Test and Evaluation Methodology for Indoor Localization Systems in Intralogistics

Dissertation (monograph) approved
by the Doctoral Degree Committee of
Hamburg University of Technology

in pursuit of the academic degree of

Doktor-Ingenieur (Dr.-Ing.)

written by

Jakob Schyga

from

Bad Oldesloe

2026

First Reviewer:	Prof. Dr.-Ing. Carlos Jahn
Second Reviewer:	Prof. Dr.-Ing. Robert Schulz

Date of oral examination: 06.10.2025

DOI: 10.15480/882.16728
ORCID: 0000-0002-1687-7654

Abstract

The research, development, and application of Indoor Localization Systems (ILSs) are hindered by the absence of adequate methods for Test and Evaluation (T&E). This dissertation tackles this challenge within the intralogistics context. It introduces methods for systematically designing test scenarios, for benchmarking, and for specifying location data requirements, which are integrated into a comprehensive T&E methodology. The contributions of this work enable stakeholders to meaningfully evaluate the performance of ILSs and identify suitable systems for intralogistics applications, ultimately fueling the ongoing transition to a more interconnected and efficient industry.

Kurzfassung

Die Forschung, Entwicklung und Anwendung von Indoor-Lokalisierungssystemen (ILSs) wird durch das Fehlen adäquater Methoden für Test- und Evaluations (T&E) behindert. Diese Dissertation widmet sich dieser Herausforderung im intralogistischen Kontext. Es werden Methoden zur systematischen Gestaltung von Testszenarien, zum Benchmarking, sowie zur Spezifikation von Anforderungen an Lokalisierungsdaten entwickelt und in eine umfassende T&E-Methodik integriert. Die Ergebnisse dieser Arbeit ermöglichen es Stakeholdern die Leistungsfähigkeit von ILSs zu beurteilen und geeignete Systeme für Anwendungen in der Intralogistik zu identifizieren. Dies trägt letztlich dazu bei, den Übergang zu einer vernetzten und effizienten Industrie voranzutreiben.

Acknowledgements

Throughout my academic journey leading to the completion of this dissertation, I have been fortunate to receive support from numerous individuals and organizations. I am thankful for their contributions and would like to express my sincere gratitude.

First, I sincerely thank my first and second examiners, Professor Carlos Jahn and Professor Robert Schulz, for their willingness to evaluate my doctoral thesis. Additionally, I am grateful to everyone who accompanied me during my time at the *Institute for Technical Logistics* (ITL) of *Hamburg University of Technology* (TUHH). In particular, I wish to express my gratitude to the late Professor Jochen Kreutzfeldt and to Dr. Johannes Hinckeldeyn for the stimulating discussions and their support for my work. Besides, their leadership was instrumental in establishing an inspiring and enjoyable work environment, which continues to thrive. In this context, I would also like to thank Philipp Braun, Hendrik Rose, Markus Knitt, Yevgen Blank, Cordula Hiller-Martens, and all my other colleagues at the ITL for their mutual support and camaraderie.

Beyond my direct colleagues at ITL, I would like to thank Swantje Plambeck from the *Institute of Embedded Systems*, as well as Christoph Eckert from *SICK AG*, and Benjamin Bruss from *Jungheinrich AG* for the collaborative engagement in the *3D_Log* project, which sparked the initial idea of this dissertation. Additionally, I am grateful for the insightful feedback on ideas and results from the members of the associated project steering committee. I would also like to thank TUHH for supporting this project as part of the I^3-program, as well as to the *Federal Ministry of Education and Research* for funding the subsequent research project *AGenC*. Furthermore, my research was facilitated by the generous scholarship provided by the *Dr. Friedrich Jungheinrich Foundation* during my tenure at ITL. Special thanks to Julia Thiel for the numerous enriching conversations and events that have contributed to my personal and professional development.

Regarding my private life, I am deeply grateful to all my friends and relatives for their encouragement and the time we have shared. Special thanks go to my girlfriend, Matilde Coin, my housemates from Veddel, and the extended circle of friends from Hamburg, Göttingen, Aveiro, and Nice. Finally, heartfelt thanks to my brother Paul, my mother Anette, and in loving memory of my late father Martin, for their companionship and support.

Preface

In the realm of engineering sciences, academic researchers tend to approach their work with a technology-driven mindset rather than a focus on practical applications and stakeholder demands. As a consequence, the requirements and limitations of the real world are often overlooked. Although a technology-driven approach is undoubtedly essential for advancing technology, it can sometimes create a disconnect between user needs and the capabilities of the systems being developed. From an application standpoint, the ideal system meets the users' specific needs, rather than merely achieving the highest performance regarding the dominant performance metric. This perspective underpins the present work, which seeks to develop an application-driven methodology for T&E of ILSs in the context of intralogistics.

Various publications have paved the way for this dissertation. The publications emerged from the engagement in multiple research projects during the author's employment at the *Institute for Technical Logistics of Hamburg University of Technology* (ITL). In particular, these include (a) the participation in a warehouse planning process for a large-equipment manufacturer, (b) the exploration of indoor localization technologies and their application to optimize intralogistics operations, as well as (c) the research on automated model generation for Cyber-Physical Systems (CPSs). A substantial portion of the dissertation's content has already been disseminated, reviewed, and discussed within the scientific community. Naturally, any reused or adapted content will be acknowledged and properly cited during this work.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AGV Automated Guided Vehicle

AMR Autonomous Mobile Robot

AoA Angle of Arrival

AR Augmented Reality

AuC Application under Consideration

ANOVA Analysis of Variance

CDF Comulative Distribution Function

CIR Channel Impulse Response

CoO Cell of Origin

CPS Cyber-Physical System

CPSL *Conference on Production Systems and Logistics*

CRUD Create, Read, Update, Delete

DoF Degree of Freedom

DSR Design Science Research

EKF Extended Kalman Filter

ELT Entity to be Localized / Tracked

EvAAL *Evaluating Ambient Assisted Living systems through competitive benchmarking*

EVARILOS *Evaluation of RF-based Indoor Localization Solutions for the Future Internet*

FoV Field of View

GNSS Global Navigation Satellite System

GPS Global Positioning System

GT Ground Truth

ICP Iterative Closest Point

IIoT Industrial Internet of Things

ILS Indoor Localization System

IMU Inertial Measurement Unit

IPIN *International Conference on Indoor Positioning and Indoor Navigation*

IPSN *International Conference on Information Processing in Sensor Networks*

IR Infrared

ITL *Institute for Technical Logistics of Hamburg University of Technology*

LiDAR Light Detection and Ranging

LLS *SICK LiDAR-LOC*

LOCU *SICK LOCU Localization Solution*

LoRaWAN *Long Range Wide Area Network*

LoS Line of Sight

NIST *U.S. National Institute of Standards and Technology*

NLoS None Line of Sight

MoCap Motion Capture

ms3 *SICK microScan3*

PoA Phase of Arrival

PTP Precision Time Protocol

RF Radio-Frequency

RFID Radio-Frequency Identification

RGB Red-Green-Blue

RGBD Red-Green-Blue-Depth

ROS Robot Operating Systems

RQ1 Research Question 1

RQ2 Research Question 2

RQ3 Research Question 3

RMSE Root Mean Squared Error

RSS Received Signal Strength

RTLS Real-Time Locating System

RTT Round Trip Time

SLAM Simultaneous Localization and Mapping

SuT System under Test

TDoA Time Difference of Arrival

ToA Time of Arrival

ToF Time of Flight

T&E Test and Evaluation

TUHH *Hamburg University of Technology*

UWB Ultra-Wideband

WGTL *Wissenschaftliche Gesellschaft für Technische Logistik*

WLAN Wireless Local Area Network

WMS Warehouse Management System

List of Symbols

A	Availability
A_{global}	Transformation matrix for global alignment
A_{local}	Transformation matrix for local alignment
E_{align}	Alignment data
E_{lat}	Latency data
E_{poses}	Evaluation poses
E_{GT}	GT experiment data
E_{SuT}	SuT experiment data
$E_{SuT,GT}$	Timestamp-aligned experiment data
I	Interest Space
M	Motion Space
O	Coordinate frame
O_{int}	Entity's Interest Frame
O_{loc}	Entity's Localization Frame
$O_{GT-global}$	Global GT frame
$O_{GT-local}$	Local GT frame
$O_{SuT-global}$	Global SuT frame
$O_{SuT-local}$	Local SuT frame
R_x^y	Rotation matrix from coordinate frame O_x to O_y
S	Safety Margin
T_x^y	Transformation matrix from coordinate frame O_x to O_y
b	Position tolerance
f_{update}	Update rate
g	Grid size

i, j, k, p	Counter referring to SuT measurement, GT measurement, evaluation pose, repetition
l, m, n, r	Total number of evaluation poses, GT measurements, SuT measurements, repetitions
$\vec{t}_x^y$	Translation vector from coordinate frame O_x to O_y
t_{del}	Time delay
$t_{err,sync}$	Synchronization error
t_{gap}	Time gap
t_{lat}	System latency
t_{net}	Network transmission time
v_{int}	Velocity in Interest Frame
$v_{max,int}$	Maximum velocity in Interest Frame
$\overline{E}_{SuT,GT}$	Timestamp-aligned experiment data associated with evaluation poses
$\widehat{E}_{SuT,GT}$	Fully aligned experiment data
$\widehat{\overline{E}}_{SuT,GT}$	Fully aligned experiment data associated with evaluation poses
$\widehat{\overline{\epsilon}}$	Localization error data
$\widehat{\overline{\epsilon}}_{rep}$	Repetition error data
$\lvert\widehat{\overline{\epsilon}}\rvert$	Absolute localization error data
$\lvert\widehat{\overline{\epsilon}}_{rep}\rvert$	Absolute repetition error data
$\widehat{\overline{\epsilon}}_h$	Horizontal position error data
$\vec{\epsilon}_{del}$	Time Delay Error
$\vec{\epsilon}_{eff}$	Effective Localization Error
$\vec{\epsilon}_{gap}$	Time Gap Error
$\vec{\epsilon}_{rel}$	Relevant Localization Error
$\vec{\epsilon}_{A,Acc/Rep}$	Percentile of the localization accuracy or repeatability corresponding to availability
$\vec{R}$	Requirement Margin vector
ϕ, θ, ψ	Roll, pitch, yaw

Glossary

Absolute Localization Process of estimating the location of an entity within a global coordinate system that is established by fixed landmarks

Action Within the*T&E 4Log Framework*, refers to a specific operation to contribute to the completion of a procedure within the T&E process

Application under Consideration Within the *T&E 4Log Framework*, refers to an application that is considered within the T&E process

Application-driven Influencing Factor Within the *T&E 4Log Framework*, refers to an attribute related to an AuC that influences the performance of an ILS

Availability Proportion of time during which location data is provided under the given requirements

Benchmarking Process of experimentation and evaluation of ILSs under predefined conditions

Black-box Testing Type of testing where a system's performance is assessed by examining its inputs and outputs, without significantly considering its inner working

Building-wide Testing Type of testing that is performed in a large-scale, application environment

Comparability In the context of T&E, the ability to compare results obtained from various experiments

Component-level Testing Type of testing that focuses on evaluating individual components of an ILS

Comprehensibility In the context of T&E, the degree to which information is presented in a clear, organized, and understandable manner

Concept Within the *T&E 4Log Framework*, refers to an abstract idea that can be employed by methods and procedures

Data Output In the context of user requirements for ILSs, describes the type of location information required for an application

Effective Localization Error Within the *T&E 4Log Framework*, refers to the total localization error relevant for the localization function

Entity to be Localized / Tracked Any entity, such as a person, object, vehicle, or robot, whose location is to be determined

Entity's Interest Frame Within the *T&E 4Log Framework*, refers to the local coordinate frame of the ELT relevant for the localization function

Entity's Localization Frame Within the *T&E 4Log Framework*, refers to the local coordinate frame of the ELT for which a location is provided by the ILS

Evaluation Pose Within the *T&E 4Log Framework*, refers to the location determining the relevant experiment data point for the determination of localization errors

Experiment Actual execution of a test scenario

Experiment Specification Concrete description of an experiment for a particular testbed

Feasibility In the context of T&E, refers to the practicality of conducting the T&E process, considering the required resources, time, and technical limitations

Ground Truth Assumed true location of an entity

Guideline Within the *T&E 4Log Framework*, refers to a rule, principle, or piece of advice that provides directional assistance within the T&E process

