlights the shortcomings of traditional data quality approaches, which treat it as a one-time check rather than an ongoing process. In the context of Industrial Data Science (IDS), this study proposes a four-phase framework for Data Analysis Quality (DAQ), focusing on continuous quality management throughout the data lifecycle from access to application, as seen in Figure 4: Four steps for categorizing the criteria of data analysis quality.

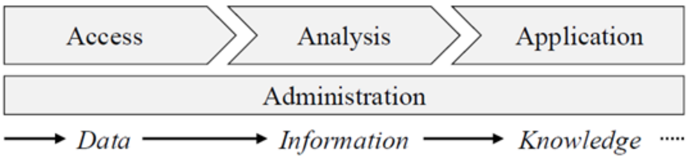

Figure 4: Four steps for categorizing the criteria of data analysis quality

By integrating quality checks at each stage, the framework ensures that data remains accurate, usable, and secure. See Figure 5: Process model for integrated Data Analysis Quality over the course of the four layers.

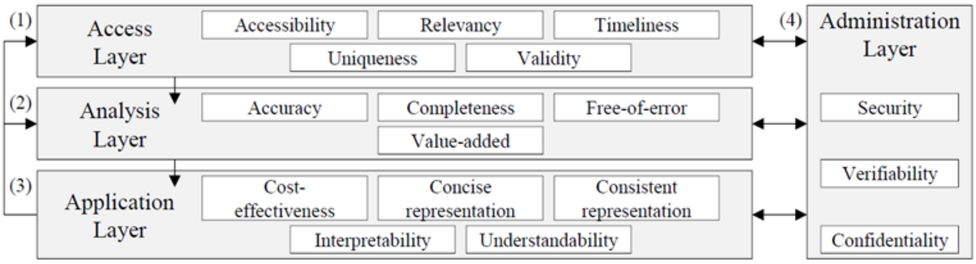

Figure 5: Process model for integrated Data Analysis Quality over the course of the four layers

The two studies share a common thread: they both advocate for more dynamic and context-aware methods of handling data quality. While [4] focuses on adaptive classification for decision-making under uncertainty, [5] offers a broader approach for maintaining data quality in industrial contexts. Together, these works suggest that the future of data quality lies in continuous, context-sensitive management that adapts to both data changes and evolving decision needs. Furthermore, the work of Goknil et al [10] presents a systematic review on data quality challenges within Cyber-Physical Systems (CPS) and the Internet of Things (IoT) for Industry 4.0, with the primary aim of analysing current research on data quality techniques in this field. The work investigated data quality issues and their sources for CPS and IoT, data quality metrics, data quality techniques (including data quality monitoring, data repair, and data cleaning), and software engineering solutions for handling data quality.

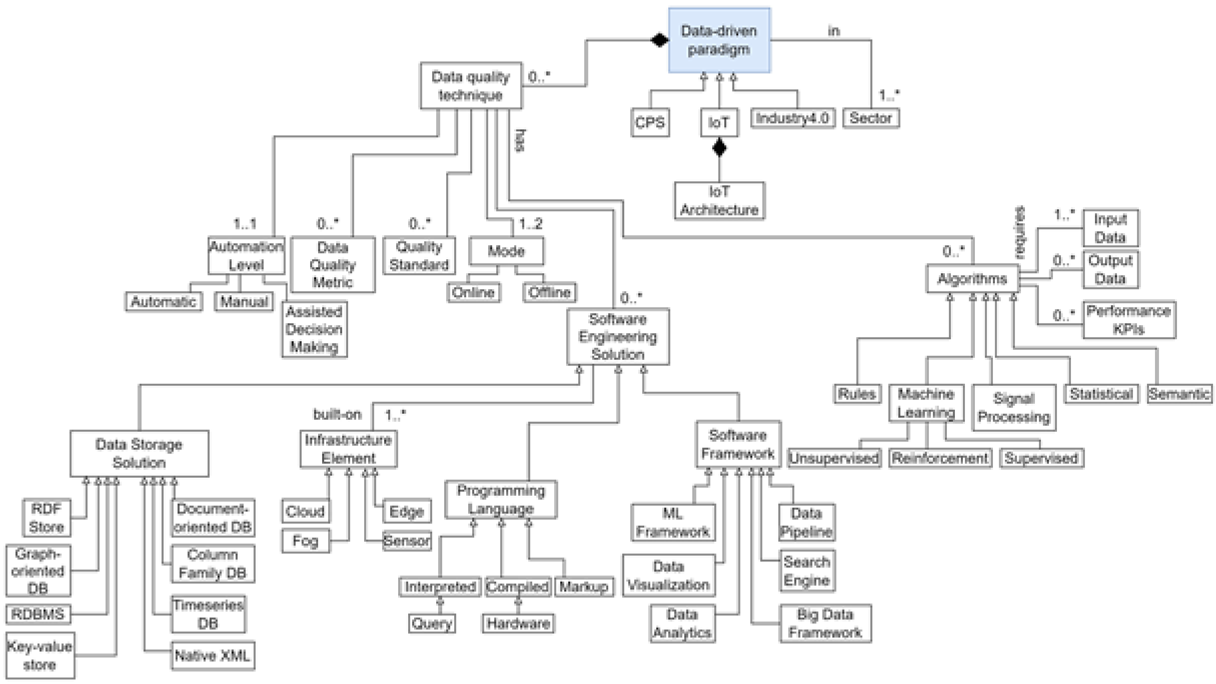

Figure 6: A taxonomy of data quality in CPS and IoT

While the taxonomy covers various aspects of the domain, they work also elaborates on data quality metrics which are "the measurements by which you assess your data" [1, p. 10]. These metrics are pivotal to assess whether your dataset contains "high-quality and low-quality data" [1, p. 10]. In fact, the authors identified several issues related to data quality, which may also be related to the domain of automated vehicles. These include the use of multiple data sources, sensor malfunction, cyber-attacks and high sampling frequency. Furthermore, the authors have compiled a list of data quality metrics derived from their literature review illustrated in Figure 7: Data quality issues and data quality metrics for CPS and IoT. These metrics may also be relevant for the project's use cases and include accuracy, completeness, confidentiality, timeliness, and validity, among others.

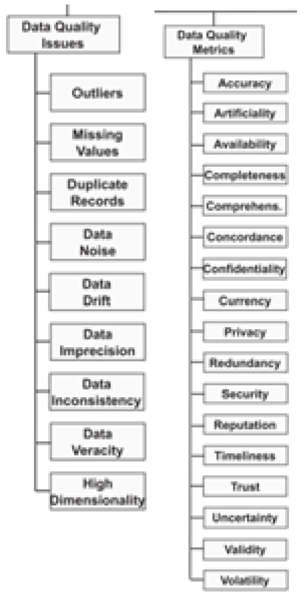

Figure 7: Data quality issues and data quality metrics for CPS and IoT

2.3.5 Key Framework for Optimizing Data Collection and Usability

The report of CCAM Data Sharing Framework [8] increases demand for data in Cooperative, Connected, and Automated Mobility (CCAM) research has highlighted the need for efficient data collection and administration. This report seeks to address these challenges by proposing a comprehensive framework of guidelines and recommendations aimed at optimizing the usability of recorded datasets.

The framework emphasizes two critical aspects:

- **Preparation and Specification of Data Collection:** To ensure datasets meet research requirements, meticulous planning and specification are crucial. Properly defining the scope and objectives of data collection helps in aligning the dataset's content with the intended research goals.
- **Facilitation of Data Reuse:** Reusability is enhanced through the provision of thorough and precise documentation. This includes detailed information on data structure, collection methods, metadata, and other key aspects of data management, ensuring clarity and accessibility for future users.

By outlining essential components that should be included in a dataset, it offers a practical foundation for evaluating dataset quality and suitability for research purposes.

Those essential components cover several questions that data provider should have answer to:

- Ownership and access to data and data tools
- Storage and download of data
- Access methods
- Cybersecurity
- Areas of use
- Post-project re-use of data

- Post-project financing

This report provides a **metadata documentation** template to fill up by providers, covering the following fields:

1. Dataset Summary (Location, date,)
2. Administrative metadata (Contact, Access rights)
3. Dataset categories for search and select (Type of dataset, type/number of vehicles, data recording protocol, etc.)
4. Structural metadata
   - Data storage format
   - data field description:

Table 1: data field description template

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			

   - Manual annotations
   - Data quality policy
5. Study design and test execution as separate documentation

This report identifies the essential elements that must be addressed to determine whether a specific dataset should be included in SYNERGIES. Particular attention is given to legal aspects, which are to be integrated to SYNERGIES's data quality assessment methodology.

2.3.6 Data Readiness Levels (DRL) for Dataset Quality Communication

Furthermore, the paper [13] advocates for "Data readiness levels" (DRL) as a common language for communicating the quality/maturity of datasets **Error! Reference source not found.** through t he following approaches.

1. Give a level of the readiness of the data in different dimensions/bands
   a. Accessibility
   b. Faithfulness
   c. Data in context
2. Can be used e.g. to assess the work needed to bring a dataset to the necessary quality level

To apply this method, different bands (dimensions) need to be defined, where each band has a score indicating the readiness to be used for some purpose. The position paper suggests the bands illustrated in Figure 8: Data readiness levels. does not clearly define the levels. This is up to the user and use-case.

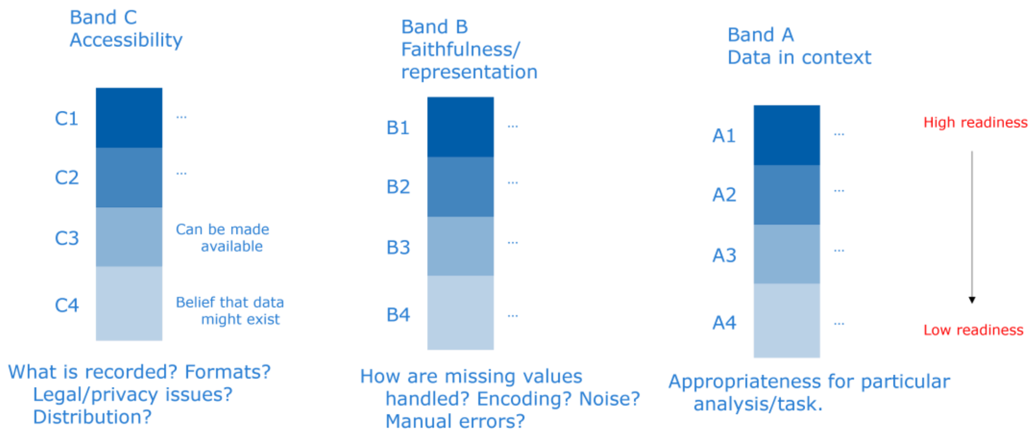

Figure 8: Data readiness levels.

2.4 Requirements collection

To develop a robust and comprehensive methodology for data quality assessment, it is actively coordinated with various tasks and work packages within the **SYNERGIES** project to gather relevant requirements. The requirements collection process ensured alignment with the project's overarching goals and the specific needs of scenario generation, interoperability, and data traceability.

Task 2.3 (Interoperability Requirements) was initially expected to provide the quality descriptors necessary for Task 3.1. However, as the descriptors have not yet been finalized in T2.3, it was decided that Task 3.1 would proceed with Deliverable D3.1 independently. To address this, the developed methodology in D3.1 has been formulated to remain **agnostic to specific metrics**, ensuring flexibility while still aligning with the overall project requirements. This approach allows the methodology to adapt seamlessly once the finalized descriptors from T2.3 become available, ensuring compatibility and robustness in assessing data quality.

Deliverable D2.1 (*High-Level Requirements*) established the foundational data quality expectations, helping us identify the most critical dimensions such as completeness, relevance, and timeliness. It is also referred to **Deliverable D2.2** (*Storylines Definition and Technical & Interoperability Requirements*), which provided critical insights into the specific storylines, scenarios, and technical constraints that the datasets must address. The definitions outlined in D2.2 shaped the understanding of data traceability, accuracy, and usability.

In addition, **Work Package 5** (WP5), which is responsible for generating and managing scenarios, contributed technical and data format requirements essential for ensuring that the assessed datasets can be seamlessly integrated into the scenario creation process. The guidelines from WP5 highlighted the importance of structured, high-quality data that supports scenario extraction, testing, and validation activities.

The requirements gathered through this collaborative effort were complemented by the findings from the **Literature Review** in this document (Section 2.2). The insights from state-of-the-art methodologies, frameworks, and metrics combined with project-specific requirements formed the basis for the two-tiered approach to data quality assessment presented in this deliverable: the **High-Level Qualification Methodology** and the **Detailed-Level Qualification Methodology**. The two-tiered approach serves distinct yet complementary purposes: the High-Level Qualification Methodology provides a way to quickly assess and identify the most promising datasets from a large number of candidates, while the Detailed-Level Qualification Methodology enables an in-depth analysis of the dataset candidates identified by the high-level methodology.

This coordinated approach ensures that the developed methodology is both theoretically grounded and practically aligned with the needs of scenario generation, enabling the seamless evaluation and integration of high-quality datasets into the **SYNERGIES** platform.

2.5 Approach to qualify datasets

The vast amount of publicly available datasets raises the need for a methodology to filter and select those that are consistent with the use cases of this project. Based on the literature review from Section 2.3 and the requirements from work packages WP2 and WP5 of this project as discussed in Section 2.4, a methodology is proposed for the qualification of datasets. It is important to note that the descriptors for data quality were not finalized in T2.3 and T2.4 at the time of developing this methodology. As a result, the approach presented here has been designed to be agnostic to specific quality metrics, ensuring flexibility and adaptability as the project evolves.

The overall methodology follows a two-stage process. The first stage is the "high-level qualification methodology" (see Section 2.5.1) which primary objective is to effectively identify datasets that are relevant for the project's use cases. This is achieved through the evaluation of a dataset based on five criteria, which were identified through a literature review. The second stage is the "Detailed-level qualification methodology" (see Section 2.5.2). This stage involves a more in-depth analysis according to the specific use case and under various qualification categories and quality metrics.

2.5.1 High-level Qualification Methodology

The High-level Qualification Methodology is designed to provide a rapid and efficient screening process for evaluating datasets. Its primary purpose is to quickly identify datasets that are relevant and suitable for further detailed analysis. This methodology ensures that only dataset with potentially high-quality are selected for in-depth evaluation, thereby optimizing resource allocation and time management. The High-level Qualification Methodology involves a systematic evaluation of datasets against a set of predefined essential criteria. Each dataset is assessed to determine its suitability based on these criteria, which are critical for ensuring the dataset's relevance, completeness, and usability. The High-level Qualification Methodology employs a binary scoring system, where each criterion is scored as "True" or "False." where the overall score for the dataset against all the criteria is based on "Pass" or "Fail". This straightforward approach allows for an efficient evaluation process, quickly identifying datasets that meet the essential quality criteria. In the following sections, the essential criteria and scoring for datasets will be presented.

2.5.1.1 ESSENTIAL High-level Qualification METHODOLOGY CRITERIA

The **Essential Screen Criteria** serve as the initial filters to quickly assess datasets for their relevance and usability in further analysis. These criteria, identified through the literature review (Section 2.3) and project requirements, ensure that only high-quality datasets proceed to detailed evaluation. By applying a binary "pass" or "fail" scoring system, the screening process streamlines dataset selection, optimizing time and resource allocation.

The agreed-upon criteria for screening include **Relevance**, **Coverage**, **Accessibility**, **Documentation**, **Metadata** and **Licensing**. Each criterion addresses a specific quality dimension critical for ensuring that datasets meet the project's needs and align with use case requirements.

Definitions for these criteria will follow, outlining their importance and evaluation dimensions.

Relevance

A criterion used to assess datasets is whether the dataset contains data that is relevant for the use cases of this project. The relevancy of a dataset is evaluated based on the following criteria and as seen in Table 2 : Relevance dimensions and criteria:

- ODD coverage: If the ODD is given for a specific use case, it can be used a guideline for dataset assessment by verifying that the dataset contains data matching the ODD and thus aligning with the use case. This includes the geographic properties, such as whether traffic data is collected in urban, rural or highway areas, environmental conditions, such as weather and lightning, as well as temporal properties such as time of day.
- Scenario availability: In the absence of an ODD description but aligning with a project's use case, the existence of scenarios in a dataset is used to assess its relevancy. The more relevant scenarios are available in the dataset, the higher the relevancy of the dataset for a particular use case.

Table 2 : Relevance dimensions and criteria

Dimension	Criteria
ODD coverage	True/False (if the relevant ODD information exists in the dataset)
Scenario availability	True/False (if the relevant scenario(s) exists in the dataset)

Coverage

The Coverage aspect checks whether the dataset contains the critical fields and sufficient amounts of data needed for robust analysis as well as with later usage in CCAM scenarios and related tools. Coverage is evaluated through the following dimensions:

- **Key Fields:** The dataset must include essential attributes that form the basis of the CCAM scenario analysis as seen in Table 3: Coverage dimensions and criteria. These include:
  - **Timestamps:** Accurate timestamps are essential for the synchronisation and matching of events, reconstruction of driving scenarios, and alignment across multiple data streams (e.g. sensor readings, vehicle dynamics). High-frequency timestamps enable detailed temporal resolution, which is critical for analysing complex scenarios such as near-collision or rapid environmental changes.
  - **Location Data:** Accurate and precise geospatial information is essential for understanding vehicle positioning and/or other traffic participants, movement, and interaction with surrounding environment. It must also include the location details of infrastructure-related devices such as traffic lights, road signs, and cameras or roadside units, which interact with vehicles. Geospatial accuracy

should meet high standards, such as sub-meter precision, to support localisation and mapping tasks in dense, urban, or dynamic environments.

- **Usability:** The dataset must support usability by minimising the gaps in the provided data. For high-level methodology only for screening purpose, missing data records should be flagged, and percentage needs to be calculated.

Table 3: Coverage dimensions and criteria

Dimension	Criteria
Timestamps	True/False (if the timestamp attribute exists in the dataset)
Location Data	True/False (if the location attribute exists in the dataset)
Usability	The attributes of timestamps and location cannot have missing values. For the remaining attributes, missing values overall cannot be higher than 5%.

Accessibility

The Accessibility aspect evaluates whether the dataset is readily available and usable for processing, analysis and integration into the project and its tools. Ensuring accessibility is crucial for the smooth execution of project workflows and effective utilisation of the dataset. The aspect is assessed based on the following criteria:

- **Availability:** The dataset should be hosted on trusted platforms or repositories with regular uptime to ensure interrupted access. A dataset should be easily accessed via standard protocols (e.g. HTTP, FTP, or APIs) without excessive delays or any technical difficulties. Access must be granted through a transparent and a simple process, such as direct download, account registration, or approval workflows.
    - **Historical Data:** The dataset should support access to historical data in at least one of the alternatives: (i) manual download – availability of CSV, JSON files or other standard formats for direct download to facilitate quick access to structured data for analysis. (ii) programmatic access – through APIs or other integration methods such as Python SDKs or REST APIs), to facilitate automated data retrieval.
    - **Real-time Data:** Datasets provide real-time information (e.g. sensor feeds, GPS, environmental data etc.) that must support streaming or querying via robust APIs or WebSocket protocols. Infrastructure for real-time data delivery should be reliable, ensuring low latency and minimal interruptions during data collection and processing. This is important for decision making and system testing where immediate access to data is needed. Real-time access should include methods/mechanisms for synchronising data with timestamps to enable seamless integration with other project components that may be collecting or analysing data simultaneously.
- **Format Compatibility:** The dataset should be provided in widely used, machine readable formats that are compatible with common tools to facilitate ease of processing, analysis and later for modelling. This ensures easy to integrate and simplifies processing pipelines (e.g., formats like CSV, JSON, Parquet or HDF5). For datasets containing multimedia components (e.g. video, images, or point clouds), standard formats such as H.265

MPEG/AVI (for video), PNG/HEIF (for images), LAS/E57/PCD/PLY (for point clouds) etc., this criterium should be used to ensure compatibility with tools and frameworks.

- **Infrastructure and Scalability:** The dataset hosting platform needs to provide a scalable infrastructure to handle varying levels of demand such as bulk downloads for large datasets or concurrent connections for real-time streams.
- **Access Permissions:** The platform/website used to download the data should provide clean and clear access permissions and specify how the dataset can be retrieved.

Table 4: Accessibility dimensions and criteria

Dimension	Criteria
Availability (Historical Data)	True/False (if the timestamp attribute exists in the dataset)
Availability (Real-time Data)	True/False (if the location attribute exists in the dataset)
Format Compatibility	True/False (is the format of the dataset are in the acceptable format list, provided by WP4)
Infrastructure and Scalability	True/False (if the platform can handle multiple concurrent calls for big data extraction)
Access Permissions	True/False (if the provider/platform documented the access permissions clearly and transparently)

Documentation

The documentation quality of a dataset is evaluated based on several dimensions. These dimensions ensure that the dataset is comprehensively described and that its use is supported by clear, detailed, and accessible resources. Below is the structured evaluation of the documentation dimension. A dataset is evaluated based on the following criteria:

- **Codebook Availability**: A codebook provides essential details about the dataset structure and how to interpret its contents.
- **Format Documentation:** Clear documentation about the dataset's format ensures compatibility and ease of integration.
- **Website Documentation:** A dedicated website simplifies dataset access and provides comprehensive usage instructions
- **Published Papers:** Publications provide in-depth insights into the dataset's methodologies and potential applications.

Table 5: Documentation dimensions and criteria

Dimension	Criteria
Codebook Availability	True/False (if the timestamp attribute exists in the dataset)
Format Documentation	True/False (if the location attribute exists in the dataset)
Website Documentation	True/False (is the format of the dataset are in the acceptable format list, provided by WP4)

Published Papers	True/False (if the platform can handle multiple concurrent calls for big data extraction)

Metadata

Metadata is a critical evaluation category since it is mandatory to understand the information in a dataset and the overall context to align with the use cases of this project. A dataset is evaluated based on the following criteria:

- **Metadata availability**: A dataset should have additional information that describes the data in the dataset and how the dataset is collected or generated. This enables the comparison of datasets in an early phase and effective dataset filtering and selection.
- **Metadata completeness**: Alongside the actual dataset, a variety of supplementary information can be provided to contextualize and structure the data. These include descriptive metadata, such as information about the authors, version and date of creation, provenance metadata, such as data sources used for collection or processing history, and annotation metadata, such as object or scenario labelling. The necessity for and scope of the metadata dimension should align with the specific use case.

Table 6: Metadata dimensions and criteria

Dimension	Criteria
Metadata availability	True/False (if metadata is available)
Metadata completeness	True/False (if the relevant information exists in the dataset)

Licensing

The licensing criterion addresses the legal constraints associated with the analysis of a dataset as well as the publication of related results. CCAM datasets are published under a variety of licenses, each requiring the identification of the specific rules and restrictions associated with them, as well as the clarification of dataset ownership, to enable an effective assessment of this aspect.