Indoor Localization System System designed to estimate an entity's location within an indoor environment

Interest Space Within the *T&E 4Log Framework*, refers to a multi-dimensional space in which the presence or absence of an ELT is of interest to an AuC

Laboratory Testing Type of testing that is performed within controlled conditions

Localization Process of estimating an entity's location within a reference space

Localization Accuracy Degree of agreement between true and estimated location of an entity within the same reference coordinate system

Localization Function Within the *T&E 4Log Framework*, refers to an operation within an application for determining the presence or absence of an entity in a specified interest space

Localization Method Approach by which the measured physical quantity is used to determine the location of an entity

Localization Repeatability Degree of agreement between location estimates at the same true location

Localization Technology Combination of measurement techniques, localization methods, and sensor technologies employed by ILS for localization

Locating Process of determining the location of a remote entity

Location Discrete value pair that provides information about an entity's coordinates within a reference space, typically associated with a specific moment in time

Location Data Series of locations within a certain time frame

Location Data Requirements User requirements referring to the location data output of an ILS

Meaningfulness In the context of T&E, describes the significance of results to the involved stakeholders

Measurement Technique Principle used to determine a physical quantity for localization

Method Within the *T&E 4Log Framework*, refers to a systematic approach employed to carry out specific actions within a procedure

Motion Space Within the *T&E 4Log Framework*, refers to a multi-dimensional space in which an ELT should be able to move without false association with an interest space

Partially Controlled Test Environment Testing environment in which some but not all test conditions can be manipulated, bridging the gap between laboratories and real-world environments

Performance Metric An indicator regarding an aspect of system performance

Pose Combined position and orientation of an entity

Positioning Process of self-localization, where an entity determines its own location using a localization system

Procedure Within the *T&E 4Log Framework*, refers to a sequence of actions or instructions designed to accomplish a specific outcome

Real-Time Locating System Subcategory of ILSs used for real-time remote locating, usually based on RF-signals

Relative Localization Process of determining a location of an entity within a local coordinate frame that is situated within a global frame of reference

Relevant Localization Error Within the *T&E 4Log Framework*, refers to the component of the Effective Localization Error resulting from localization accuracy or localization repeatability

Repeatable Testing Type of testing with the ability to conduct identical experiments multiple times to ensure consistent outcomes

Replicability In the context of T&E, the ability to achieve consistent results in different test facilities

Reproducibility In the context of T&E, the ability that independent researchers can obtain consistent results from experiments

Requirement Margin Within the *T&E 4Log Framework*, refers to the acceptable threshold of Effective Localization Error to ensure reliable localization functions

Safety Margin Within the *T&E 4Log Framework*, refers to a multi-dimensional space additionally considered within the requirement margin

Sensor Technology Underlying methods and principles used by the localization sensor to determine the measurand

Stakeholder In the context of T&E, refers to individuals, groups, or organizations that participate in the process of T&E or have an interest in the results

System Latency Time delay between the actual measurement and the provision of location data by an ILS

System under Test Specific localization system to be examined through empirical experimentation

System-level Testing Type of testing that evaluates the performance of an ILS as a holistic unit

Technical Influencing Factor Within the *T&E 4Log Framework*, refers to a characteristic of technical nature that influences the performance of an ILS

Test and Evaluation Empirical examination of ILSs based on experiments to draw conclusions about a system's performance or its suitability for a specific task

Test Scenario Specific conditions under which an ILS is tested

Testbed Within the *T&E 4Log Framework*, refers to a partially controlled test environment that includes the test volume, test environment, and GT, providing the setting for conducting experiments

Time Delay Error Within the *T&E 4Log Framework*, refers to the component of the Effective Localization Error resulting from the time delay of the location data

Time Gap Error Within the *T&E 4Log Framework*, refers to the component of the Effective Localization Error resulting from the time gap of the location data

Tool Within the *T&E 4Log Framework*, refers to a practical instrument facilitating the T&E process

Transferability In the context of T&E, the ability of T&E results to apply to real-world application scenarios

Update Rate Rate at which an ILS provides new location data, also referred to as update frequency

User Requirements Specifications that an ILS must meet to be suitable for a particular task or application

White-box Testing Type of testing that requires a comprehensive understanding of an ILS's inner workings, significantly influencing the testing procedure

1 Introduction

The ongoing shift toward an interconnected, efficient, and flexible industry is fueled by the advancement of Cyber-Physical Systems (CPSs), such as Indoor Localization Systems (ILSs) [1]. Although localization technologies are already significantly impacting people's daily lives, they have just begun to disrupt the industry, particularly in indoor environments [2]. The possible applications are especially manifold in intralogistics, where everything ultimately revolves around the transport of personnel, goods, industrial trucks, and other assets. Examples discussed in the scientific literature range from visualization of material flows [3, 4] to automation of booking processes [5] and industrial truck operations [6, 7]. Consequently, in intralogistics ILSs represent a vital instrument to remain competitive amidst rising customer demands [8].

Various indoor localization technologies have evolved towards or have already reached market maturity [9]. However, to fully exploit the potential of these technologies and enable widespread adoption in intralogistics and other domains, several challenges must still be overcome. Key challenges involve Test and Evaluation (T&E), which refers to the experimental examination of ILSs to draw conclusions about the performance of a system or its suitability for a certain task [10, 11]. Figure 1.1 illustrates the fundamental principle for selecting a suitable system for an application with T&E. An ILS is considered suitable for an application if the performance of the system satisfies user requirements, whereby T&E is used to determine the performance of the system and potentially identify and match requirements. T&E is crucial for a system user or integrator to facilitate an informed system selection, as well as for researchers and developers to empirically validate concepts, identify relevant challenges, and consequently design appropriate products [12, p. 15]. If suboptimal decisions are made, substantial opportunity costs can arise.

This dissertation seeks to address significant challenges related to T&E in the context of intralogistics. The main practical motivation behind this research is to facilitate a reliable comparison of system performance and assist in making informed decisions regarding system selection. This will contribute significantly to increasing stakeholder confidence while improving research and system development, thus stimulating the adoption of ILSs and their myriad applications in intralogistics [11, p. 209].

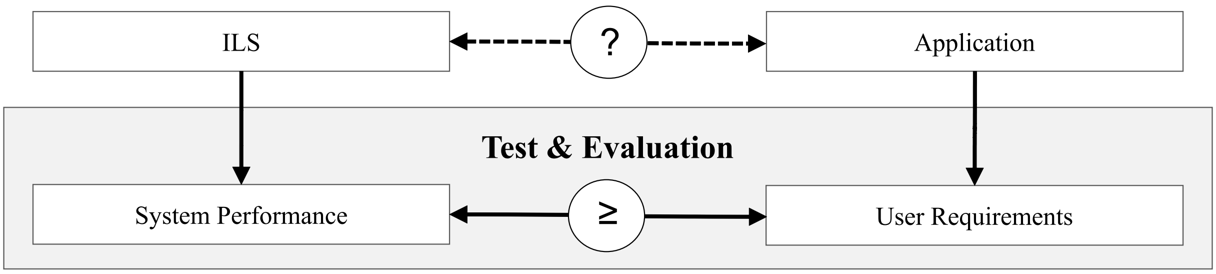


Figure 1.1: Finding a suitable ILS for an application through T&E

1.1 Problem Description and Research Objective

When designing a T&E procedure, many questions arise, such as how to deploy an ILS, set up a test environment, execute an experiment, which performance metrics to determine, or how to determine them. The answers to these questions significantly impact the T&E results, which according to Lymberopoulos *et al.* [13] must be realistic and comparable. For stakeholders to interpret and compare system performance and suitability effectively, it is imperative to apply common concepts, methods, and metrics [14, 15]. Methodologies combine these elements into an overarching strategy and are therefore crucial to establish a consensus on how to conduct T&E [16, p. 551]. Adopting such a methodology could yield numerous benefits, including the following.

- Providing system users with guidance to assess and ultimately select the most suitable system for their specific applications [17]
- Assisting system developers to identify optimization potentials and foster a deeper understanding of system behavior within realistic scenarios [14]
- Empowering system providers to state realistic and relevant performance metrics to their customers [17]
- Increasing confidence of stakeholders in tested systems [11, p. 209]
- Reducing the need for expertise on how to conduct T&E [18, p. 1]
- Facilitating the access, development, and understanding of compliant software tools [18, 19]
- Establishing a standard, potentially serving as a basis for jurisdiction [17, p. 1]

Due to their manifold potential benefits, methodologies, including the *Evaluating Ambient Assisted Living systems through competitive benchmarking* (EvAAL) *Framework* [20], the *Evaluation of RF-based Indoor Localization Solutions for the Future Internet* (EVARILOS) *Benchmarking Handbook* [21], and the *ISO/IEC 18305:2016 International Standard* [17] have been proposed. However, they have been applied primarily in the context of associated indoor localization competitions [13, 22–24] and have not been widely adopted in practice, particularly in the domain of intralogistics.

One of the reasons for the limited utilization of existing methodologies is their focus on building-wide testing, which requires significant resources and lacks transferability to other environments [18, p. 1]. As a cost-effective alternative, tests are often performed in test halls that can be adjusted, for example, by placing specific objects within the test area [25–28]. In this work, this approach is referred to as testing in partially controlled test environments, emphasizing that some, but not all, influencing factors can be controlled. T&E in such settings potentially provides results with higher repeatability, replicability, and comparability. However, the need for methodological approaches that produce outcomes reflecting the system's behavior in real-world application scenarios is even more evident. Since the basic structure of test halls commonly resembles compartments in warehouses and production halls, T&E in partially controlled test environments holds particularly great potential for intralogistics. However, so far, no adequate methodology exists.

Furthermore, existing methodologies overlook specific requirements and characteristics of the various application domains. Designing a universally applicable methodology that yields comparable and realistic results in such varied scenarios is a significant challenge [11, p. 209]. The *EVARILOS Benchmarking Handbook* addresses this issue by focusing on Radio-Frequency (RF)-based solutions [21]. However, from a user perspective, system performance takes precedence over functionality. Thus, focusing on a particular application domain is a more appropriate approach. Especially in the case of intralogistics, the specific characteristics diverge significantly from those of conventional tracking applications, necessitating a tailored methodology. Nonetheless, no existing methodology centers on intralogistics or related domains.

The research problem addressed in this work is summarized as follows. Methodologies are essential to establish common concepts for T&E and to achieve significant results. Especially for intralogistics, T&E in partially controlled test environments is a promising approach. However, existing methodologies focus on building-wide testing. In addition, they do not adequately address the particularities of the intralogistics domain. Therefore, the preliminary research objective of this dissertation is the development of a novel T&E methodology for partially controlled test environments, focusing on the application of ILSs in the context of intralogistics. The research objective is considered preliminary, since the examination of the state of the art in the following chapter will reveal gaps in the existing literature, ultimately leading to the identification of research questions and delineation of this preliminary research objective.

1.2 Research Design

The practical nature of the research objective requires an applied research design [30, p. 3]. This dissertation adopts the Design Science Research (DSR) paradigm according to Hevner *et al.* [31], which emphasizes the creation of novel artifacts to produce meaningful insights about their design. The artifact of this dissertation is the T&E methodology to be developed. Figure 1.2 illustrates the three-cycle view proposed by Hevner [29], in the context of this research. Accordingly, the T&E methodology is designed during the so-called design cycle by justifying various design decisions based on stakeholder demands within the intralogistics domain, as well as existing knowledge related to indoor localization and T&E (grounding). The rigor cycle provides the knowledge base for the design cycle and is furthermore used to identify theoretical contributions as additions to the existing knowledge base. The relevance cycle aims to ensure the

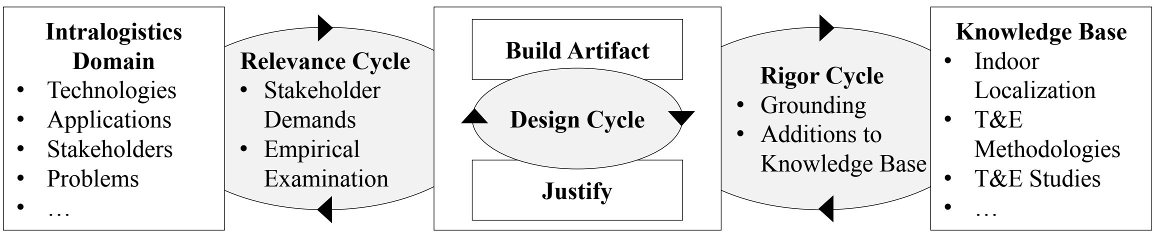


Figure 1.2: Three-cycle view of DSR in the context of this work (in accordance with Hevner [29, p. 2])

practical relevance of the artifact and the newly generated knowledge. This requires an analysis of the application domain to identify relevant research topics and stakeholder demands. To evaluate the fulfillment of stakeholder demands, an empirical examination must be performed.

Additionally, the research design of this study can be partially characterized by the concept of participatory research. As defined by Vaughn *et al.* [32], participatory research involves conducting a systematic investigation in direct collaboration with those affected by the subject under investigation. The foundations of this work were built within a two-year collaborative research project that explored various aspects of indoor localization for intralogistics applications, with T&E as a key topic. Two industrial partners were involved in the project, which can be described as system developer and system integrator of ILSs. The regular presentation and discussion of interim results within the project team and in a project steering committee naturally influenced the course of this research. this influence of market players is essential for the development of a research artifact and the generation of relevant knowledge for the application domain. Concurrently, this approach allowed ongoing research to have a direct impact on practice.

1.3 Structure of Dissertation

The structure of the present work follows a typical design for a thesis following the paradigm of DSR. Figure 1.3 provides an overview of the structure with the corresponding research results of each chapter and their dependencies. The elements corresponding to the three-cycle view of DSR are partially reflected in the results of the individual chapters. In the following, the content of each chapter is briefly introduced.

So far, the subject has been introduced. After motivating the topic of this dissertation, the research problem has been narrowed down and a preliminary research objective has been defined. In addition, the concepts of DSR and participatory research have been introduced in the context of this work to provide the research's theoretical basis.

The knowledge base for this work is provided in Chapter 2. The fundamental terms, indoor localization technologies, intralogistics applications, and user requirements of indoor localization are presented in Section 2.1. Next, Section 2.2 presents the state of the art of T&E by providing a taxonomy of T&E approaches, limitations and challenges, common practices, and a discussion of existing methodologies. This leads to the identification of the research questions, as well as the delineation of the research objective in Section 2.3.

Chapter 3 deals with the design of the T&E methodology, referred to as the *T&E 4Log Framework*. For this purpose, concrete design guidelines are initially established (Section 3.1). Next, the perspective of stakeholders is analyzed and their demands are defined in Section 3.2. Then the design of the *T&E 4Log Framework* is presented in Section 3.3. Following a clarification of the framework's scope and an overview of its architecture, the framework's components are described sequentially.

The *T&E 4Log Framework* is then used to collect empirical data for subsequent discussion. This empirical examination, presented in Chapter 4, includes three studies based on experiments conducted within the facility of the ITL of TUHH. Firstly, the *T&E 4Log Framework* is employed for an exemplary case study to examine the performance and suitability of two

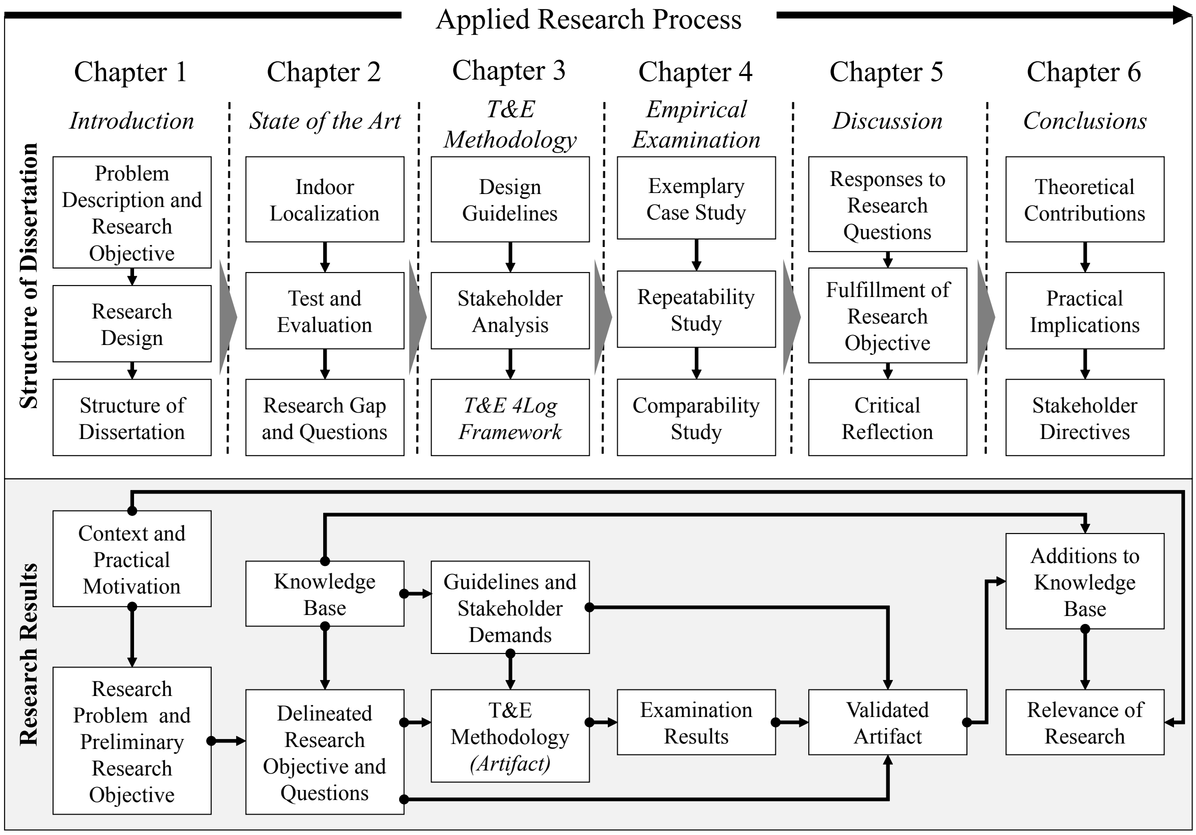


Figure 1.3: Structure of this dissertation with respective research results

distinct ILSs for localizing mobile robots. The other two studies are dedicated to analyzing key stakeholder demands. One explores the repeatability of experiments, while the other assesses the comparability of results under varying experiment conditions.

Based on the outcomes of the empirical examination, Chapter 5 engages in a discussion of the applicability and utility of the *T&E 4Log Framework* and its components. The chapter begins by addressing the responses to each research question individually and then evaluates how well the overall research objective has been achieved. The validation of the artifact is then completed by critically reflecting on the research process.

Finally, Chapter 6 pinpoints the theoretical contributions as additions to the knowledge base and highlights the practical relevance of this research. This dissertation closes by providing directives for stakeholders to benefit from this research.

2 State of the Art

A thorough understanding of the current state of the art in both indoor localization and T&E is essential to identify gaps in the current research literature, recognize the scientific and practical relevance of closing these gaps, and build a knowledge base for the design of the research artifact. The two topics are addressed separately in this chapter, with Section 2.1 providing an overview of the state of the art related to indoor localization and Section 2.2 examining the current state of the art in T&E. This leads to identifying the research questions and delineating the preliminary research objective in Section 2.3.

2.1 Indoor Localization

The rise of satellite-based localization systems such as the Global Positioning System (GPS) and other Global Navigation Satellite Systems (GNSSs) have made localization ubiquitous in daily life, allowing services such as vehicle navigation, delivery tracking, or object finding, often through a single device such as a smartphone. However, satellite-based localization systems come with certain limitations. Firstly, its accuracy does not meet the requirements of many modern applications, such as vehicle automation in road traffic [33]. Secondly, obstacles such as walls, roofs, and floors obstruct or absorb satellite radio signals, making GNSS-based systems unreliable in indoor environments [34, p. 1]. The resulting demand for accurate, reliable, and cost-effective localization solutions for indoor environments has stimulated ongoing research and development of ILSs, which is reflected in the steady growth of publications dealing with the topic of indoor localization [14, p. 1].

To provide a solid foundation for this research, basic terms are initially defined in Section 2.1.1. Section 2.1.2 then provides an overview of relevant indoor localization technologies and Section 2.1.3 of typical intralogistics applications. Finally, the scientific literature on the user requirements of ILSs is summarized in Section 2.1.4.

2.1.1 Definitions of Basic Terms

Driven by divergent working principles and application domains, indoor localization technologies have been developed from several mostly distinct research fields. As a consequence, terms are often used inconsistently. To prevent misinterpretations, the unambiguous definitions of the basic terms for the understanding of the present work are set out below. Additionally, the definitions of the terms presented in this section are listed along with others in the glossary beginning on page xxv.

Much of the research on indoor localization focuses on position-determination capabilities. Hence, the term localization often refers to estimating an entity's position. However, with growing demand, determining an entity's pose, which combines position and orientation, is becoming increasingly important. Thus, in this work, localization refers to estimating an entity's location, with location denoting a discrete value pair of coordinates describing an entity's position and/or orientation within a reference space. Typically, the location is associated with a specific moment in time. A series of locations within a certain time frame is referred to as location data.

The entity to be localized is called the Entity to be Localized / Tracked (ELT), following the *ISO/IEC 18305* [17], whereby the term tracking emphasizes that the entity's pose is captured over time. In some cases, ILSs provide additional information such as velocity, acceleration, or position predictions [12, p.19]. If not, they can often be derived from location data, as shown in a previous publication [4].

Unlike geographic localization systems, ILSs typically provide location data in Cartesian coordinates. Thus, an entity's location in three-dimensional space is expressed by (x, y, z, ϕ, θ, ψ), including both position (x, y, z) and orientation (ϕ, θ, ψ) data, with the z-axis vertically aligned. Orientation components are referred to as roll (ϕ), pitch (θ), and yaw (ψ). Consequently, the location of an entity comprises up to six Degrees of Freedom (DoFs). If certain directions of linear or rotational motion are restricted, the DoFs are reduced accordingly. Figure 2.1 shows the location of a cuboid (O_{cuboid}) and a sphere (O_{sphere}) within the reference coordinate system (O_{ref}).

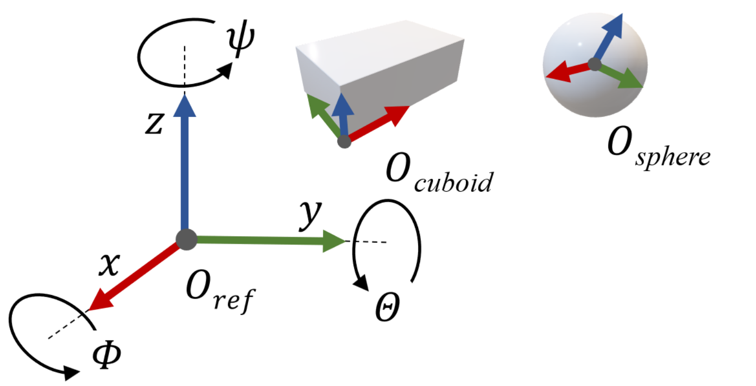


Figure 2.1: Cuboid (O_{cuboid}) and sphere (O_{sphere}) within Cartesian coordinate system (O_{ref})

Location data is typically categorized according to the type of reference coordinate system into absolute and relative location data. Absolute localization refers to the estimation of a location within a global coordinate system established by fixed landmarks. In contrast, relative localization involves the location determination within a local coordinate frame that is situated within a global reference frame [12, p. 27]. For example, if the reference frame in Figure 2.1 is defined by landmarks, localizing the cuboid or sphere within this frame is called absolute localization. Conversely, relative localization represents the location of the sphere relative to the cuboid or vice versa. In a typical intralogistics scenario, the global coordinate system is defined by some kind of warehouse structure. Absolute localization then concerns localizing entities such as a forklift or pallet within the industrial environment, while relative localization might involve determining a pallet's location with respect to a forklift truck.

Furthermore, positioning and locating are commonly employed terms in the literature, often used analogously to localization. This work adopts the definitions from the *ISO/IEC 18305* [17], according to which positioning refers to self-awareness, i.e., when the ELT uses the localization system to determine its own location. A typical example of this in an industrial context is the self-localization of a mobile robot. In contrast, locating is used when determining the location of a remote entity, such as in typical asset tracking applications. Hence, the term localization encompasses both positioning and locating.