As part of the high-level methodology, a straightforward approach is necessary to determine whether a dataset is usable. To this end, five distinct licensing *states* have been identified based on a thorough literature review and practical experiences:

1. **Unknown**: The license is not clearly disclosed in the dataset's description or metadata, and/or the dataset owner is not adequately identified.
2. **Closed**: Access to the dataset is completely restricted for SYNERGIES partners.
3. **Restricted to Partners**: Access is closed to the public but open to selected actors who are SYNERGIES partners. This included dataset accessible only for academic partners.
4. **Open with Restrictions**: The dataset is publicly accessible, but its use comes with specific restrictions, such as financial obligations (e.g., royalties) or legal limitations (e.g., prohibition of commercial use).
5. **Open**: The dataset is fully accessible without any identified restrictions.

To maximize the opportunities for SYNERGIES partners to utilize valuable datasets, the licensing criterion is considered TRUE as long as access is open for at least one partner, regardless of any associated restrictions. Thus, datasets classified under the licensing states **(3)** Restricted to Partners, **(4)** Open with Restrictions, or **(5)** Open meet the licensing criteria.

This approach allows datasets with varying access rules and publication restrictions to qualify under the licensing criterion. A more thorough evaluation of the license ruling the dataset will take place during the detailed assessment phase of the methodology.

2.5.1.2 Scoring for datasets

The **High-Level Qualification Methodology** employs a systematic and efficient scoring system to evaluate datasets against the essential screening criteria outlined in Section 2.4.1.1. The purpose of the scoring system is to quickly identify datasets that meet the minimum quality requirements for further detailed analysis, ensuring that resources are directed towards relevant and usable data.

The scoring system uses a **binary approach**, where each dataset is assessed against the predefined criteria and assigned a score of **"Pass" or "Fail"** for each dimension. The key dimensions evaluated include **Relevance**, **Completeness**, **Accessibility**, **Licensing**, **Metadata**, **Documentation**, and **Compatibility**. This straightforward scoring method ensures clarity, transparency, and ease of implementation across diverse datasets.

Each criterion is evaluated as follows:

- **Pass**: The dataset meets the minimum requirement for the given criterion.
- **Fail**: The dataset does not meet the minimum requirement, indicating it may not be suitable for further detailed evaluation.

For a dataset to qualify for the next stage of assessment, it must achieve a **"Pass" score for all mandatory criteria**. This strict screening ensures that only high-quality datasets proceed to the **Detailed Level Qualification Methodology** (Section 2.4.2). In cases where certain criteria are non-mandatory, exceptions may be considered depending on the specific use case and overall dataset relevance.

To validate the robustness of this scoring approach, the **High-Level Qualification Methodology** was applied to an initial selection of **14 datasets**. Each dataset was systematically evaluated against the essential criteria, and the results were recorded in the **D3.1 Excel sheet**. This practical application demonstrated the efficiency of the scoring system in identifying datasets that meet the baseline requirements, while also highlighting areas for potential improvement in dataset quality.

The binary scoring system, combined with the clear and objective criteria, ensures a consistent and scalable approach to dataset screening. It facilitates a rapid yet reliable evaluation process, enabling the SYNERGIES project to integrate high-quality datasets into its workflows and scenario generation processes

2.5.1.3 Apply High-level qualification methodology to 14 datasets

The high-level qualification methodology consisted of applying 6 mandatory criteria to assess different types of datasets (Accident, Synthetic, fixed and moving point of view).

In order to test the usability of the high-level methodology, the methodology has been applied to 14 datasets and showed a success rate of 78.5%, where 3 datasets failed to pass this qualification for the following reasons:

- Dataset number 14 (Fixed point of view, UAVId): the coverage didn't pass because the FPS is provided but not the timestamp nor the location, where data volume is also small.
- Dataset number 5 (Synthetic, CarlaScenes): the accessibility didn't pass because FTP server is down, the Dataset couldn't be downloaded, furthermore, the owners of the dataset do not respond to the community, the link sources are not maintained for 2 years
- Dataset number 6 (Synthetic, VIRTUALKITTI 2):
  - the relevance didn't pass because the infrastructure information (lane marking, traffic signs, road decomposition, etc.) are not provided, it needs effort to reconstruct them from the images.
  - The coverage didn't pass because 5 scenes of 30 min is considered not enough to create an interesting volume of scenarios

The following Figure 9: High level Methodology Results shows the results of this high-level qualification methodology:

		Criteria						
		Mandatory	Mandatory	Mandatory	Mandatory	Mandatory	Mandatory	
	Type	Relevance	Coverage	Accessibility	Documentation	Metadata		Score
Dataset Name 1 : GIDAS	Accident Data	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	Pass
Dataset Name 2 : TASC	Accident Data	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	Pass
Dataset Name 3 : DLR Urban Traffic Dataset	Fixed point of view (drone, infra, etc..)	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	Pass
Dataset Name 4 : open dataset Road Safety Data - STATS19, 2020	Accident Data	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	ss
Dataset Name 5 : CarlaScenes	Synthetic Data	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	Fail
Dataset Name 6 : VIRTUAL KITTI 2	Synthetic Data	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	Fail
Dataset Name 7 : UDRIVE	Moving point of view (On borad ego centered)	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	Pass
Dataset Name 8 : VOIESUR	Accident Data	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	Pass
Dataset Name 9 : highD	Fixed point of view (drone, infra, etc..)	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	Pass
Dataset Name 10 : Argoverse 2	Moving point of view (On borad ego centered)	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	Pass
Dataset Name 11 : Standing General Order on Crash Reporting	Accident Data	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	Pass
Dataset Name 12 : RADIATE	Moving point of view (On borad ego centered)	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	Pass
Dataset Name 13 : highD	Fixed point of view (drone, infra, etc..)	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	Pass
Dataset Name 14 : UAVId	Fixed point of view (drone, infra, etc..)	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	Fail
							Success Rate:	78,57%

Figure 9: High level Methodology Results

As seen in Figure 9: High level Methodology Results, several of the selected 14 datasets did not meet the criteria for various reasons:

1- "VIRTUAL KITTI 2" Dataset:

- **Relevance Concerns**: Key infrastructure details, such as lane markings, traffic signs, and road decomposition, are not included in the dataset, limiting its usefulness for certain applications.
- **Insufficient Data Volume**: The dataset contains only 5 scenes, which represent approximately 30 minutes of driving. With a 10 Hz recording frequency, it includes just 21,260 images in total. This amount of data is considered

insufficient to create a diverse and interesting range of scenarios for testing or training.

2- "CarlaScense" Dataset

The dataset fails to meet accessibility criteria, as its server is currently down, the dataset owner has not responded to the community, and the source links have not been maintained for over two years.

3- **UAVId Dataset**:

This dataset lacks important information, such as timestamps and location data, and has a small data volume, limiting its applicability for thorough analysis.

2.5.2 Detailed Level Qualification Methodology

The **Detailed Level Qualification Methodology** represents the next step following the **High-Level Qualification Methodology** outlined in Section 2.4.1. While the high-level methodology serves as an initial screening tool to identify datasets meeting essential quality criteria, the detailed methodology involves a more in-depth evaluation of datasets. This stage provides a comprehensive assessment of dataset quality, ensuring alignment with project requirements and the needs of the **SYNERGIES** platform.

This methodology evaluates datasets across a set of **key quality categories** identified during the literature review and requirements collection phases. These categories include, but are not limited to, **accuracy, completeness, consistency, timeliness, coverage, and relevance**. Unlike the binary scoring system used in the high-level methodology, the detailed assessment employs **quantitative metrics and thresholds** to provide a granular evaluation of each dataset.

It is important to note that this section presents a **working version** of the detailed methodology, which is still under development. As the **SYNERGIES** project progresses, additional requirements and inputs from other tasks and work packages—such as Task 2.3 (Interoperability Requirements), Work Package 5 (Scenario Generation), and future deliverables—will be essential for refining this methodology. The ongoing nature of this work ensures that the methodology remains adaptable and responsive to evolving project needs.

Future updates to this methodology will focus on integrating specific quality descriptors, metrics, and thresholds as they become available. This iterative approach allows for continuous improvement, ensuring that the **Detailed Level Qualification Methodology** provides a robust, reliable, and comprehensive framework for evaluating datasets used in scenario generation, testing, and validation within the **SYNERGIES** platform.

In summary, this section outlines the foundational elements of the detailed methodology while acknowledging that further refinements and updates will be required to address the full scope of project requirements.

2.5.2.1 ESTABLISH GENERAL QUALITY CATEGORIES

To evaluate the quality of a given dataset and allow for the selection of relevant ones that align with the project's use cases, seven data quality categories were identified for in-depth analysis. The criteria relevance, coverage, licensing, which are employed in the high-level qualification methodology and introduced in Section 2.4.1.1 to evaluate datasets, are also used within the detailed-level qualification methodology as categories. Furthermore, the following additional categories are used for dataset assessment:

1. **Completeness** is used to assess whether a dataset contains all relevant information [10]. In terms of real-world traffic data collected via (quasi-) infrastructure and vehicles, as well as synthetically generated data this relates to the availability of information to represent scenarios using the five-layer model [18]. This includes information about the road, traffic infrastructure, their temporal modification, objects and the environment. In terms of accident data this category is used to assess whether the dataset includes information about all critical fields that are required for scenario analysis, including time of accident, the location and the accident severity.
2. **Accuracy** is used to assess whether the data is error-free and adheres to its intended truth or standard. It can be defined as the measure of the correctness, reliability, and exactness of data, typically quantified as the ratio of correct values to the total values in a dataset. In terms of real-world traffic data, accident data, as well as synthetically generated data, this relates to the correct representation of real-world concepts, phenomena and events.
3. **Consistency** is used to assess the extent to which data maintain uniformity, coherence, and adherence to semantic rules across observations, formats, and time. It ensures compatibility with previous datasets and logical, temporal, spatial, and structural alignment, to enable reliable use within the project's scope.
4. **Timeliness** is used to assess whether the data are processed in a timely manner, kept up to date and suitable for their intended use given their age. It considers factors such as the rate of change of the system (volatility), the frequency of data updates (currency) and the age of the data in relation to its relevance and application.

It has been determined that data quality categories, their associated metrics, and thresholds may vary across different datasets. To address this, we have identified four distinct dataset types, each with metrics tailored to its specific characteristics and environment. These dataset types are:

- **Fixed Point of View dataset**: A dataset collected from a stationary or fixed location, where the sensor or recording device does not move. Examples include surveillance camera footage or environmental monitoring systems from fixed points
- **Moving Point of View Dataset** : A dataset captured from a moving sensor or recording device, such as cameras mounted on drones, vehicles, or handheld devices. These datasets often involve dynamic perspectives and changing fields of view.
- **Accident Dataset**: dataset specifically focused on recording and analysing incidents or accidents, such as road traffic collisions or workplace safety incidents. These datasets often contain event-specific data like timestamps, locations, and causative factors.
- **Synthetic Dataset**: dataset generated artificially, often through simulation or computational models, rather than collected from the real world. These datasets are used to model scenarios, train machine learning models, or fill gaps where real-world data is unavailable

2.5.2.2 Quality metrics and Threshold for Fixed Point of View datasets

Completeness

Scope: The completeness metric is intended to evaluate whether the data set includes all the elements that are required to describe the CCAM scenario under study. For data types related to fixed methods, such as drones or infrastructure, the following metrics are used to monitor and evaluate accuracy, which is necessary to ensure the accuracy of the CCAMs landscape in providing a detailed picture.