In localization, multiple error components contribute to the deviations between the actual and estimated location of an entity. As described in the international vocabulary of metrology provided by the *Joint Committee for Guides in Metrology* [35], localization accuracy refers to the level of agreement between these two locations with the localization error describing the distance between them. Localization errors can be expressed element-wise or in combination, such as the horizontal position error, combining the horizontal position error components. When it comes to evaluating ILSs, localization accuracy is generally considered the predominant performance criterion. It is usually expressed by statistical measures such as the mean or median. However, in practice, localization accuracy is typically provided at the 95 % confidence level, expressed by the 95th percentile [12, p. 17].

2.1.2 Indoor Localization Technologies

Since the underlying working principle of an ILS largely determines its performance characteristics, a fundamental understanding of indoor localization technologies is essential for this research work. The scientific community has not yet agreed on a strict taxonomy to describe generic indoor localization technologies [36, pp. 4]. Often, the term indoor localization technology is used synonymously with the underlying sensor technology. However, since the underlying sensor technology is not sufficient to describe the working principle of an ILS, in this work the localization technology is considered a combination of the following three main component types of indoor localization technologies.

1. The **measurement technique** determines which physical quantity is measured in which way

2. The **localization method** describes how the measured quantity is used to determine the location of an ELT

3. The **sensor technology** encompasses the underlying methods that are used by the localization sensor to determine the measurand

An ILS is furthermore defined by its soft- and hardware implementation and possibly its deployment in the field. The relationship of the introduced terms is illustrated in Figure 2.2.

In addition to the presented taxonomy, there are other commonly used approaches to categorize ILSs that are not mutually exclusive. One such approach distinguishes between device-based and device-free localization [37]. Accordingly, device-based localization relies on an active device to determine its location, while device-free localization of an ELT is achieved without the

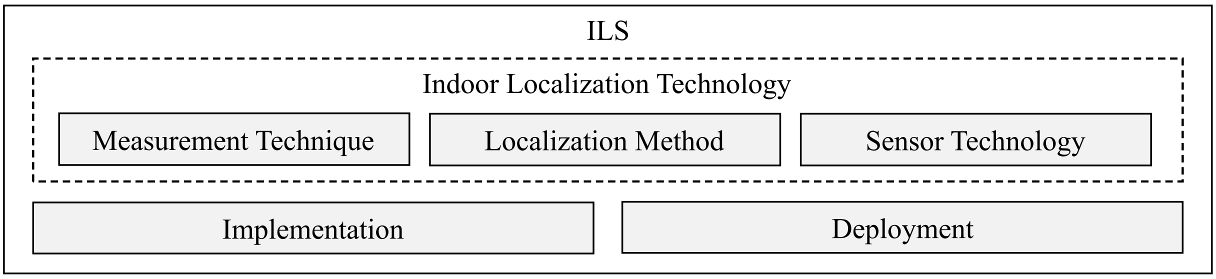


Figure 2.2: Components making up an ILS

need for a device attached to the ELT. In addition, marker-based localization refers to device-free localization that depends on passive markers attached to the ELT. Furthermore, indoor localization technologies are commonly divided into infrastructure-based and infrastructure-free solutions, depending on whether they rely on the deployment of dedicated infrastructure [38].

An overview of common components of indoor localization technologies is provided in Table 2.1. There are various combinations suitable for indoor localization, each with its own advantages and drawbacks. As noted in Bousdar Ahmed *et al.* [14], there is no individual technology that dominates the market. Instead, there has been a trend towards the hybridization of ILSs, which involves the integration of multiple components of the same component type. In particular, combining components with complementary working principles can significantly enhance accuracy and robustness. This approach is also known as sensor data fusion. A comprehensive overview of the methods for fusion of sensor data in the context of localization is provided by Guo *et al.* [39]. In the following, components of indoor localization technologies that are commonly used in industrial contexts are briefly introduced.

Measurement Techniques

Besides the measurement techniques for indoor localization presented as follows, others exist, such as Round Trip Time (RTT), Doppler Ranging, or Phase of Arrival (PoA). However their relevance for indoor localization in the context of intralogistics is limited. Therefore, for these,

Table 2.1: Components making up an indoor localization technology

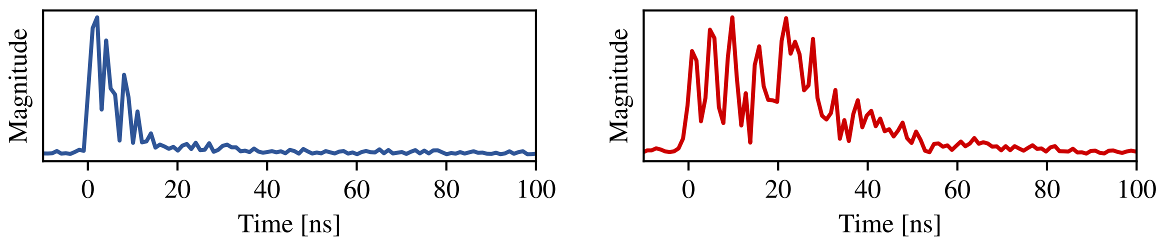
Component type	**Examples**
Measurement technique	Received Signal Strength (RSS), Time of Flight (ToF) / Time of Arrival (ToA), Time Difference of Arrival (TDoA), Angle of Arrival (AoA), wheel odometry, ...
Localization method	Cell of Origin (CoO) / proximity detection, lateration, angulation, map matching / fingerprinting, dead reckoning, ...
Sensor technology	Wireless Local Area Network (WLAN), Bluetooth, Radio-Frequency Identification (RFID), UWB, LiDAR, Vision, Inertial Measurement Unit (IMU), ...

reference is made to the fundamental literature for indoor localization as provided by Mautz [12] or Samama [40].

Received Signal Strength (RSS) The RSS represents the strength of a signal received by a receiver from an emitter. Depending on the underlying sensor technology, the RSS is affected by various environmental factors, such as obstructions or interferences. If environmental influences can be neglected or predicted, propagation models can be employed to determine the distance between the receiver and the emitter [12, p. 28].

Time of Flight (ToF) / Time of Arrival (ToA) The ToF (also called ToA) technique measures the time it takes for a signal to travel from an emitter to a receiver. The distance can then be estimated by multiplying the measured time by the known speed of signal propagation [36, p. 5]. ToF measurements require precise time synchronization between the emitter and the receiver, which can be challenging to achieve in practice. Moreover, accurate distance estimations based on ToF measurements are affected by the environment, for example, through absorption, reduced propagation speed, and multipath propagation, which refers to the signal taking multiple paths between the emitter and the receiver, leading to different time delays [12, pp. 28]. It is particularly problematic in the case of RF-based localization in so-called None Line of Sight (NLoS) situations. In contrast to Line of Sight (LoS), NLoS refers to a situation in which the direct path between the emitter and the receiver is occluded.

To detect and mitigate the impact of NLoS situations on the system performance, the Channel Impulse Response (CIR) can be used for pulsed signals [41]. It describes the response of a signal sent as an impulse, thus characterizing its propagation. Figure 2.3 depicts the CIR for two Ultra-Wideband (UWB) impulses in case of LoS (Figure 2.3 (a)) and NLoS (Figure 2.3 (b)). The graph for the NLoS case shows that simply looking at the highest or earliest peak to determine the distance between the transmitter and the receiver can be misleading. Instead, the full CIR can be considered for identifying NLoS and optimizing the localization accordingly.



(a) (b)

Figure 2.3: CIRs for (a) LoS and (b) NLoS situations (data provided by Kolakowski [42])

Time Difference of Arrival (TDoA) The TDoA technique is similar to the ToF method and therefore suffers similar influences. However, to mitigate the issue of accurate time synchronization, the differences in the arrival times of signals from different emitters are measured. This measurement can thus be used to estimate the differences between the distances from a receiver to different emitters. Possible positions can be formed in the form of hyperbolic lines, and the position of an emitter then results from the intersection of several lines [12, p. 30].

Angle of Arrival (AoA) The AoA technique measures the angle at which a signal arrives from an emitter at a receiver. Figure 2.4 illustrates the working principle in the case of RF-based localization. Based on the time or phase difference of arrival of an incoming signal at individual antennas of an antenna array, the angle Θ can be determined based on the distance d between to antennas [43, p. 2572]. Similar to ToF and TDoA, the quality of AoA measurements suffer from absorption and multipath propagation in NLoS situations. For RF-based localization, AoA is often combined with other methods to improve accuracy and robustness. Monocular cameras also employ AoA measurements for localization purposes, which are determined from the position of a pixel in the image plane [12, p. 31].

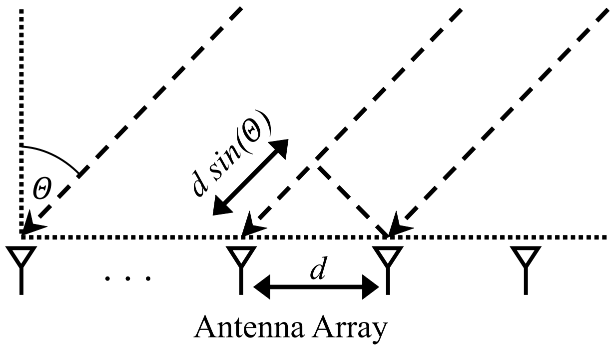


Figure 2.4: AoA technique with an antenna array (based on Zafari *et al.* [43, p. 2572])

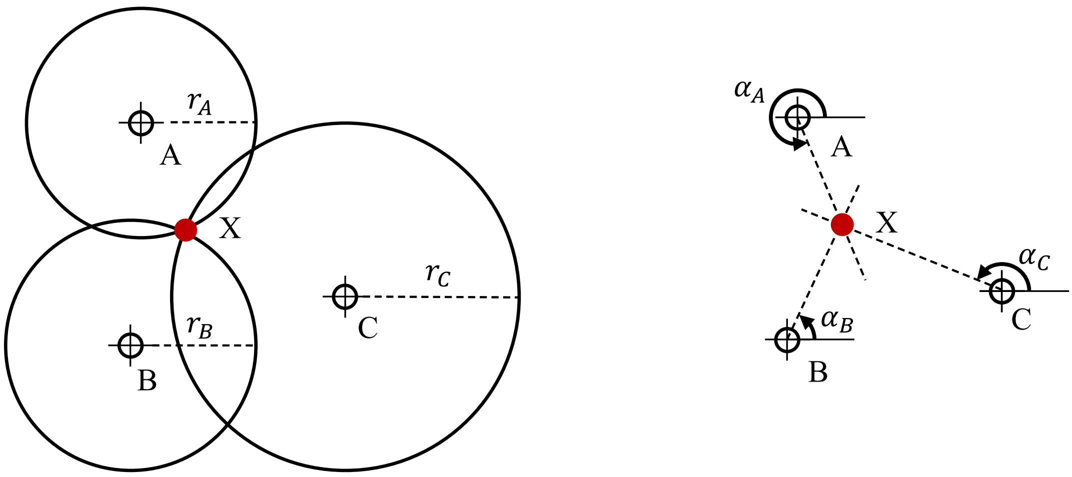
The presented measurement techniques provide physical quantities that describe intensities such as for RSS, distances for ToF or TDoA, and angles for AoA. In addition, the relative position, orientation, or velocity of vehicles can be estimated based on wheel odometry, i.e. by measuring the rotations of the wheels with known radii [44]. Wheel odometry is subject to certain sources of error, such as wheel slippage or uneven terrain. The orientation of an entity is often estimated based on the earth's magnetic field [12, pp. 100]. Furthermore, angular velocity and linear acceleration are commonly estimated through the concept of inertia using Newton's second law, which describes the relationship between force, mass, and acceleration [45]. In the following, localization methods are employed to estimate an entity's location based on the provided measurements.

Localization Methods

Various methods can be used to determine the location of an entity based on physical quantities, such as those provided by the measurement techniques described above. In the following, the most relevant localization methods are presented within the context of this research.

Proximity Detection / Cell of Origin (CoO) With proximity detection or CoO, the location of an entity is simply determined by its presence in a particular area, sometimes referred to as a cell. This method usually assumes that the entity's coordinates are identical to those of the sensor node associated with the strongest RSS [46]. Hence, the accuracy of this method is strongly dependent on the density of the sensor network and the characteristics of signal propagation within the environment [17]. This type of detection serves as a straightforward localization method, typically employed for applications with low accuracy requirements [12].

Lateration Lateration can be used for localization that relies on distance estimates obtained from RSS or ToF measurements. Figure 2.5 (a) illustrates the fundamental principle of lateration in the horizontal plane. To determine the position of an entity (X), the measured distances (r_A, r_B, r_C) from multiple reference points (A, B, C) are used. The position is ideally obtained from the intersection of circles (or spheres in 3D) with radii equivalent to the measured distances. In practice, errors can occur, for example, due to deviations between the determined and true positions of the reference points. Often, more than the required amount of measured distances to the reference points are available, making the problem overconstrained. In this case, a least squares optimization approach is typically applied [17].



(**a**) (**b**)

Figure 2.5: Localization through (a) lateration and (b) angulation (based on Farid *et al.* [47, p. 2572])

Angulation Angulation can be used for localization based on angle measurements. The method's working principle is illustrated in the horizontal plane in Figure 2.5 (a). Angulation determines an entity's position by identifying the intersection of multiple lines, each representing an angle (α_A, α_B, α_C) between the entity (X) and a reference point (A, B, C) with known position [12, p. 31]. Furthermore, the orientation of the entity in space can be estimated [48].

Map Matching / Fingerprinting Map matching is a localization method that involves comparing sensor data with a map of the environment. When applied to RF-based solutions, it is usually referred to as fingerprinting, which is used predominantly in conjunction with RSS [12, p. 32]. The process is illustrated in Figure 2.6 (a). Here, the RSS measurements are recorded and associated with locations in an offline phase. In the online phase, the RSS values are compared to the pre-recorded fingerprints to determine the corresponding location.

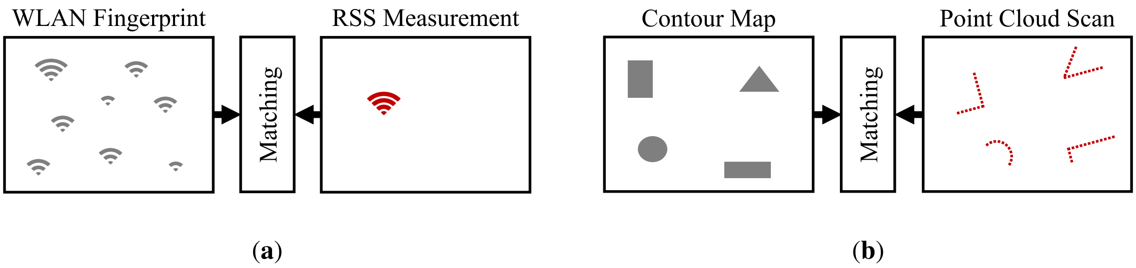


Figure 2.6: Localization through (a) WLAN-based fingerprinting with RSS and (b) LiDAR-based map matching

Furthermore, map matching is commonly used in mobile robotics for robot positioning based on numerous measurements of the distance and angle measurements of the spatial environment, represented as a point cloud (Figure 2.6 (b)). This point cloud can be used for localization by matching sensor data with an existing map or to generate a contour map of the environment [49].

A significant drawback of this method is its reliance on an environmental map, which results in a dependency on the environment and environmental changes. Regular map updates can help mitigate the challenges posed by dynamic environments. A widely used approach to generate spatial maps for localization purposes is Simultaneous Localization and Mapping (SLAM). In SLAM, map matching is utilized for both localization and generation of a contour or feature map of the environment without relying on a pre-existing map. For an in-depth understanding of SLAM's working principles, Thrun *et al.* [50] and Aulinas *et al.* [51] are recommended resources.

Dead Reckoning Another commonly used localization method is dead reckoning. It involves estimating an entity's movement based on its previously known or assumed location and is typically implemented based on measurements of the entity's acceleration and/or velocity. Consequently, in contrast to the methods previously presented, dead reckoning provides relative location information. Thus, dead reckoning is often combined with other absolute localization methods through sensor data fusion techniques to improve their accuracy and robustness [36, p. 9].

A popular algorithm for dead reckoning is the Kalman filter. This algorithm functions by predicting an entity's location based on a model of its motion and its previous location, subsequently updating the prediction using new sensor measurements. The Extended Kalman Filter (EKF) is a variant of the Kalman filter designed for non-linear models. The Kalman filter and, in particular, the EKF are also frequently used as an algorithm to fuse sensor data from different sources. These and other widely used algorithms based on dead reckoning, such as particle filtering, are elaborated in detail by Panigrahi *et al.* [49].

The presented localization methods can be employed to estimate an entity's position and partially its orientation based on different kinds of measurands. However, if orientation data are not provided directly, it can usually be calculated in a subsequent step using knowledge of the position of multiple points of the ELT with known local coordinates [52].

Sensor Technologies

A diverse range of sensor technologies can be harnessed to determine physical quantities such as ToF or AoA with the previously introduced measurement techniques. The type and quality of sensor data are dependent on the underlying sensor technology, resulting in typical combinations with measurement techniques and localization methods. This section introduces the prevalent sensor technologies for indoor localization, categorizing the typically associated ILSs into infrastructure-based and infrastructure-free, as well as device-based and device-free solutions. Furthermore, rough indications are made regarding the absolute position accuracy and costs for ILSs based on the respective sensor technology.

Wireless Local Area Network (WLAN) Primarily designed for network communication, indoor localization can be implemented using standard WLAN. Hence, WLAN-based ILSs can often be built on top of preexisting WLAN-infrastructure. In WLAN localization, a mobile device emits a radio signal that complies with the IEEE 802.11 standard, which is then received by a router [53]. The location of the mobile device can be determined by measuring the RSS and employing different methods, such as proximity detection, lateration or fingerprinting [43].

Bluetooth Bluetooth is a globally recognized standard (IEEE 802.15.1) for facilitating personal area networks among end devices, operating with signal pulses at a frequency of 2.4 GHz [54]. The technical infrastructure of this technology is both easily accessible and cost-effective [34]. Although Bluetooth is compatible with a variety of measurement techniques, such as ToF and AoA the majority of existing solutions depend on RSS measurements [43].

Radio-Frequency Identification (RFID) A localization system based on RFID consists of one or more RFID readers and RFID tags. The RFID tag can either actively emit signals or be stimulated by signals sent by the reader, thus obviating the need for an additional external energy source [55]. Localization with RFID is usually achieved using the proximity detection method. Due to the cost-effective, flexible, and scalable nature of RFID technology, it is frequently used for simple localization and identification tasks [56].

Ultra-Wideband (UWB) With UWB, short pulses of broadband radio signals are emitted periodically by so-called UWB tags within a frequency spectrum of 3.1 GHz to 10.6 GHz. Due to the broad frequency band, UWB signals exhibit increased resistance to electromagnetic interference and remain less affected by the absorptions of obstacles [57]. Usually, UWB-based localization is achieved with ToF, TDoA, and/or AoA measurements. Furthermore, short pulses allow for the detection of NLoS by analyzing the CIR [43].

In industrial contexts, ILSs based on RF signals such as WLAN, Bluetooth, or UWB are often used for real-time remote locating, forming a subcategory of ILSs called Real-Time

Locating Systems (RTLSs). Contrary to GNSS-based systems, RTLSs allow placing the emitter and receiver of the signals within the indoor environment, leading to increased accuracy and reliability of the location data, and thus allowing a wide range of indoor applications. They usually require fixed nodes with known locations as well as a mobile node attached to the ELT and are thus device- and infrastructure-based. All of these RF-based systems can be affected by radio interference and occlusion from metal objects [17]. Depending on the measurement technique, localization method, implementation, and deployment, the position accuracy of WLAN-, RFID- and Bluetooth-based ILSs is typically in the range of meters, while UWB-based ILSs commonly reach values less than one meter, as indicated in several literature surveys [36, 52, 58]. However, compared to the typically low-cost and scalable RF-based ILSs, the costs of UWB-based systems are comparatively high [52].

Other RF-based sensor technologies, such as *ZigBee*, *Long Range Wide Area Network* (LoRaWAN), or 5G exist and can also be used for localization purposes. Although the specifications of the radio signals differ, the fundamental principles employed for localization remain consistent across these technologies, allowing the findings of this work to apply to them as well. For a detailed overview of the specifications of these sensor technologies, the literature surveys provided by Zafari *et al.* [43] and by Peral-Rosado *et al.* [59] are suggested.

Light Detection and Ranging (LiDAR) In LiDAR-based localization, LiDAR-scanners are typically mounted on the ELT to enable device-based, infrastructure-free localization. LiDAR-scanners operate by emitting high-frequency laser pulses, which, upon reflection by objects, are registered by a dedicated detector within the scanner. The distance to the reflected point is determined with the ToF technique [60]. Conventional LiDAR-scanners incorporate a rotating mirror, which periodically alters the orientation of the laser pulse, thereby enabling distance measurements with an angle of up to 360° around the sensor's central axis. Multilayer scanners typically integrate multiple emitters and receivers, enabling a 3D representation of the surveyed environment. The map matching method can then be used to achieve localization, whereby SLAM-algorithms are typically applied for map recording [61].

LiDAR-based ILSs usually provide position and orientation data with high accuracy. The results of a comparative study of several SLAM algorithms, based on LiDAR data from experiments in a warehouse scenario, show values for the position accuracy in the range of centimeters [62]. By providing reflective infrastructure elements, the accuracy and robustness of localization can be further increased. However, the costs for a LiDAR scanner are typically high, leading to reduced scalability. Therefore, solid-state LiDAR sensors have emerged in recent years as a cost-effective alternative to conventional mechanical systems [63].

Vision Different types of vision technologies can be utilized for localization purposes. Monocular cameras, for example, rely on detecting features or markers within the image plane to compute the AoA. Frequently, 2D markers such as AprilTags are used to facilitate localization and identification [64]. For device-free remote locating with a monocular camera, the angulation method estimates the position of a detected feature or marker based on AoA measurements from multiple cameras with known locations [65]. On the other hand, device-based self-localization can be achieved by attaching a monocular camera to the ELT. AoA measurements are then used for positioning by applying the map matching method, based either on a preexisting feature map or by employing SLAM [66].

Besides monocular Red-Green-Blue (RGB) cameras, Red-Green-Blue-Depth (RGBD) cameras can be used for localization, providing additional depth information. For example, stereo cameras consist of two synchronized camera frames, typically mounted at a fixed distance apart. By comparing the disparities between the two images from slightly different viewpoints, depth information can be derived. On the other hand, ToF or structured light cameras capture both color and depth information in a single frame. Structured light cameras typically employ an infrared projector and sensor to measure depth by analyzing the pattern distortion. Device-free, as well as device-based localization, can be achieved based on RGBD-cameras using methods similar to monocular cameras. However, localization can be facilitated by utilizing additional depth information. An in-depth overview of vision-based localization is provided by Wu *et al.* [67].

Finally, Infrared (IR) cameras are commonly utilized for device-free locating, in the pursuit of high-performance localization. The cameras emit high-frequency IR flashes that are reflected by passive IR reflectors. These are attached to the ELT to capture distinct and identifiable features within the camera image. This technology, also known as optical passive motion capture, generates highly accurate location data with up to six DoFs. Nonetheless, due to the significant investment costs associated with this technology, it is primarily used in laboratory settings [68].

The usage of cameras for localization is diverse. Thus, accuracy specifications vary by several orders of magnitude. For example, Knitt *et al.* [69], when locating a pallet with a monocular camera and feature-based localization, achieve a position accuracy in the range of a few decimeters. However, a study conducted by Bostelman *et al.* [70] on the performance of optical passive motion capture systems for the tracking of Automated Guided Vehicles (AGVs) shows position errors in the range of millimeters. A significant disadvantage with vision-based localization is that localization methods typically require direct LoS and their sensitivity to poor lighting conditions, applying to both over- and under-exposure of the scene. In the case of optical motion capture systems, sunlight and IR-reflective surfaces are a major concern [17].

Inertial Measurement Unit (IMU) Inertial Measurement Units (IMUs) are self-contained devices that typically consist of accelerometers, gyroscopes, and magnetometers. Accelerometers measure linear acceleration along the three orthogonal axes, while gyroscopes measure angular velocities [45]. Both are based on inertia measurements. Magnetometers, when included, measure the strength and direction of the Earth's magnetic field, providing a global reference about the entity's orientation.