Table 7: Metrics definition and requirements for Completeness for Fixed Point of View datasets

Metrics	Definition	Requirements
Temporal Completeness	This metric is intended to measure whether data has been consistently captured over time. It ensures that the dataset includes the entire time series, with no differences in temporal coverage.	The time difference between the start and end date periods should be calculated and multiplied by the sample size. Example. For data collected hourly over 2 days, the data must be 48 rows (i.e. 24 hours x 2 days). The actual number of rows should match the expected number, and any difference indicates missing data.
Attribute Completeness	This metric determines the number of fields in the data set, ensuring that all required fields are included and presented in each data entry.	The required data fields must be defined first before this metric can be analysed. A list of all required attributes should be available for comparison to ensure completeness.
Spatial Completeness	This metric analyses the extent to which the location of interest will completely span the dataset and ensures that there are no differences in the coverage of the target locations.	For data sets with more than one site, each field must be checked as a complete space to ensure that there is no missing data in any of the regions captured in the data set.
Data Annotation Completeness	This metric measures the proportion of data points recorded or labelled, which is particularly important for supervised machine learning projects or analytics that require data interpretation.	The percentage of labelled data should be checked against the entire dataset to ensure that a sufficient portion of the data is labelled.
Annotation Categorical Completeness	These metric checks whether the data set includes all annotation classes required for analysis.	Predefined interest groups are required to compare with those in the data set.
Trajectory Completeness	This metric ensures that the data set includes complete trajectories, which are important for understanding object motion or its behaviour over time.	A threshold for complete access (e.g. minimum number of waypoints) should be defined based on specific requirements.

Below are the formulas and thresholds used evaluate and calculate the completeness metrics.

Table 8: formulas and thresholds used evaluate and calculate the completeness metrics for Fixed Point of View datasets

Metrics	Formula	Threshold
Temporal Completeness	$\frac{Number\ of\ Valid\ Data\ Points\ (Timestamps)}{Expected\ Data\ Points\ (by\ sampling\ frequency)} x100$	100%
	Rationale: This metric is needed for time-series data to ensure that all required timestamps are captured without records. Differences in timestamps can lead to incomplete or inconsistent event assessments, reducing the confidence in the conclusions drawn from the data, so the threshold is defined as 100% .	
Attribute Completeness	$\frac{Number\ of\ Fields\ with\ No\ Mis\sin g\ ,\ Valid\ Values}{Total\ Required\ Fields} x100$	100%
	Rationale: Ensuring that all necessary fields are available assures that the database has all the necessary information needed for an accurate analysis and reconstruction Missing fields can lead to incomplete analysis or indicate that the information is not well understood, so the threshold is defined as 100% .	
Data Annotation Completeness	$\frac{Number\ of\ Annotated\ Points}{Total\ Data\ Points} x\ 100$	≥95%
	Rationale: A valid specification ensures that the data set is useful for applications that rely on recorded data. The higher absolute description increases the utility of the dataset for model training and analysis. For the threshold, some allowance (5%) is made for omitted or invalid specifications due to human or system error.	
Annotation Categorical Completeness	$\frac{Number\ of\ Categories\ \Pr e\ sent\ in\ an\ Attribute}{Total\ Number\ of\ Required\ Categories} x100$	100%
	Rationale: All required categories must be documented to fully represent the dataset. Missing categories may result in incomplete analyses or underrepresentation of important features. The required classes must exist in the data structure; Therefore, the threshold is defined as 100%.	
Trajectory Completeness	$\frac{Number\ of\ Complete\ Trajectories\ (min.\ waypoints)}{Total\ Number\ of\ Trajectories} x100$	≥75%
	Rationale: Incomplete trajectories may limit the usefulness of the dataset for situations where object movements or interactions must be tracked. Given that a larger portion of the route has a larger number of waypoints, it is ensured that the data set supports the scenario reconstruction. Allowing for about 25% of incomplete trips allows for flexibility in road information (e.g., nuisance signs or short messages) and ensures that most roads are usable.	

Accuracy

Scope: The accuracy aspect checks whether the data accurately represents the real-world values or events it is intended to capture. Accurate datasets are essential to generate reliable insights and build robust models in different contexts. The following metrics are used to assess accuracy:

Table 9: Metrics definition and requirements for Accuracy for Fixed Point of View datasets

Metrics	Definition	Requirements
Temporal Accuracy	It accurately measures the recorded timestamp at an expected or defined time.	The true moment must first be defined, and the overlap or deviation must be quantified.
Spatial Accuracy	It measures the deviation of the recorded GPS coordinates from the actual location on the ground with fixed observation points (e.g. fixed infrastructure or sensors).	Actual locations of fixed locations should be available for comparison.
Sensor Measurement Accuracy	Study the deviation of recorded sensor data from known ground truth values.	The ground truth must be available for the sensor measurement accuracy, or it must be defaulted for comparison.
Unit Accuracy	Verify that the recorded values are in the correct units and conform to the required measurement parameters.	A default acceptable range and units should be set to detect differences and out-of-range values.
Non-default Valued Attributes Accuracy	It ensures that the required attribute contains valid parameters rather than default or placeholder values (e.g., "N/A", "0", "-999", etc.).	Default values must be specified and omitted in fields required by default rules or ranges.
Pixel Accuracy	It measures the agreement between pixel values in image data and ground truth or expected values (e.g., colour, intensity, spatial fidelity, etc.) during calibration testing.	This can take the form of histogram-based methods (e.g., entropy or standard deviation per pixel bin) to validate pixel assembly and quality.

Below are the formulas and thresholds used evaluate and calculate the accuracy metrics.

Table 10: formulas and thresholds used evaluate and calculate the accuracy metrics for Fixed Point of View datasets.

Metrics	Formula	Threshold
	$\frac{Total\ Valid\ Recorded\ Time\ (within\ expected\ range)}{Expected\ Total\ Time} x100$	≥99%

Timestamp Accuracy	**Rationale:** The 100% requirement does not apply here due to possible minor delays that add up during recording; The 99% threshold ensures reliability by accounting for trivial errors.	
Spatial Accuracy	$\frac{\sqrt{\sum(Latitude\ Deviation^2\ +\ Longitude\ Deviation^2)}}{Number\ of\ Points}$	≤ 0.5 meter
	Rationale: The addition of the Pythagorean theorem to calculate the division ensures that the space is symmetrical. Strict limits are needed for autonomous vehicles where small errors in a location can have important consequences.	
Sensor Measurement Accuracy	$100 - \frac{\sum\lvert Measured\ Value\ -\ True\ Value\rvert}{Sum\ of\ True\ Values}x100$	≥95%
	Rationale: The 95% limit balances high accuracy with acceptable variations due to sensor noise or environmental factors. The theory highlights the stark difference between overcontrol and underestimation.	
Unit Accuracy	$\frac{Number\ of\ Correct\ Unit\ Values}{Total\ Number\ of\ Values}x100$	100%
	Rationale: There is no tolerance for incorrect units, as inconsistent units can lead to incorrect decisions and results. To ensure that, the units must be cleaned prior to use.	
Non-default Valued Attributes Accuracy	$\frac{Number\ of\ NonDefault\ Records}{Total\ Number\ of\ Records}x100$	≥99%
	Rationale: Although 100% is ideal, a small error (1%) is allowed to account for unavoidable edge cases such as occasional data gaps during compilation A high emphasis period who are still true.	
Pixel Accuracy	$\frac{Number\ of\ Correct\ Pixels\ (by\ \exp e\ cted\ value)}{Total\ Pixels}x100$	≥95%
	Rationale: A higher threshold for the main function of pixel stability in absolute terms. The features are basically in the comparison of pixel values with predicted values based on colour depth, spatial complexity, or ground truth context.	

Consistency

Scope: Consistency ensures that data maintains logical, temporal, spatial, and organizational consistency in viewing and reading. This aspect is critical for ensuring that data sets can be reliably used for situational analysis, modelling and decision-making processes. The following metrics are used to assess stability:

Table 11: Metrics definition and requirements for Consistency for Fixed Point of View datasets

Metrics	Definition	Requirements

Temporal Consistency	It checks whether the data exhibit an expected logical or physical pattern over time without sudden or unexpected anomalies.	The maximum allowable rate of change Δ_{max} needs to be defined for key numerical attributes. Definition of Δ_{max} is to be based on physical or logical limits in measured attributes. Domain knowledge or historical data can be used to set realistic requirements for acceptable change rates. Later, identification of transitions exceeding Δ_{max} will be done.
Spatial Consistency	It measures the extent to which spatial information overlaps and consistently integrates in observations or time series.	First, a reference map or data set is needed to capture the accuracy of the design. The vertical deflection should be negligible (0.5 m) for stationary sensors or 1 meter – depends on how it is defined in "spatial accuracy". For spatial data capturing moving areas (e.g. drone footage), constant overlap between frames ensures accurate spatial coverage. For drone imagery, the area covered must meet a preset threshold (95% coverage in observations).
Sensor Consistency	It ensures that sensor readings are stable and reliable under the same environmental and operating conditions.	It is important to define an acceptable variable for the sensors.
Annotation Consistency	It considers whether the dataset description is applied consistently to records or conforms to predefined criteria.	Annotation guidelines or standard classifications for categories are needed.
Format Consistency	Ensure that all data set records conform to the default data structure, including fields (numbers, datetime, etc.) and allowed formats.	All attributes (e.g. ISO 8601 for time stamps) should be defined as the gold standard.
Object Resolution Consistency	This metric ensures that the features captured in the dataset meet the minimum requirements needed for the intended analysis or use.	Minimum resolution requirements can be defined first before this metric is viewed. The data set must contain at least 95% of the frames that meet the specified resolution threshold.

Below are the formulas and thresholds used evaluate and calculate the consistency metrics.

Table 12: formulas and thresholds used evaluate and calculate the consistency metrics for Fixed Point of View datasets.

Metrics	Formula	Threshold
Temporal Consistency	$\frac{Number\ of\ Transitions\ Within\ Defined\ \Delta_{max}}{Total\ Transitions} x100$	≥98%
	Rationale: Transition is a change between consecutive data points. $\lvert Value_{t+1} - Value_t \rvert$ will be calculated to get the absolute difference between consecutive values. Then the results will be compared with predefined Δ_{max}. Upper limit (98%) of the range but ensuring minimal variation when accounting for noise or acceptable excesses. It ensures smooth transformation of data, which is important for time-series analysis, traffic behaviour modelling, and contextual applications.	
Spatial Consistency	For stationary sensors: $\frac{Number\ of\ Consistent\ Locations}{Total\ Recorded\ Locations} x100$ For drone footage: $\frac{Number\ of\ Overlapping\ Areas\ Across\ Frames}{Total\ Recorded\ Locations} x100$	For stationary sensors: 100% For drone footage: ≥95%
	Rationale: A 100% threshold is mandatory for stationary sensors because any deviation indicates an error in data logging or sensor placement. Deviations here should be negligible (0.5 m or 1 m). For drone photography, it's a threshold ≥95% have some variance due to environment, sensor noise, or motion artifacts. Here, the acceptable overlap threshold is used as 95% - the spatial overlap between consecutive frames.	
Sensor Consistency	$\frac{Number\ of\ Stable\ Readings}{Total\ Sensor\ Readings} x100$	≥98%
	Rationale: Logic Stable reading: The reading does not deviate beyond an acceptable range of variation (e.g. ±2% under the same conditions). The 98% threshold ensures that the sensors provide highly reliable readings and allows for minor variations due to sensor accuracy limitations or environmental noise Status is important to ensure that the data is accurate and reliable especially for safety-critical applications.	
Annotation Consistency	$\frac{Number\ of\ Records\ with\ Consistent\ Annotations}{Total\ Annotated\ Records} x100$	≥98%
	Rationale: A threshold of ≥ 98% ensures high precision in the data set, reduces uncertainty and improves model reliability. The 2% allowance is for occasional differences due to human error or edge issues.	
Format Consistency	$\frac{Number\ of\ Records\ Conf or \min g\ \ to\ Expected\ Format}{Total\ Records} x100$	100%
	Rationale: Format inconsistencies can lead to errors in data processing, analysis, and modelling. The 100% threshold is necessary to ensure that all records follow the default formats.	

Object Resolution Consistency	$\frac{Number\ of\ Frames\ Meeting\ Min.\ Resolution}{Total\ Number\ of\ Frames} x100$	≥95%
	Rationale: In many applications, high-resolution data are critical for accurate analysis, especially when details or small features need to be detected. This metric ensures that the data set meets the resolution standards required for effective processing and analysis. The 95% threshold ensures that most data sets are usable while allowing for small discrepancies.	

Timeliness

Scope: The Timeliness evaluates the extent to which a dataset is recent, enabling its effective use as scenario dataset. It indicates how up to date the dataset is to ensure the contained scenarios are still relevant for analysis and usable for the intended purposes. The following metrics are used to assess Timeliness.

Table 13: Metrics definition and requirements for Timeliness for Fixed Point of View datasets

Metrics	Definition	Requirements
Data freshness	This metric measures the degree to which data is updated within a specific time window. It ensures that the scenarios captured the traffic behaviours, and the sensor data can be used for relevant analyses	The freshness window needs to be defined (in minutes, hours, days, years…) depending on the data and the location.
Sensor EoL	This metric measures the timeliness of sensors used in captured dataset. It ensures that the sensor data is usable and reproducible	Need to categorize sensors into critical/non-critical and focus on sensors where technology updates have made the old technology obsolete

Below are the formulas and thresholds used evaluate and calculate the timeliness metrics.

Table 14: formulas and thresholds used evaluate and calculate the timeliness metrics for Fixed Point of View datasets

Metrics	Formula	Threshold
Data freshness	$\frac{Current\ time - time\ of\ last\ update}{Freshness\ window}$	≤1
	Rationale: Ensuring the dataset is maintained and constantly updated within a predefined freshness time window.	
Sensor EoL	$Current\ time - latest\ time\ of\ EoL\ time\ of\ critical\ sensors\ used$	≤5 years

	Rationale: Ensuring the raw sensor data formats and annotations are still relevant and reproducibility of data capture.

Relevancy

Scope: Relevancy evaluates the extent to which a dataset is relevant for representations of real-world scenarios (of the selected list) and for scenario extraction task. The following metrics are used to assess Relevancy.

Table 15: Metrics definition and requirements for Relevancy for Fixed Point of View datasets

Metrics	Definition	Requirements
Relevant Env Conditions	This metric measures the degree to which dataset cover relevant environmental conditions comparing to the conditions of interest	Need to define ODD environmental parameter ranges, and extract from the dataset environment related labels
Relevant scenarios	This metric measures the degree to which dataset captures relevant scenario(s)	Need to define what are relevant scenarios
Proportion of relevant data	These metric measures percentage of relevant scenarios related data contained within the dataset	Need to define what are relevant scenarios

Below are the formulas and thresholds used evaluate and calculate the relevancy metrics.

Table 16: formulas and thresholds used evaluate and calculate the relevancy metrics for Fixed Point of View datasets

Metrics	Formula	Threshold
Relevant Env Conditions	$Intersection\ (dataset\ env\ ranges, required\ env\ params\ ranges)$	>0
	Rationale: Ensuring the dataset is usable for the ODD.	
Relevant Scenarios	$Number\ of\ relevant\ scenarios$	>0
	Rationale: If a scenario is included, it should be in the list of relevant scenarios and have valid/complete data to describe (e.g. trajectory shall cover the key events of the scenario).	
Proportion of relevant data	$\frac{Total\ of\ time\ durations\ of\ relevant\ data}{Dataset\ recording\ duration} \times 100$	≥50%
	Rationale: To balance the amount of efforts used to identify/extract relevant scenarios from a dataset and the extracted volume of relevant data.	

Coverage

Scope: Coverage evaluates the extent to which a dataset includes all relevant data to serve its intended purposes. For the fixed point of view datasets, this includes how the sensor settings capture the traffic scenes, the coverages for timeseries analyses

Table 17: Metrics definition and requirements for Accuracy for Fixed Point of View datasets

Metrics	Definition	Requirements
Target Area Coverage	This metric measures the degree to which the dataset covers the geographical or physical area of interest.	Need to define the geographical and physical area of interest for the scenarios, and to extract geographical information from the dataset
Temporal Relevancy	This metric measures the degree to which data reflects the time periods critical to the intended analysis	Need to define the expected time duration and frequencies
Event identification	This metric measures the degree to which data captures key events of interest scenarios	Need to define the required types of events for each and every scenario of interest
Measurement coverage	This metric measures how many sensors are used to capture the data	Need to extract information about sensor setups from the dataset

Below are the formulas and thresholds used evaluate and calculate the coverage metrics.

Table 18: formulas and thresholds used evaluate and calculate the coverage metrics for Fixed Point of View datasets

Metrics	Formula	Threshold
Target Area Coverage	$\frac{Area\ of\ interest\ covered}{Total\ target\ area} \times 100$	≥80%
	Rationale: ensures that the dataset captures specific spatial regions required. Example: How far from the intersection a dataset has to cover to provide meaningful intersection related scenarios	
Temporal Relevancy	$\frac{Critical\ time\ period\ covered}{Total\ required\ time\ periods} \times 100$	≥90%
	Rationale: ensures the dataset captures data from the required time periods, avoiding irrelevance due to outdated or improperly timed observations.	
	$\frac{Number\ of\ event\ types\ appear\ in\ the\ dataset}{Required\ number\ of\ event\ types} \times 100$	100%

Event identification	**Rationale:** guarantees that the dataset focuses on capturing meaningful events	
Measurement coverage	$\frac{Total\ sensor\ hours}{Dataset\ recording\ duration\ (h)} \times 100$	≥50%
	Rationale: Ensure that the interested scenarios have been captured from different perspectives	

2.5.2.3 Quality metrics and Threshold for Moving Point of View datasets

Completeness

Scope: *Ensuring mandatory data for all 5 layers of scenario description are included in addition to timestamps*

Table 19: Metrics definition and requirements for Completeness for Moving Point of View datasets

Metrics	Definition	Requirements
Temporal Completeness	This metric measures the degree to which data is consistently captured over the observation period. It ensures that the dataset includes a complete time series, with no gaps in the temporal coverage.	The time difference between the start and end datetime needs to be calculated and multiplied by the sampling range. E.g. for hourly collected data over 2 days, there must be 48 data rows (i.e. 24 hours x 2 days). The actual number of rows should match the expected number, with any discrepancies indicating missing data.
Attribute Completeness	This metric quantifies the proportion of available attributes in the dataset, ensuring that all required attributes are included and presented in each data entry.	The required data fields must be pre-defined before this metric is checked. A list of all essential attributes must be available for comparison to ensure completeness.
Data Annotation Completeness	This metric measures the proportion of data points that have been annotated or labelled, which is particularly important for supervised machine learning tasks or analysis requiring data interpretation.	Required data fields needs to be pre-defined before this metric is checked.

Annotation categorical completeness	This metric assesses whether the dataset includes all categories of annotations that are required for the analysis.	Pre-defined categories of interest are needed to compare their presence in the dataset.

Table 20: formulas and thresholds used evaluate and calculate the Completeness metrics for Fixed Point of View datasets

Metrics	Formula	Threshold
Temporal Completeness	$\frac{Number}{Total\ target\ area} \times 100$	≥80%
	Rationale: ensures that the dataset captures specific spatial regions required. Example: How far from the intersection a dataset has to cover to provide meaningful intersection related scenarios	
Attribute Completeness	$\frac{Critical\ time\ period\ covered}{Total\ required\ time\ periods} \times 100$	≥90%
	Rationale: ensures the dataset captures data from the required time periods, avoiding irrelevance due to outdated or improperly timed observations.	
Data Annotation Completeness	number of annotated points / total data points x 100	95%
	Rationale: Sufficient annotation ensures that the dataset is usable for tasks that rely on labelled data. A high degree of annotation completeness increases the utility of the dataset for model training and evaluation. For threshold, some allowance (5%) is given for unannotated or invalid annotations due to human or system errors	
Annotation categorical completeness	number of categories present in an attribute / total number of required categories x 100	≥50%
	Rationale: For the dataset to be fully representative, all expected categories must be annotated. Missing categories could lead to incomplete analysis or underrepresentation of important features. The required categories must be in the dataset; thus, the threshold is defined as 100%.	

Accuracy

Scope: *Ensuring data has attributes that correctly represent the true value of the intended attribute of a concept or event in a specific context of use.*

Table 21: Metrics definition and requirements for accuracy for Moving Point of View datasets

Metrics	Definition	Requirements
object classification accuracy	the accuracy of the object classification system compared to a ground truth reference (potentially per class)	Ground truth must be available or must be predefined for comparison to validate the sensor measurements.

Sensor Measurement Accuracy	Measures the deviation of recorded sensor data from known ground truth values.	
Range Accuracy	Measures the proportion of recorded values that fall within a predefined minimum and maximum range.	The predefined min and max ranges must be carefully defined based on domain-specific requirements. Any values outside the range should be flagged for review.
Non-default valued attributes accuracy	Ensures that required attributes contain valid measured values rather than default or placeholder values (e.g., “N/A”, “0”, “-999”, etc.)	Identification and exclusion of default values in required fields by predefined rules or ranges is necessary.
Coherent values	Amongst signals with redundancy are values coherent. (e.g. speed x yaw_rate ≈ lateral_accel ; dSpeed/dT ≈ longitudinal_accel)	list of "redundant signals" (i;e) relations that can be found amongst signal, allowing to check for coherence

Below are the formulas and thresholds used evaluate and calculate the metrics for accuracy for moving points of view datasets.

Table 22: formulas and thresholds used evaluate and calculate the metrics for accuracy for moving points of view datasets.

Metrics	Formula	Threshold
object classification accuracy	100*number of objects with valid classification/ number of objects in total	≥80%
	Rationale: overview of how well the system discriminates between the different object types	
Sensor Measurement Accuracy		≥90%
	Rationale: A threshold of 95% balances high precision with acceptable variability due to sensor noise or environmental factors. The formula emphasizes the absolute difference to handle both over- and underestimation.	
Range Accuracy	Number of values within range / total number of values x 100	99%
	Rationale: A threshold of 99% ensures that most values are within the acceptable range while allowing for small deviations.	
Non-default valued attributes accuracy	number of categories present in an attribute / total number of required categories x 100	≥50%

	Rationale: For the dataset to be fully representative, all expected categories must be annotated. Missing categories could lead to incomplete analysis or underrepresentation of important features. The required categories must be in the dataset; thus, the threshold is defined as 100%.	
Coherent values	Number of available redundant signals which values are coherent / Number of available redundant signals x100	≥50%
	Rationale: Disagreement between values in redundant signals is a good way to detect errors in a dataset and is, therefore, a useful tool to assess the quality of the dataset.	

Consistency

Scope *Ensuring the collected data is consistent through records of dataset*

Table 23: Metrics definition and requirements for consistency for Moving Point of View datasets

Metrics	Definition	Requirements
data format Consistency	Ensures that all dataset records conform to a predefined data structure, including field names, types (numeric, datetime, etc.) and units and permissible formats.	Predefined data structure (corresponding to documentation and metadata)
Temporal Consistency	Evaluates whether data exhibit expected logical or physical patterns over time without abrupt or unexpected anomalies.	the max allowable rate of change (delta_max) needs to be defined for key numerical attributes. definition of delta_max is to be based on physical or logical limits in measured attributes. Either domain knowledge or historical data can be used to establish thresholds for acceptable change rates.
Sensor consistency	extent to which sensor readings are consistent under the same conditions across time	access to the data
Annotation Consistency	Measures whether the dataset's annotations are applied uniformly across records and align with predefined standards.	Annotation guidelines or a standard taxonomy for categories is needed.