As stated in the *ISO/IEC 18305* [17], the quality and price of accelerometers and gyroscopes can vary widely, which is reflected in the quality of the data output. IMU data are commonly employed by dead reckoning algorithms to improve localization in conjunction with other sensors [12, pp. 100]. Since the localization of a purely IMU-based dead reckoning system is subject to drift, an indication of absolute accuracy is not meaningful. Accelerometers and gyroscopes commonly suffer from misalignment and require regular calibration, while magnetometers naturally suffer magnetic interferences, such as by the presence of ferrite materials [17].

In addition to the named sensor technologies, wheel encoders are commonly employed for the localization of vehicles and mobile robots based on wheel odometry. Analogous to IMU-based systems, ILSs based solely on wheel odometry do not allow absolute localization. Various other sensor technologies can be employed for localization, including magnetic sensors, tactile sensors, and sound sensors. For further insights, the comprehensive reviews provided by Mautz [12] or Samama [40] are recommended.

Throughout this section, it became apparent that ILSs are generally heterogeneous and complex CPSs. The technology of an ILS consists of a combination of one or multiple measurement techniques, localization methods, and sensor technologies, each of which exhibiting unique characteristics. Rough indications of the absolute position accuracy based on the results of different studies have been provided. However, the comparability of such studies is naturally limited, as the data were recorded and evaluated under significantly varying conditions. The issue will be discussed in detail later in this work.

2.1.3 Intralogistics Applications

To emphasize the diversity and benefits of the use of ILSs, this section presents exemplary intralogistics applications of ILSs that have been discussed in the scientific literature. The applications are grouped according to the type of ELT, as summarized in Table 2.2. Finally, the assignment of typically applied sensor technologies is briefly discussed.

Personnel Localizing personnel in intralogistics can enhance safety, enable monitoring of staff movements for process optimization, and support human workers in their daily operations. Exemplary applications from the literature include trajectory analysis for process optimization and optimized navigation of workers within the facility [71], Augmented Reality (AR) for picking support [72], or contact tracing for preventing the spread of diseases [73]. Data privacy is a sensitive issue for localizing human workers and must be treated accordingly. This is particularly relevant for camera-based localization.

Goods and Load Carriers Goods comprise items or materials that require transportation, storage, or handling within an intralogistics facility. Load carriers such as pallets, crates, or

Table 2.2: Exemplary applications of ILSs for different entity types with references from the scientific literature

Personnel	Goods & Load Carriers	Industrial Trucks	Mobile Robots
• Trajectory analysis [71] • Navigation [71] • Picking support [72] • Contact tracing [73]	• Automated booking [5] • Value-stream mapping [74] • Material flow analysis [3] • Order fulfillment [3]	• Order allocation [5] • Operation analysis [4] • Collision avoidance [75] • Semi-automatic pallet pick-up [7]	• AMR use cases [76] • Automated Commissioning [77] • Autonomous forklift [6] • Inventory management [78]

containers facilitate the storage and movement of these goods. Localizing goods and load carriers can support efficient inventory management, material handling, storage space optimization, and reduces search times. For example, Hesslein *et al.* [5] present an application for automated booking processes based on the association of load carriers to storage compartments. An application for digital value-stream mapping with an UWB-based ILS is presented by Sullivan *et al.* [74]. Finally, Mütze *et al.* [3] showcase several use cases, including material flow analysis for layout planning and support of order fulfillment.

Industrial Trucks Industrial trucks, such as forklifts and tow trucks, play a crucial role in intralogistics, facilitating movement, transportation, and handling of certain goods and load carriers within industrial environments. Their localization allows for a multitude of improvement potentials within intralogistics operations. Hesslein *et al.* [5], for example, propose a location-dependent order allocation algorithm to prevent waiting times for industrial trucks in narrow aisles. Furthermore, a paper was presented, describing the analysis of the operation of industrial trucks to identify inefficiencies and potential safety hazards [4]. Thiede *et al.* [75] present a camera-based collision avoidance application between human and industrial trucks. Finally, Molter *et al.* [7] present a driver assistance system for semi-automatic pallet pickup based on camera technology.

Mobile Robots The *DHL Trend Radar* 2022 identified mobile robots as the technology trend with the most significant impact in the logistics industry [79, pp. 116]. Localization represents a key capability of mobile robots. In industrial settings, the term AGV refers to a mobile robot with basic line-following capabilities, or in the case of Autonomous Mobile Robots (AMRs), extended decision-making abilities for partial autonomy [80]. Unger *et al.* [76] evaluate ten factory use cases for AMRs, from simple material transport to advanced tasks like packaging. The wheeled robot "Stretch" was showcased by *Boston Dynamics* for palletizing and commissioning [77]. Behrje *et al.* [6] proposed a vision-based localization system for autonomous forklift trucks. Finally, Beul *et al.* [78] demonstrates an automated inventory management application based on drones.

The multitude of potential intralogistics applications of ILSs offers substantial benefits to improve or enable various aspects of intralogistics operations. This statement is underpinned by a survey conducted by Thiede *et al.* [71]. The assessment of the relevance of various applications through production and warehouse managers highlights the interest in particular for tracking goods, load carriers, and industrial trucks, as well as for mobile robotics applications, and automated booking processes. Certain applications or application categories are commonly linked to particular sensor technologies. For example, due to the portable size of the localization devices and the scalability of the technology, tracking goods, load carriers, or personnel is generally achieved by employing some kind of RTLSs. More performant and costly systems, such as those based on LiDAR-scanners, are instead employed for automation tasks or safety-critical driver assistance systems, while cameras, offering the greatest versatility, are found in all discussed application fields.

2.1.4 User Requirements

User requirements in the context of indoor localization are specifications that an ILS must meet to be suitable for a particular task or application. Thoroughly examining user requirements of ILSs is essential not only for system users to assess the suitability of the system, but also for researchers and system developers to align their efforts accordingly [12, p. 15]. Requirements are typically divided into functional and non-functional. While functional requirements describe the intended behavior or specific features of a system, non-functional requirements outline the characteristics or attributes, such as performance, reliability, usability, and maintainability [81, 82]. Since this work deals with the evaluation of ILSs based on experimental investigations, the focus here is on non-functional requirement parameters corresponding to a system's localization performance, which can be determined appropriately through experimental investigations in the context of this work. These types of user requirements related to the data output of an ILS are referred to as location data requirements. To enable the development of a T&E methodology, this section provides an overview of the scientific literature dealing with location data requirements of ILSs. Firstly, the parameters for location data requirements relevant to this work are presented. Subsequently, the methodological approaches for their specification are briefly discussed. A significant portion of the content of this section has already been part of a previous publication, which was presented at the *Conference on Production Systems and Logistics* (CPSL) 2023 [83].

Parameters for Location Data Requirements

The most holistic examination of requirements for ILSs was conducted by Mautz [12] as part of a comprehensive literature survey. The author compiled a list of clearly defined user requirement parameters across the four categories: "positioning", "human-machine interface", "security and privacy", and "costs". This work focuses on the "positioning" category, which corresponds to the previously defined category of location data requirements. The related parameters are shown in Figure 2.7. The most relevant parameters for this work are briefly described in the following, based on Mautz [12, pp. 15], while highlighting their significance for intralogistics applications. In Figure 2.7, these parameters are highlighted in gray.

Data Output The data output refers to the DoFs of the provided location data. It is a crucial requirement parameter as it determines the type of location information needed for an application. For example, basic asset tracking applications in intralogistics may only require horizontal position data (x, y), while safety-critical driver assistance systems like collision avoidance typically demand 3-DoFs pose data (x, y, ψ).

Localization Accuracy The term localization accuracy was previously introduced in Section 2.1.1. Since it directly relates to the quality of a location estimate and varies considerably in practice, localization accuracy is often considered the most critical requirement parameter. Simple tracking applications typically require a horizontal position accuracy within a few meters, while sophisticated automation solutions may demand centimeter-range accuracy or better.

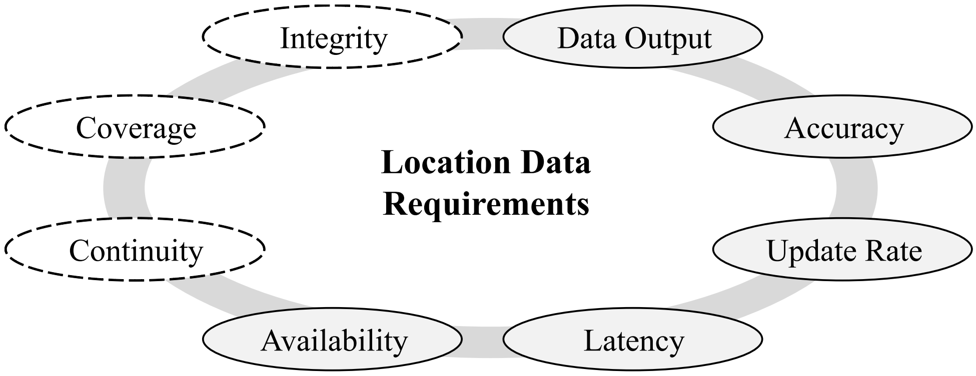


Figure 2.7: Parameters for location data requirements described by Mautz [12] for the category "positioning"

Update Rate Update rate, or update frequency, refers to the rate at which an ILS provides location updates. Generally, ILSs offer location updates at a constant rate f_{update}, with the time gap between consecutive measurements given by $t_{gap} = \frac{1}{f_{update}}$. During this time, an application must rely on previous measurements, since location changes between two measurements remain unknown. Therefore, a higher update rate enables more frequent and up-to-date information about the ELT's location. Besides a constant update rate, location data can also be supplied upon external request or triggered by an event [12, p. 19].

System Latency System latency denotes the time delay (t_{lat}) between the actual measurement and the provision of the location update, with the associated timestamp typically set to the point in time of measurement [12, p. 19]. Thus, system latency indicates the ability of an ILS to detect and report the location change of an ELT. Figure 2.8 illustrates the difference between the time gap and the time delay. Lower latency indicates that the system can react more quickly to location changes, which is crucial for real-time applications such as collision avoidance.

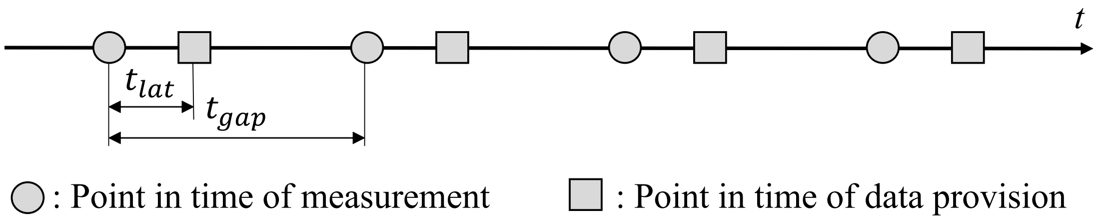


Figure 2.8: Time gap (t_{gap}) and time delay (t_{lat}) for an ILS with constant update rate

Availability Availability in the context of localization systems is defined as the proportion of time during which location data is provided under the given requirements. In practice, this value is predominantly determined by operational constraints, such as maintenance tasks. According to Mautz, values below 95 % corresponds to low availability, while values above 99.9 % indicate high availability [12, p. 18].

Furthermore, Mautz identifies the parameters "continuity", "coverage", and "integrity" as relevant location data requirement parameters. However, they can hardly be determined in a meaningful way based on empirical experiments in partially controlled test environments and are therefore not further considered in this work. In the following, the approaches presented in the literature for specifying location data requirements are briefly discussed.

Approaches for the Specification of Location Data Requirements

There exist generic guidelines and best practices for requirement engineering and management, such as the *ISO/IEC/IEEE 29148* [81] and the *ISO/IEC 25010* [82], offering systematic approaches to define, document and validate requirements in general. Although these generic and high-level methodologies are undoubtedly important, incorporating domain- and technology-driven considerations provides more concrete guidance and facilitates better decision making.

Several scientific publications address the specification of requirements for ILSs as part of a broader system selection approach. For example, Jordan *et al.* [84] propose a requirement-based matching for CPSs that employs a set of morphologies. However, this method only focuses on the identification of appropriate technologies based exclusively on functional requirements.