Below are the formulas and thresholds used evaluate and calculate the metrics.

Table 24: formulas and thresholds used evaluate and calculate the metrics for Consistency for moving points of view datasets.

Metrics	Formula	Threshold
Data format compliancy	Number of records respecting data format/total number of records x 100	≥95%
	Rationale: Ensuring that all records respect predefined structure and identify which are not, Not conform record cannot be processed.	
Temporal Consistency	number of transitions within defined delta_max / total transitions x 100	≥99%
	Rationale: Transition is a change between consecutive data points. Absolute difference between Value_(t+1) and Value_t will be calculated to get the absolute difference between consecutive values. Then the results will be compared with predefined delta_max. A higher threshold (98%) ensures minimal anomalies while accounting for rate but acceptable noise or outliers.	
Sensor consistency	stable sensor readings / total sensor readings x 100	95%
	Rationale: verfies that the sensors provides stable and reliable data	
Annotation Consistency	number of records with consistent annotations / total annotated records x 100	≥98%
	Rationale: A threshold of 98% ensures high-quality annotations across dataset, reducing ambiguity and improving the reliability of models. 2% allowance is for occasional inconsistencies caused by human error or edge cases.	

Timeliness

Scope *Ensure that data is up to date*

Table 25: Metrics definition and requirements for Timeliness for Moving Point of View datasets

Metrics	Definition	Requirements
Dataset Freshness	The delay from the date when the dataset was captured until it was available	access to the year of data collection

Below are the formulas and thresholds used evaluate and calculate the metrics.

Table 26: formulas and thresholds used evaluate and calculate the Timeliness metrics for Moving Point of View datasets

Metrics	Formula	Threshold
Dataset Freshness	date(now) - date(collection)	≥95%
	Rationale: give an insight if the knowledge we can extract from dataset is still relevant	

Relevancy

Scope *Ensuring the dataset contains information and scenarios that are useful for the project*

Table 27: Metrics definition and requirements for Relevancy for Moving Point of View datasets

Metrics	Definition	Requirements
Relevant scenarios	Number of scenarios of interest represented	List of relevant scenarios and the toolchain enabling processing the dataset to get the expected output

Below are the formulas and thresholds used evaluate and calculate the metrics.

Table 28: formulas and thresholds used evaluate and calculate the Relevancy metrics for Fixed Point of View datasets

Metrics	Formula	Threshold
Relevant scenarios	Number of relevant scenarios	>0
	Rationale: If a scenario is included, it should be in the list of relevant scenarios and have valid/complete data to describe (e.g. trajectory shall cover the key events of the scenario)	

Coverage

Scope *Ensure that dataset covers a variety of scenarios needed in SYNERGIES project*

Table 29: Metrics definition and requirements for Coverage for Moving Point of View datasets

Metrics	Definition	Requirements
Relevant ODD conditions	Ensure the dataset covers the intersections of following parameters time of day, seasons, weather types, road types	Need to define ODD environmental and road types as parameter ranges

Object coverage	Ensure the dataset contains detection of relevant objects (i.e: pedestrian, car, bike, e-scooter...)	list of relevant objects for the project's use cases

Below are the formulas and thresholds used evaluate and calculate the metrics.

Table 30: formulas and thresholds used evaluate and calculate the coverage metrics for Fixed Point of View datasets

Metrics	Formula	Threshold
Relevant ODD conditions	Number of relevant scenarios	>0
	Rationale: If a scenario is included, it should be in the list of relevant scenarios and have valid/complete data to describe (e.g. trajectory shall cover the key events of the scenario)	
Object coverage	100*number of object category detected in dataset/ number of objects category considered in synergies	>50%
	Rationale: Ensure that dataset contained detection and classification of a large variety of object, including new VRU in urban area	

2.5.2.4 Quality metrics and Threshold for Quality metrics and Threshold for Accident datasets

Category: Completeness

Scope: Ensuring accident data includes all critical fields (e.g., time, location, severity) and key scenarios necessary for meaningful analysis

Table 31: Metrics definition and requirements for Completeness for accident datasets

Metrics	Definition	Requirements
Field Completeness Score	Percentage of mandatory fields filled, including accident time, location, vehicle type, and severity.	Total number of mandatory fields per record. Count missing fields for all records.
Record Completeness Score	Proportion of records where all mandatory fields are populated.	Records with all mandatory fields populated. Count total records.

Minimal Volume Requirement	Number of accident records required for reliable statistical analysis and AI training.	Define the minimum volume threshold for the dataset. Count the total number of records in the dataset.

Below are the formulas and thresholds used evaluate and calculate the metrics:

Table 32: formulas and thresholds used evaluate and calculate the metrics for completeness for accident datasets.

Metrics	Formula	Threshold
Field Completeness Score	1 – (Missing fields/Total mandatory fields) × 100	≥ 95%
	Rationale : Ensures all critical fields required for accident analysis are available for each record.	
Record Completeness Score	(Complete records/Total records) × 100	≥ 90%
	Rationale: Identifies records that are complete versus those with multiple missing fields.	
Minimal Volume Requirement	Check if: Total records ≥ Minimum required volume	≥ 10,000 for national datasets
	Rationale: Guarantees sufficient dataset size for statistical reliability and training models.	

Category: Accuracy

Scope: Ensuring accident data accurately reflects real-world events with minimal errors.

Table 33: Metrics definition and requirements for accuracy for accident datasets

Metrics	Definition	Requirements
Field Accuracy Score	Percentage of fields verified as correct by cross-referencing trusted sources (e.g., police reports, IoT data).	Verify fields against trusted references (e.g., police reports, IoT data). Count correct entries per field.
Aggregate Accuracy Score	Measures overall correctness by assessing records against established benchmarks or trusted references.	Complete records against trusted sources. accurate records and total records.

Benchmark Accuracy Score	Proportion of dataset entries matching benchmarks (e.g., historical averages, known distributions).	Benchmarks for critical distributions (e.g., average accident frequencies).
Outlier Rate	Proportion of records containing implausible values (e.g., speeds > 300 km/h or invalid timestamps).	Thresholds for implausible values (e.g., impossible speeds).

Below are the formulas and thresholds used evaluate and calculate the metrics:

Table 34: formulas and thresholds used evaluate and calculate the metrics for accuracy for accident datasets.

Metrics	Formula	Threshold
Field Accuracy Score	(Correct entries per field / Total entries) × 100	≥ 98% for critical fields
	Rationale: Critical fields like time, location, and severity must be error-free for reliable analysis**.**	
Aggregate Accuracy Score	(Accurate records / Total records) × 100	≥ 95%
	Rationale: Reflects cumulative correctness across all fields within each record.	
Benchmark Accuracy Score	(Matching records to benchmarks / Total records) × 100	≥ 85%
	Rationale: Verifies the dataset adheres to expected distributions, minimizing systematic errors.	
Outlier Rate	(Outlier records / Total records) × 100	≤ 2%
	Rationale: Detects anomalies that could indicate errors during data capture or entry.	

Category: Consistency

Scope: Ensuring accident data is logically and temporally coherent.

Table 35: Metrics definition and requirements for Consistency for accident datasets

Metrics	**Definition**	**Requirements**
Duplicate Record Rate	Percentage of duplicate records in the dataset.	Duplicate entries in the dataset.
Geographic Consistency	Percentage of records where the location aligns across related fields (e.g., GPS data vs. police-reported location).	Cross-check location fields (e.g., GPS vs. police reports). Count consistent and total records.

Data Format Consistency	Percentage of records conforming to standard data formats for fields (e.g., ISO 8601 for timestamps, lat/long for locations).	Standard formats for critical fields. Validate records against these standards.
Attribute Value Range Consistency	Percentage of records with field values falling within expected ranges (e.g., speeds < 200 km/h, valid dates).	Valid ranges for key fields (e.g., speed, time). count valid records.

Below are the formulas and thresholds used evaluate and calculate the metrics:

Table 36: formulas and thresholds used evaluate and calculate the metrics for Consistency for accident datasets.

Metrics	Formula	Threshold	Rationale
Duplicate Record Rate	(Duplicate records / Total records) × 100	≤ 1%	Avoids skewing results due to duplication.
Geographic Consistency	(Records with matching geographic data / total records) × 100	≥ 96%	Ensures accident location data is consistent across multiple fields or sources.
Data Format Consistency	(Records following standard formats / Total records) × 100	≥ 98%	Enforces standardized data formats, preventing errors during analysis or system integration.
Attribute Value Range Consistency	(Records within expected ranges/Total records) × 100	≥ 99%	Ensures all data fields have plausible and expected values, reducing anomalies.

Category: Timeliness

Scope: Ensuring accident data is current and relevant for analysis.

Table 37: formulas and thresholds used evaluate and calculate the metrics for Timeliness for accident datasets.

Metrics	Definition	Requirements
Age of data	Average age of accident	Determine the year of each accident. Compute the difference between the current year and accident year.

Below are the formulas and thresholds used evaluate and calculate the metrics:

Table 38: formulas and thresholds used evaluate and calculate the metrics for Timeliness for accident datasets.

Metrics	Formula	Threshold	Rationale
Age of data	Current year - Year of accident	< 5 years	Ensures the dataset focuses on recent and relevant accident data for analysis.

Category: Relevancy

Scope: Ensuring the dataset covers scenarios that are critical for accident analysis and modelling.

Table 39: Metrics definition and requirements for Relevancy for accident datasets

Metrics	Definition	Requirements
Scenario Relevance Score	Proportion of records representing high-priority accident scenarios (e.g., multi-vehicle crashes, adverse weather).	Define high-priority scenarios (e.g., adverse weather, multi-vehicle crashes). Identify records belonging to these scenarios.
Geographic Relevance Score	Percentage of records originating from the target geographic area (e.g., a specific city, region, or road network).	Define the target geographic area. Identify records belonging to the area.
Accident Type Diversity Score	Proportion of records covering a variety of accident types (e.g., single vehicle, rear-end collisions, rollovers).	Define accident type categories Classify records by accident type.
Environmental Context Relevance	Percentage of records capturing environmental factors (e.g., weather, road conditions, lighting).	Identify relevant environmental factors (e.g., weather, lighting). Count records containing these fields.

Below are the formulas and thresholds used evaluate and calculate the metrics.

Table 40: formulas and thresholds used evaluate and calculate the metrics for Relevancy for accident datasets.

Metrics	Formula	Threshold	Rationale

Scenario Relevance Score	(Records for high-priority scenarios / Total records) × 100	≥ 85%	Focuses on scenarios that are critical for safety analysis or AI model training (e.g., lane merging, emergency braking).
Geographic Relevance Score	(Records from target geography / Total records) × 100	≥ 95%	Ensures the dataset focuses on the geographic area under study, avoiding irrelevant entries.
Accident Type Diversity Score	(Records across accident types / Total records) × 100	≥ 85%	Ensures datasets provide a balanced representation of accident types critical for safety studies.
Environmental Context Relevance	(Records with contextual data / Total records) × 100	≥ 90%	Provides essential context for understanding accident causes and preventing similar events.