Moreover, Gladysz *et al.* [85] outline a procedure for selecting a suitable ILS for intralogistics applications. The initial step of this process involves defining the requirements by detailing the business case and establishing the acceptable limits of the requirements. The authors illustrate the utilization of this approach through a case study that involved a forklift control and diagnostic tool in a cold chain warehouse. However, the authors do not elaborate on the underlying logic of how the requirements are specified, leaving a gap in understanding this crucial aspect.

Finally, Mautz [12] offers a method for deriving user requirements for ILSs, independent of the application domain. The consecutive steps of this method are (1) “Definition of potential user groups”, (2) “Definition of potential services”, (3) “Definition of high-level functions”, (4) “Definition of required parameters”, (5) “Data acquisition”, (6) “Detailed description of user requirements”, and (7) “Summary of user requirements in table form”. In this approach, Step 5 “Data acquisition” is simply executed by conducting user surveys. However, it is unknown how users specify requirements. Hence, similar to the method presented by Gladysz *et al.*, the procedure of Mautz remains at a high level and lacks concrete guidance on how to specify location data requirements.

A study on the quantification of the accuracy requirements of localization systems for autonomous road vehicles was presented by Reid *et al.* [33]. The authors’ procedure is based on geometric considerations that ensure the allocation of the vehicle on the road under the premise that the failure rate corresponds to an improvement in road safety. Although the authors provide an interesting approach for autonomous road vehicles, most follow-up considerations are only valid for this particular use case.

Finally, Hohenstein *et al.* [86] conducted a survey to evaluate the suitability of 25 RTLSs for localizing forklift trucks, based on the criteria “localization accuracy”, “outdoor capability”, “flexibility”, and “scalability”. The localization accuracy was defined as the 95th percentile of the horizontal position error and the size of the object to be localized was identified as the relevant criterion for quantifying the accuracy requirement. A case study demonstrated how the horizontal position accuracy requirement with respect to the center of a pallet is determined by half the width of the storage location during an automated booking process. The authors provide rough range estimations for the required localization accuracy of five common intralogistics areas or objects, including storage area, storage aisle, and storage location. However, the authors do not provide further insight into how to transfer these basic considerations to other intralogistics applications.

Although several publications discuss the performance requirements of ILSs, they lack depth in terms of quantifying location data requirements. The relationship between different system performance indicators and location data requirements is barely explored. The approaches presented by Reid *et al.* [33] and Hohenstein *et al.* [86] to determine location data requirements based on the dimensions of the ELT or the area of interest are intriguing. However, to create substantial added value for stakeholders, these basic considerations must be further developed and integrated into a systematic approach.

2.2 Test and Evaluation

T&E serves to determine the performance and potentially the suitability of an ILS for specific tasks, based on empirical experiments. To accomplish this, an ILS or its components must be deployed within a designated environment. The deployed system is referred to as the System under Test (SuT). Throughout the experimental phase, data is gathered from the SuT while the ELT is moved through the environment. The conditions outlining an experiment are described in a scenario, while an experiment refers to the actual execution of a given scenario. Upon successful completion of an experiment, various performance metrics are computed from the collected data, whereby the absolute localization accuracy is typically considered the dominant characteristic.

A term strongly related to T&E and commonly used in the literature is benchmarking. However, benchmarking is used mainly when there is an emphasis on the comparability of the results. In this work, the terms are distinguished as follows. Although benchmarking encompasses the set-up, execution, and evaluation of an experiment, T&E additionally includes steps for defining scenarios, specifying user requirements, and determining the system's suitability for an application.

The available literature on T&E in the field of intralogistics is limited. Hence, this section offers a comprehensive overview of the current state of the art in T&E in general. Initially, a taxonomy for T&E approaches is presented. Next, in Section 2.2.2, insight is provided on how T&E is performed in scientific research, thereby pinpointing the key limitations and the main challenges. Subsequently, commonly applied practices related to T&E are presented in Section 2.2.3. Finally, Section 2.2.4 elaborates on the existing methodologies. The material covered in this section has been previously addressed to some extent in a previous publication titled "Meaningful test and evaluation of indoor localization systems in semi-controlled environments" [87].

2.2.1 Taxonomy of Approaches

The *ISO/IEC 18305* [17] provides a straightforward taxonomy for T&E approaches, based on idealized edge cases. In this work, this taxonomy is used to categorize existing research and to delineate the research objective. In practice, however, the categorization is often not that clear, and combinations or compromises are common. The edge cases that serve as the basis for the taxonomy according to the *ISO/IEC 18305* are described as follows, highlighting their strengths and weaknesses.

Black-box vs. White-box Testing [18, p. 8] Testing approaches for ILSs can be divided into black-box and white-box testing, which differ in the extent to which knowledge about the system's internal mechanisms influences the testing process. With black-box testing, performance is solely evaluated by examining the system's in- and outputs, without taking the underlying algorithms or internal structures into account. This approach emphasizes the investigation of the performance of the system in real-world scenarios and supports the comparability between various ILSs. As end-users are more concerned with fulfilling performance requirements rather than understanding the system's inner workings, for them black-box testing is typically the preferred approach. In contrast, white-box testing involves a comprehensive understanding of an ILS's internal mechanisms, significantly influencing the testing procedure. This approach implies that the design of T&E procedures is tailored for the technology of the SuT. Therefore, the results lose generality, thus reducing the comparability between systems based on different technologies. It is the approach typically employed when specific properties of a technology need to be examined [17, p. 10].

System- vs. Component-level Testing [17, p. 9] System-level testing involves assessing the performance of the entire system as a holistic unit, considering the combined operation of all its components. It is the approach, usually chosen in conjunction with black-box testing to allow the evaluation of an ILS for a real-world application (Figure 2.9 (a)). Furthermore, this is typically the case for examining commercial systems whose inner workings frequently remain proprietary. It is also possible to combine system-level and white-box testing (Figure 2.9 (b)) if the testing procedure should be influenced by the ILS's inner workings. Component-level testing focuses on individual parts or modules of the system, such as sensors or algorithms. By examining these components separately, researchers and developers can identify specific areas of improvement, optimize individual elements, and deepen the understanding of how each component contributes to the overall performance of the system. Most often, but not exclusively, it is used in conjunction with white-box testing (Figure 2.9 (c)).

Repeatable vs. Non-Repeatable Testing [17, p. 10] Repeatable testing emphasizes the ability to perform the same experiment multiple times under identical conditions, ensuring consistent and comparable results. This allows for a more differentiated analysis of the system's performance, eliminating random factors that may influence the outcomes. Hence, repeatable testing

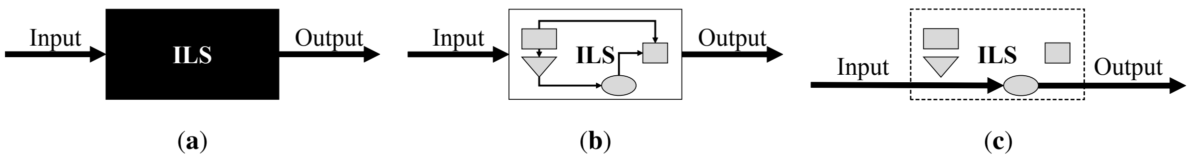


Figure 2.9: Approaches for T&E: (a) Black-box & System-level testing, (b) White-box & System-level testing, (c) Component-level testing

enables the evaluation of influences on the system's performance by conducting experiments under varying influencing factors. Non-repeatable testing involves testing the SuT under different, often unpredictable conditions. Although this approach may not offer the same level of consistency as repeatable testing, it can provide valuable insights on how the system performs in real-world scenarios.

Building-wide vs. Laboratory Testing [17, pp. 10] Building-wide testing involves evaluating the performance of an ILS in a large-scale, real-world environment. This approach provides a more realistic assessment of the system's capabilities, as it naturally accounts for various influencing factors, such as multipath propagation, interference, and dynamic obstacles typically encountered in real-world applications. However, building-wide testing can be resource-intensive and time-consuming, while resulting in poor repeatability. Laboratory testing focuses on evaluating ILSs within a controlled environment, such as a lab or a smaller confined space. This approach allows for control over test parameters, enabling increased repeatability, and allowing study of the effects of specific factors on the system's performance. Moreover, laboratory testing is typically less expensive than building-wide testing. Its main drawback consists of the limited transferability of the results, as the controlled environment may not accurately represent the complexities and challenges encountered in real-world scenarios.

Reference ILS vs. Offline Surveyed Points [17, p. 11] As discussed previously, absolute accuracy is typically considered the dominant performance characteristic of an ILS. In order to evaluate a system's accuracy, its true location must be known. This so-called Ground Truth (GT) can either be provided by a reference ILS or a set of offline surveyed points, with the same or more DoFs and significantly higher accuracy than the SuT. Offline surveyed points denote a collection of predetermined location data points. During the experiment, these points are sequentially visited by the ELT, and the corresponding data generated by the ILS is recorded. This approach is often chosen for simple investigations or when no reference ILS is available with satisfying accuracy. Alternatively, employing a reference ILS as a GT provides additional information, allowing the analysis of the entire trajectory, as well as spatiotemporal relationships.

In conclusion, the *ISO/IEC 18305* presents a classification framework for T&E, which takes the following factors into account: the extent to which the inner workings of the SuT impact the testing procedure, whether the entire system or just its individual components are examined, the ability to obtain similar results when repeating experiments, the testing environment with respect to the size and control of test conditions, and the manner in which the GT is supplied. This taxonomy serves as a foundation for understanding and comparing T&E procedures in the upcoming sections.

2.2.2 Limitations and Challenges

Research papers dealing with the development of indoor localization systems commonly incorporate some form of evaluation, often through empirical experiments. Adler *et al.* [88] and Bousdar Ahmed *et al.* [14] reviewed and compared such evaluations. The main findings of these reviews are briefly discussed in the following, along with the key limitations mentioned in the literature. Subsequently, the most relevant challenges for T&E are pointed out.

Adler *et al.* [88] conducted a comparative analysis of T&E procedures based on 183 randomly selected publications from the proceedings of the *International Conference on Indoor Positioning and Indoor Navigation* (IPIN) conferences between 2010 and 2014. The study found that 77 % featured physical experiments ranging from simple office walks to real-world tests in various buildings. Although absolute position accuracy is often considered the main performance criterion, there are significant inconsistencies in the choice of metrics and GT determination. Along with the lack of transparency highlighted by the authors, this variance in applied procedures compromises the comparability of the results. Nevertheless, results are frequently summarized and compared in the scientific literature to provide indications of system performance [12, 36, 43, 57]. However, their quantitative results must be treated with caution [89, p. 1161]. Adler *et al.* [88, p. 44] furthermore notes, that the results of such studies often do not translate well to real-world scenarios. In this work, this limitation is described by the term transferability.

Bousdar Ahmed *et al.* [14] carried out a more recent study focusing on T&E procedures based on *ISO/IEC 18305* characteristics. The research analyzed publications dealing with ILSs for pedestrian localization. Although often not even explicitly stated, the examined experiments exhibited considerable diversity in terms of building type, number of floor levels, type of motion, number of evaluation points, applied accuracy metrics, and GT type. The authors emphasize the importance of dedicating more effort to developing and documenting adequate T&E procedures, which will enhance the repeatability and reproducibility of the T&E results. While repeatability is given by the consistency of the results of the experiment when performed multiple times on the same system under identical conditions [17], reproducibility ensures that the results can be obtained by independent researchers [14, p. 486]. Thus, it is a key pillar of science in confirming or comparing research works. In the case of T&E, reproducibility includes replicability, which refers to obtaining consistent results in different test facilities [14, p. 486].

In summary, the key limiting factors for T&E identified in the existing literature are the lack of repeatability, comparability, transferability, and reproducibility of results. These are partially in conflict with each other. For example, to achieve repeatable and reproducible experiments, controlled testing conditions are required in similar test facilities, whereas transferability is attained when conditions resemble real-world application scenarios, which are typically unique in their concrete manifestation [14, p. 488]. As a result, together these limitations can only be overcome to a certain extent. Finding a suitable trade-off to provide significant results based on stakeholder interests ultimately increases the meaningfulness of the results.

The stated limitations are the consequence of several challenges regarding T&E(Figure 2.10). When designing a T&E procedure or developing a methodology, one must be aware of these challenges. Hence, they are outlined as follows.