Category: Coverage

Scope: Ensuring the dataset covers target area population and the right accident type

Table 41: Metrics definition and requirements for Coverage for accident datasets

Metrics	Definition	Requirements
Target area coverage	Degree to which the dataset covers the geographical area.	Write here … What it needs for data processing., e.g., Record level. Attribute level.
Target population coverage	Degree to which the dataset covers the target population.	

Below are the formulas and thresholds used evaluate and calculate the metrics.

Table 42: formulas and thresholds used evaluate and calculate the metrics for coverage for accident datasets.

Metrics	Formula	Threshold	Rationale
Target area coverage	Target coverage = area of interest covered / Total target area x 100	No threshold	Defines coverage of database in relation to area.
Target population coverage	Population coverage = population of interest covered /Total target population x 100	No threshold	Defines coverage of database in relation to population.

2.5.2.5 Quality metrics and Threshold for Quality metrics and Threshold for Synthetic datasets

Completeness

Scope: Completeness of synthetic data is described in terms of having values for all expected attributes and related entity instances in a specific context of use. In other terms, it evaluates the proportion of missed values and their impact on the Scenario Source Data parameters for the creation of a scenario and the context of use.

Table 43: Metrics definition and requirements for Completeness for Synthetic datasets

Metrics	Definition	Requirements
Scenario Source Data model completeness	The proportion of available attributes to complete scenario source data model	Mandatory attributes for the scenario construction shall be determined to be able to determine if the scenario has complete information.
Raw values completeness	The proportion of filled values	Ensure that the content of the dataset is not truncated and has complete set of values.

Below are the formulas and thresholds used evaluate and calculate the metrics.

Table 44: formulas and thresholds used evaluate and calculate the metrics for Completeness for Synthetic datasets.

Metrics	Formula	Threshold	Rationale
Scenario Source Data model completeness	(number of corresponding attributes available in the raw data / number of attributes in the scenario source data model) x 100	100% for mandatory attributes	Raw data will be transformed into Scenario Source Data which has a data model designed to construct a scenario. Ensuring data for 5 layers of scenario description are included
Raw values completeness	(number of filled items / number of total items) x 100	>= 95%	Raw data completeness ensures that the scenarios created from the raw data will be complete.

Accuracy

Scope: Ensures that in a specific context the synthetic raw data is precise and correct the reality of the ground truth to a certain extent

Table 45: Metrics definition and requirements for Accuracy for Synthetic datasets

Metrics	Definition	Requirements
Precision of the positioning and dynamics values	Proportion of attributes available with the mandatory precision.	The mandatory precision for dataset attributes must be specified.
Sampling frequency	Sampling frequency of available data	Sampling frequency shall be available in the dataset and equal to 10Hz to be able to extract all actions and events.
Granularity of objects classification	Level of which the objects are defined and classed in a well structure hierarchy	Object representation shall be based on an ontology to be able to cover all aspects of scenarios detected from any source of dataset

Below are the formulas and thresholds used evaluate and calculate the metrics.

Table 46: formulas and thresholds used evaluate and calculate the metrics for Accuracy for Synthetic datasets.

Metrics	Formula	Threshold	Rationale
Precision of the positioning and dynamics values	(number of accurate attributes / total number of attributes) x 100	100%	The values used for scenario reconstruction must be precise enough to capture the scenery and dynamics of the scenario
Sampling frequency	sampling frequency	10Hz	Under a certain sampling frequency, behaviours of actors and the sequence of events can be missed.
Granularity of objects classification	(Number of object class in raw data structure / Number of object class in scenario source data structure) x 100	Depends on the object class	The description of the scenarios is based on an ontology for objects representation. We need to have a classification of objects that fits this ontology to ensure it is thin enough to get scenarios characteristic elements.

Consistency

Scope: Ensure that the quality of the synthetic data always behave and perform in similar way. It will ensure that the core elements are coherent among themselves using physical rules.

Table 47: Metrics definition and requirements for Consistency for Synthetic datasets

Metrics	Definition	Requirements
Spatial data coherence (image generation based)	Core elements coherence among the objects generated and its own attributes in the frame. This metric evaluate the spatial position of object based on its attributes given a referential position	Ensure that the objects in the dataset faithfully represent real data in order to create good quality scenarios
Temporal data coherence (video based generation)	core element temporal coherence in tracking approach. An object in a generated video shouldn't ' have a strange behaviour along the frame. We introduce the flow, a pixelwise distance between two consecutive frame of a generated video	Ensure that the objects in the dataset are not acting weird from one frame to another in order to create good quality scenarios

Below are the formulas and thresholds used evaluate and calculate the metrics.

Table 48: formulas and thresholds used evaluate and calculate the metrics for Consistency for Synthetic datasets.

Metrics	Formula	Threshold	Rationale
Spatial data coherence (image generation based)	Multiple detection model as judge: Cn (such as yolo) IoU : intersection over union Total object coherence score = IoU(Cn_bbox(generated images)), ground_truth_position)	The higher the better	Multiple detection model as judge is a known method to get rid of individual model bias and some involved detection using object's attributes description
Temporal data coherence (video-based generation)	flow = generated_video_frame(t + delta_t) - generated_video_frame(t)	depend of the video subject the lower the better	An object shouldn't disappear or be generated in random positions along the video

Timeliness

Scope: Ensuring the extraction of scenarios not to be short-cutted and have a full description of scenarios

Table 49: Metrics definition and requirements for Timeliness for Synthetic datasets

Metrics	Definition	Requirements
Scenario Timeliness	Proportion of uncomplete instances of scenarios	Scenario duration in the raw data shall be calculated and shall represent the whole duration of the scenario when extracted to ensure that the scenario exclude uncut recordings
Scenario timelapse	ensure parser: synthetic data -> scenario source data, respect event time occurrence	Event time occurrence shall be respected by verifying between synthetic data and its projection as scenario source data event occurrence time

Below are the formulas and thresholds used evaluate and calculate the metrics.

Table 50: formulas and thresholds used evaluate and calculate the metrics for Timeliness for Synthetic datasets.

Metrics	Formula	Threshold	Rationale
Scenario Timeliness	(Duration of scenarios extracted / total duration of scenarios detected) x 100	>= 95%	For each scenario, we need to have the data for the whole duration of the scenario for it not to be truncated.
Scenario timelapse	lapse =Σ(time_scenario_input_ data_attribut_event_occurence - time_synthetic_data_ attribut_event_occurence)	The lower the better	I shall be able to evaluate the timelapse between the original synthetic dataset and its projection in scenario input data format. (Working with simulated data, no real sensors are used)

Coverage

Scope: Ensuring how much information a dataset is providing by measuring the proportion of scenario source data (input data) core elements of a scenario retrievable from synthetic raw data.

Table 51: Metrics definition and requirements for Coverage for Synthetic datasets

Metrics	Definition	Requirements
Entity coverage	Determining how much entities a certain dataset is covering	Entity coverage shall be calculated to check how much the synthetic raw data is covering regarding the main actors of the scenario
Scenario coverage	Determining how much the scenarios extracted from a dataset is covering	Scenario coverage shall be calculated to check how much the synthetic raw data is covering regarding the distances covered in the scenarios
Infrastructure elements coverage	Determining what the dataset is providing regarding the infrastructure	Infrastructure elements coverage shall be calculated to check how much the synthetic raw data is covering regarding the infrastructure elements covered in the scenarios
Manoeuvre coverage	Determine what types of manoeuvres the dataset is providing	Manoeuvre coverage shall be calculated to check how much the synthetic raw data is covering regarding the behaviors covered in the scenarios

Below are the formulas and thresholds used evaluate and calculate the metrics.

Table 52: formulas and thresholds used evaluate and calculate the metrics for Coverage for Synthetic datasets.

Metrics	Formula	Threshold	Rationale
Entity coverage	(number of entities in the dataset / number of entities in Scenario source data model) x 100	the higher the better	Determine how much a dataset is providing to be able to cover most of the ODDs with scenarios
Scenario coverage	(total number of Km the scenario represent / total number of Km the dataset represents) x 100	the higher the better	Determine how much a dataset is providing to be able to cover most of the ODDs with scenarios

Infrastructure elements coverage	(number of infrastructure elements / total number of IE in the SSD model) x 100	the higher the better	Determine how much a dataset is providing to be able to cover most of the ODDs with scenarios
Manoeuvre coverage	(number of manoeuvre types / total number of IE in the SSD model) x 100	the higher the better	Determine how much a dataset is providing to be able to cover most of the ODDs with scenarios

Relevancy

Scope: Ensure that the synthetic dataset contains information relevant to create scenarios.

Table 53: Metrics definition and requirements for Relevancy for Synthetic datasets

Metrics	Definition	Requirements
Data scenario relevancy	Making sure that the defined core elements of a scenario exist in a dataset.	Define the scenario core elements to create relevant scenarios

Below are the formulas and thresholds used evaluate and calculate the metrics.

Table 54: formulas and thresholds used evaluate and calculate the metrics for Relevancy for Synthetic datasets.

Metrics	Formula	Threshold	Rationale
Data scenario relevancy	Can apply a priority coefficient to the important and less important attribute of a scenario	All of the core elements exist to create scenario	When working with synthetic datasets I shall be able to determine if the content of the raw data can generate relevant scenarios

2.5.2.6 Quality metrics and Thresholds for all datasets

Licensing

Scope *Identifying restrictions of use, and associated cost to process the dataset*

Table 55: Metrics definition and requirements for Licensing for all datasets

Metrics	Definition	Requirements

Readability	Does the access to the dataset required the use of mandatory tools?	Proper documentation
License	which license rule use of dataset	Proper documentation
Restriction: Access to partners	Ensure that dataset is accessible to all partners. If not, identify restrictions	Proper documentation
Restriction: Commercial Use	Is commercial use allowed or not	Proper documentation
Data exploitation Cost: Royalties	Estimation off cost, if any	Proper documentation
Data exploitation Cost: Men month	Estimation of the needed time to process the dataset	Proper documentation
Data exploitation Cost: Compliance cost	Estimation of the cost associated to the use of the dataset (I.e.: Stockage of data, resources)	Proper documentation

Below are the formulas and thresholds used evaluate and calculate the metrics.

Table 56: formulas and thresholds used evaluate and calculate the metrics for Licensing for all datasets.

Metrics	Formula	Threshold
Readability	YES/NO	NA
	Rationale: identifying restrictions of future use	
License	among List of possible License	NA
	Rationale: identifying restrictions of future use	
Restriction: Access to partners	List of restriction access (i.e to academics); List of specific partners with access.	At least one partner has access
	Rationale: Identifying which partners is allowed to use dataset	
Restriction: Commercial Use	YES/NO	NA

	Rationale: identifying restrictions of future use	
Data exploitation Cost: Royalties	estimation, in €	NA
	Rationale: Helps the decision-making process in determining whether to use this dataset or not.	
Data exploitation Cost: Men month	estimation, in €	NA
	Rationale: Helps the decision-making process in determining whether to use this dataset or not.	
Data exploitation Cost: Compliance cost	estimation, in €	NA
	Rationale: Helps the decision-making process in determining whether to use this dataset or not.	

2.5.2.7 Detailed Level Methodology - Next step

As outlined earlier, the detailed-level methodology is currently a work in progress and has not yet been finalized due to pending inputs from T2.3 and WP5. The next step in developing this methodology is to identify common metrics that can be applied across multiple datasets. Once the required inputs are received from the relevant tasks, the methodology will be updated accordingly. These updates may include refining metrics, providing clear definitions, and establishing appropriate thresholds to ensure consistency and applicability.

2.6 Summary of metrics

The metrics defined for high level methodology are Relevance, Coverage, Accessibility, Documentation, Metadata and Licensing.

Table 57: Summary of all metrics for detailed level methodology below provides the summary of all metrics defined for detailed-level methodology. Please note that this is not a final version but rather a working document, subject to change

Table 57: Summary of all metrics for detailed level methodology

Metrics	Category	Dataset Type
Temporal Completeness	Completeness	Fixed Point of View, Moving Point of View
Attribute Completeness		Fixed Point of View, Moving Point of View
Spatial Completeness		Fixed Point of View
Data Annotation Completeness		Fixed Point of View, Moving Point of View
Annotation Categorical Completeness		Fixed Point of View, Moving Point of View
Trajectory Completeness		Fixed Point of View
Field Completeness Score		Accident data
Record Completeness Score		Accident data

Minimal Volume Requirement		Accident data
Scenario Source Data model completeness		Synthetic datasets
Raw values completeness		Synthetic datasets
Temporal Accuracy	Accuracy	Fixed Point of View
Spatial Accuracy		Fixed Point of View
Sensor Measurement Accuracy		Fixed Point of View, Moving Point of View
Unit Accuracy		Fixed Point of View
Non-default Valued Attributes Accuracy		Fixed Point of View, Moving Point of View
Pixel Accuracy		Fixed Point of View
object classification accuracy		Moving Point of View
Sensor Measurement Accuracy		Moving Point of View
Range Accuracy		Moving Point of View
Coherent values		Moving Point of View
Precision of the positioning and dynamics values		Accident datasets
Sampling frequency		Accident datasets
Granularity of objects classification		Accident datasets
Precision of the positioning and dynamics values		Synthetic datasets
Sampling frequency		Synthetic datasets
Granularity of objects classification		Synthetic datasets
Temporal Consistency	Consistency	Fixed Point of View, Moving Point of View
Spatial Consistency		Fixed Point of View dataset
Sensor Consistency		Fixed Point of View, Moving Point of View
Annotation Consistency		Fixed Point of View, Moving Point of View
Format Consistency		Fixed Point of View dataset, Moving Point of View, Accident data
Object Resolution Consistency		Fixed Point of View dataset
Duplicate Record Rate		Accident datasets
Geographic Consistency		Accident datasets
Attribute Value Range Consistency		Accident datasets
Data Coherence		Accident datasets
Spatial data coherence (image generation based)		Synthetic datasets
Temporal data coherence (video-based generation)		Synthetic datasets
Data freshness	Timeliness	Fixed Point of View, Moving Point of View
Sensor EoL		Fixed Point of View
Age of data		Accident datasets

Scenario Timeliness		Synthetic datasets
Scenario timelapse		Synthetic datasets
Target Area Coverage	Coverage	Fixed Point of View
Temporal Relevancy		Fixed Point of View
Event identification		Fixed Point of View
Measurement coverage		Fixed Point of View
Relevant ODD conditions		Moving Point of View
Object coverage		Moving Point of View
Target area coverage		Accident datasets
Target population coverage		Accident datasets
Entity coverage		Synthetic datasets
Scenario coverage		Synthetic datasets
Infrastructure elements coverage		Synthetic datasets
Manoeuvre coverage		Synthetic datasets
Relevant Env Conditions	Relevancy	Fixed Point of View
Relevant scenarios		Fixed Point of View, Moving Point of View, Synthetic datasets, Accident datasets
Proportion of relevant data		Fixed Point of View
Geographic Relevance Score		Accident datasets
Accident Type Diversity Score		Accident datasets
Environmental Context Relevance		Accident datasets
Readability	Licensing	Fixed Point of View, Moving Point of View, Synthetic datasets, Accident datasets
Restriction: Access to partners		Fixed Point of View, Moving Point of View, Synthetic datasets, Accident datasets
Restriction: Commercial Use		Fixed Point of View, Moving Point of View, Synthetic datasets, Accident datasets
Data exploitation Cost: Royalties		Fixed Point of View, Moving Point of View, Synthetic datasets, Accident datasets
Data exploitation Cost: Men month		Fixed Point of View, Moving Point of View, Synthetic datasets, Accident datasets
Data exploitation Cost: Compliance cost		Fixed Point of View, Moving Point of View, Synthetic datasets, Accident datasets

3 CONCLUSIONS

This deliverable presents a comprehensive methodology for data quality assessment within the framework of the SYNERGIES project, addressing key challenges in Connected, Cooperative, and Automated Mobility (CCAM) systems. By developing robust methodologies and metrics, this work supports the creation of high-quality scenarios essential for advancing CCAM systems.

The proposed two-tiered approach comprising the High-Level Qualification Methodology and the Detailed Level Qualification Methodology. The approach demonstrates a systematic and scalable solution for evaluating datasets. The High-Level Qualification Methodology allows for rapid screening of datasets based on essential criteria, ensuring efficient resource allocation. Meanwhile, the Detailed Level Qualification Methodology provides an in-depth analysis of dataset quality dimensions such as accuracy, completeness, and timeliness, fostering alignment with CCAM use cases. is important to note that the Detailed Level Qualification Methodology is a working version and still under development. As the SYNERGIES project progresses, it will be refined to remain adaptable to evolving project needs

For the evaluation, 14 datasets were selected to assess their quality. These 14 datasets were chosen to evaluate the effectiveness of the High-Level Qualification Methodology. This methodology was employed to test how efficiently it can screen datasets based on essential criteria, ensuring proper resource allocation while focusing on key aspects of dataset quality. The evaluation of the 14 datasets, which resulted in a success rate of 78.5%, demonstrated the utility of the High-Level Qualification Methodology. However, it also revealed several common challenges that led to some datasets failing the evaluation. These challenges included missing critical information, such as lane markings, traffic signs, or road infrastructure details; insufficient data volume, with limited scene coverage or low recording frequency; and accessibility issues, such as broken links or unresponsive dataset owners. These findings highlight the need for improvement in areas such as metadata completeness, data volume, and dataset accessibility.

The next steps following this deliverable will focus on the continued development and refinement of the detailed level methodology outlined. These steps will include incorporating feedback from various work packages, including WP4, WP5, and WP6, as well as collaborating with Task 3.2, Task 3.3, and Task 3.4 to deliver the necessary training for utilizing the methodology.

This deliverable lays a strong foundation for the SYNERGIES project by enabling the integration of diverse data sources into interoperable scenario pipelines. Moreover, this work directly supports several key objectives of the SYNERGIES project. First, the developed methodologies will contribute to the creation of a unified European Scenario Dataspace through ensuring high-quality, and standardized datasets. Second, by improving the quality and reliability of scenario datasets, this work supports the robust validation of CCAM systems, enhancing safety and regulatory compliance and thus, advancing safety assurance and validatio

4 REFERENCES

[1] S. K. Pradhan, H.-M. Heyn and E. Knauss, "Identifying and managing data quality requirements: a design science study in the field of automated driving," *Software Quality Journal,* 2023.

[2] L. L. Pipinol, Y. W. Lee and R. Y. Wang, "Data Quality Assessment," *Communications of the ACM,* pp. 211 - 218, 2002.

[3] J. Wang, Y. Liu, P. Li, S. Aggarwal and Z. Lin, "Overview of Data Quality: Examining the Dimensions, Antecedents, and Impacts of Data Quality," *Journal of the Knowledge Economy,* 2023.

[4] J. Declerck, D. Kalra, R. V. Stichele and P. Coorevits, "Frameworks, Dimensions, Definitions of Aspects, and Assessment Methods for the Appraisal of Quality of Health Data for Secondary Use: Comprehensive Overview of Reviews," *National Library of Medicine,* 2024.

[5] L. Sahi, R. Laborde, M.-A. Kandi, M. Sibilla, G. Macilotti, B. Abdelmalek and A. Ferreira, "Towards Reliable Collaborative Data Processing Ecosystems: Survey on Data Quality Criteria," *IEEE 22nd International Conference on Trust, Security and Privacy in Computing and Communications,* 2023.

[6] B. S. A. E. H. A. E. Firas Bayram, "DQSOps: Data Quality Scoring Operations Framework for Data-Driven Applications," *EASE '23: Proceedings of the 27th International Conference on Evaluation and Assessment in Software Engineering,* 2023.

[7] F. Mohammed, A. S. M. Kayes, E. Pardede and W. Rahayu, "A Framework for Measuring IoT Data Quality Based on Freshness Metrics," *IEEE,* 2020.

[8] I. Viedt, J. Mädler, V. Khaydarov and L. Urbas, "Prescriptive and descriptive quality metrics for the quality assessment of operational data," *Gesellschaft für Informatik, Bonn,* 2022.

[9] L. C. Günther, E. Colangelo, H.-H. Wiendahl and C. Bauer, "Data quality assessment for improved decision-making: a methodology for small and medium-sized enterprises," *Procedia Manufacturing,* 2019.

[10] V. Simard, M. Rönnqvist, L. Lebel and N. Lehoux, "A Method to Classify Data Quality for Decision Making Under Uncertainty," *ACM Journal of Data and Information Quality,* 2023.

[11] N. West, J. Gries, C. Brockmeier, J. C. Göbel and J. Deuse, "Towards integrated Data Analysis Quality: Criteria for the application of Industrial Data Science," *IEEE,* 2021.

[12] P. N. S. S. D. P. H. N. K. J. P. A. S. A. A. A. Z. Arda Goknil, "A Systematic Review of Data Quality in CPS and IoT for Industry 4.0," *ACM Computer Survey,* 2023.

[13] F. s. action, "CCAM Data Sharing," Connected Automated Driving, 2024.

[14] N. D. Lawrence, "Data Readiness Levels," 2017.

[15] G. Bagschik, T. Menzel and M. Maurer, "Ontology based Scene Creation for the Development of Automated Vehicles," *IEEE,* 2018.

[16] V. Simard, M. Rönnqvist, L. Lebel and N. Lehoux, "A Method to Classify Data Quality for Decision Making Under Uncertainty," *ACM Journal of Data and Information Quality,* 2023.

[17] E. Y. J. F. X. Z. W. Z. Y. C. B. L. Z. A. C. K. Mingyu Liu, "A Survey on Autonomous Driving Datasets: Statistics, Annotation Quality, and a Future Outlook," *IEEE,* 2024.

[18] C.-H. H. a. H. Y. Chih-Hong Cheng, "Quantitative Projection Coverage for Testing ML-enabled Autonomous Systems," *Automated Technology for Verification and Analysis - 16th International Symposium,,* 2018.

[19] S. Bhandari, N. Ranjan, Y.-C. Kim, J.-D. Park, K.-I. Hwang, W.-H. Kim and Y.-S. Hong, "An Automatic Data Completeness Check Framework for Open Government Data," *Open Access Journal,* 2021.

A. ABBREVIATIONS AND DEFINITIONS

Term	Definition
ADAS	Advanced Driver Assistance Systems
AIMSUN	Advanced Interactive Microscopic Simulator for Urban and Non-Urban Networks
API	Application Programming Interface
CARLA	An open-source simulator for autonomous driving research
CCAM	Connected, Cooperative, and Automated Mobility
CRISP-DM	Cross Industry Standard Process for Data Mining
CSV	Comma-Separated Values
DAQ	Data Analysis Quality
DCR	Data Completeness Ratio
DQ	Data Quality
DRL	Data Readiness Level
FPS	Frames Per Second
FTP	File Transfer Protocol
H.265 MPEG	High-Efficiency Video Coding (video compression standard)
IDS	Industrial Data Science
IoT	Internet of Things
JSON	JavaScript Object Notation
ML	Machine Learning
MLOps	Machine Learning Operations
ODD	Operational Design Domain
REST	Representational State Transfer
SDK	Software Development Kit
SUMO	Simulation of Urban MObility (an open-source traffic simulation tool)
UAV	Unmanned Aerial Vehicle