- **Heterogeneity of technologies:** The heterogeneity of indoor localization technology has been emphasized in Section 2.1.2. As ILSs based on various technologies are subject to different influencing factors, designing a T&E procedure that enables a fair comparison between systems based on distinct technologies presents a major challenge.
- **Complexity of systems:** As highlighted by Potortì *et al.* [11, p.209], ILSs are generally

complex CPSs. Thus, numerous factors can potentially affect the performance of an ILS or its components. Moreover, many systems rely on the specific deployment, as well as soft- and hardware configurations. This is particularly true for systems operating in a network, as it is common for RTLSs [17]. Consequently, replicating similar conditions in real-world scenarios, other test environments, or even the same environment is typically challenging, limiting repeatability, reproducibility, and transferability.

- **Diversity of applications:** Location data can enhance or enable various applications, as previously discussed in the context of intralogistics (Section 2.1.3). Many more relevant applications exist in other domains, such as locating firefighters in burning buildings, customers in supermarkets, or miners in underground mines [17]. Different applications and application domains entail significantly different influences on system performance. As a consequence, T&E results are difficult to transfer to other application scenarios or to compare with the results obtained from experiments targeting different application domains [13, p. 185].

- **High costs and effort for T&E:** As noted by van Haute *et al.* [18, p.1], substantial spatial, financial, and organizational resources, as well as extensive metrology, statistical, data processing, and application domain expertise are necessary to conduct sophisticated T&E. This is particularly true for infrastructure-based or high-precision ILSs in building-wide testing. To avoid high costs and effort, simpler proof-of-concept tests are often conducted, which do not adequately address the identified key limitations [89].

Recognizing the heterogeneity of technologies, the complexity of systems, the diversity of applications, and the high costs and effort associated with T&E is crucial designing more effective T&E procedures that better address the identified key limitations according to stakeholder interests.

2.2.3 Common Practices

To overcome the presented limitations, different practices have gained popularity, each possessing distinct strengths and weaknesses. The most common practices are briefly presented below.

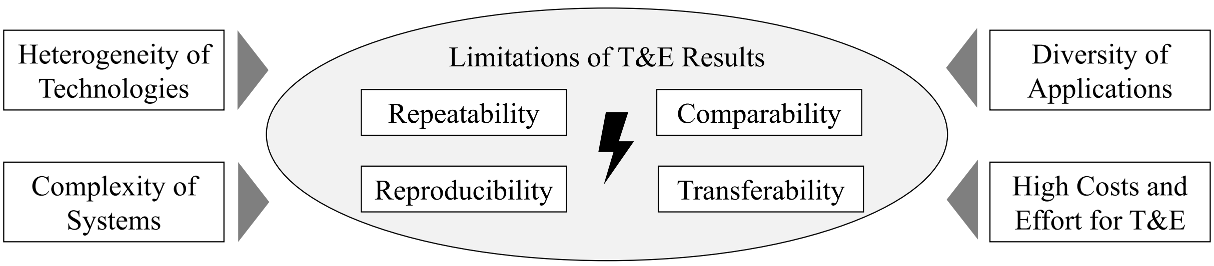


Figure 2.10: Challenges and limitations of T&E

Simulations Simulation models are used to generate test data that provide valuable insights into complex systems, thus offering cost-effective and efficient means to evaluate indoor localization solutions. Simulations are especially beneficial during the design process as they allow for the assessment of system components before the solution is fully implemented. However, since they represent abstractions of the real world, simulations naturally entail limitations. While they offer valuable insights by providing data with high repeatability and reproducibility, their transferability is generally limited, especially for complex CPSs, which are potentially influenced by diverse influencing factors [90, 91].

Common Datasets The success of shared datasets in various research fields has also inspired the indoor localization community to adopt benchmarking approaches based on pre-recorded datasets [92]. Researchers typically collect extensive measurement data from real-world experiments to share them with the community. Numerous scientific publications present datasets for a range of localization technologies, including RSS values for WLAN fingerprinting [93, 94], camera data for vision-based localization [95], or LiDAR, camera, and IMU data for SLAM-based localization [96, 97]. Benchmarking based on shared datasets has proven to be highly beneficial to the research community, as it allows for the investigation of different algorithms with enhanced comparability, based on meticulously collected measurement data. Nevertheless, stakeholders are limited to the available data and T&E cannot be performed at the system level of an ILS.

Indoor Localization Competitions Numerous indoor localization competitions have been held online or offline since the initial IPIN conference in 2010. The *EvAAL Competition*, which took place alongside the IPIN conference in 2011, focused on basic T&E within a common environment [22]. This annual competition has been renamed to *IPIN Competition* in 2014 [98] and now features various evaluation tracks, each dedicated to a particular use case or technology. For example, a track for assessing ILSs in robot localization was introduced in 2016 [99]. The *PDR Challenge in Warehouse Picking* occurred in 2017, followed by the *xDR Challenge for Warehouse Operation* in 2018, both targeting the evaluation of ILSs based on dead reckoning in warehouse settings [100]. Between 2017 and 2018, the *U.S. National Institute of Standards and Technology* (NIST) organized the *PerfLoc Competition*, which focused on evaluating localization algorithms utilizing smartphone sensor data [23, 101]. Furthermore, the *Microsoft Indoor Localization Competition* was first hosted in 2014 in conjunction with the *International Conference on Information Processing in Sensor Networks* (IPSN) [13]. A comprehensive comparison of these competitions is offered by Ichikari *et al.* [100, p. 3]. Although competitions provide increased comparability of results between ILSs tested within the same event, they lack control over data collection and comparability with experiments obtained in other contexts [20].

The presented practices can be used to tackle some of the previously identified key limitations for T&E. They are not mutually exclusive and do not necessarily aim to replace conventional T&E. Instead, the advancement of methodological approaches for T&E and the practices presented partly complement each other.

2.2.4 Methodologies presented in the Literature

To design an effective T&E procedure, several decisions must be made. However, due to the challenges discussed above, making these decisions to generate meaningful outcomes is a complex task. To facilitate systematic T&E, methodologies incorporating concepts, procedures, and metrics can be used. If a methodology becomes established, a multitude of benefits result from the consensus on how to conduct T&E, as previously pinpointed in the introduction of this work.

This section presents the T&E methodologies for ILSs found in the literature, namely the *EvAAL Framework* [20], the *EVARILOS Benchmarking Handbook* [21], and the *ISO/IEC 18305:2016* [17]. For every methodology, the fundamental T&E approaches and main characteristics are discussed while highlighting strengths, and limitations. In Table 2.3, a comparative overview of the T&E approaches and main characteristics is provided.

***EvAAL Framework* [20]** Developed in association with the IPIN competitions, the *EvAAL Framework* provides a set of rules for T&E to generate comparable results under complex and realistic conditions. Based on the taxonomy presented in Section 2.2.1, this framework can be classified as a methodology for non-repeatable, black-box, system-level, building-wide testing with offline surveyed points. Its core criteria mandate the natural movement of the ELT carrying the localization device, a realistic test environment, an appropriate GT accuracy, and the use of the 75th percentile of the Euclidean point error as a performance metric for absolute horizontal position accuracy.

In scientific practice, the *EvAAL Framework*'s simplicity supports the establishment of essential rules that foster more comparable and transferable T&E results. As a result, the framework and associated competitions have contributed significantly to the advancement of indoor localization technology. Often, its core criteria are met without explicit reference to the framework. However, the strengths of the framework also reveal its limitations. Since the rules are few and not strictly defined, the analysis of the influencing factors, the comparability of experiments, and the transferability to real-world scenarios that differ from the test case are limited. The framework was not designed with a focus on the thorough examination of commercial systems but rather for the comparison of prototypical solutions under realistic conditions. For example, considering only the 75th percentile of horizontal position error is insufficient for most industrial applications.

While the *EvAAL Framework* is well-suited to be applied in indoor localization competitions for prototypical solutions, for which it was originally designed, its results may offer limited insight into the behavior and performance of ILSs beyond the specific test case. Nonetheless, this early research contributed significantly to the consideration of the comparability and transferability of T&E results.

***EVARILOS Benchmarking Handbook* [21]** The *EVARILOS Benchmarking Handbook* offers a scheme for designing experiments and evaluating results using multiple metrics. A so-called "benchmarking scenario" is supposed to yield repeatable results for an ILS, comprising of the following.

(a) An "environment specification" describing the building and interference specifications,

Table 2.3: Comparison of existing T&E methodologies based on black-box, system-level testing

	EvAAL Framework **[20]**	***EVARILOS Benchmarking Handbook*** **[21]**	***ISO/IEC 18305:2016*** **[17]**
Repeatable / non-repeatable testing	Non-repeatable testing	Repeatable testing	Non-repeatable testing
Building-wide / laboratory testing	Building-wide testing	Focus on building-wide testing	Building-wide testing
Reference ILS / offline surveyed points	Offline surveyed points	Both	Both
Building specifications	Large space similar to application environment	Building types classified according to material, room number, and size	Classification in one of five building types
Provision of scenarios	None	Not provided. Scenarios consist of the definition of the environment specification, experiment configuration, and benchmark	14 scenarios for the localization of "persons" and "objects" are proposed by describing motion and building types
ELT specifications	Person carrying the localization device	Depending on application. No categorization provided	Categories provided ("person", "object", "robot")
Specification of path and test points	Points along realistic path	Grid-based point sampling, random point sampling, use case specific points	Randomly but uniformly distributed
Motion specification	Natural movement of human actor. Standing still at test points	Natural movement under different velocities to compute sensitivity to changes	Categorization of motion into types (walking, crawling, ...)
Considered performance characteristics	Absolute position accuracy	Room accuracy, absolute position accuracy, latency, energy efficiency, and deployment metrics	Floor/zone accuracy, absolute position accuracy, relative position accuracy, latency, setup time, coverage, location-specific accuracy, availability
Metrics applied for absolute position error	75th percentile of horizontal position error	Mean horizontal and spherical position error	Horizontal, vertical, and spherical error, covariances, Root Mean Squared Error (RMSE), mean error, mean and variance of magnitude
Consideration of influences	Influence by choice of building and path	RF-interference, environment, mobility, and scalability controlled and influence computed as changes from a reference scenario	Challenging conditions regarding the technology of the SuT. Provision of failure modes for localization technologies provided
Considerations of application requirements	None	Consideration for final score calculation	None

(b) An "experiment configuration" detailing the selected evaluation points,

(c) A "benchmark" outlining procedures for obtaining location estimates and performance metrics.

In addition to commonly used metrics for absolute accuracy, the handbook accounts for system latency, energy efficiency, complexity, robustness to radio interferences or environment changes, as well as velocity sensitivity, and the number of localization devices. Robustness is expressed using straightforward statistical measures, such as standard deviation or sensitivity values, by comparing the results of two or more benchmarking scenarios. Subsequently, rules for calculating an ILS's final score for rankings are provided by weighing relevant factors for a considered application. Furthermore, the handbook presents an extensive overview of various methods for sampling evaluation points, discussing different strategies for random and grid-based point sampling.

The *EVARILOS Benchmarking Handbook* presents a sophisticated methodology, striving for repeatable testing, while considering various metrics and influences for RF-based systems. Despite the availability of software tools [18, 19], documented usage of the *EVARILOS Benchmarking Handbook* is limited to the EVARILOS project, and in a preliminary version at the *Microsoft Indoor Localization Competition* in 2014 [102]. Developed primarily for RF-based ILSs, it overlooks the peculiarities of other prevalent technologies. Moreover, by attempting to encompass various application domains, the unique characteristics of the specific domains are neglected.

***ISO/IEC 18305:2016* [17]** Based on the findings of the EVARILOS project, the *ISO/IEC 18305* was published in 2016 to establish a methodology for evaluating generic ILSs focusing on potential system users. This standard, similar to the *EvAAL Framework* and the *EVARILOS Benchmarking Handbook*, emphasizes black-box, system-level testing. However, unlike the *EVARILOS Benchmarking Handbook*, the *ISO/IEC 18305* specifically focuses on non-repeatable, building-wide testing.

Besides providing taxonomies for ILSs and for T&E, the standard provides a set of performance metrics, considerations, scenarios, and reporting requirements. Accordingly, numerous metrics related to the absolute accuracy of an ILS should be reported for absolute accuracy: "means and absolute means of various errors", "covariance", "variance", "RMSE", "circular and vertical error 95 %". Furthermore, "coverage", "set-up time", and "latency" are considered. Regarding latency, the standard distinguishes between "push" and "pull", referring to location data sent continuously or upon request, without specifying the measurement or computation method. Considerations include building types, mobility effects, and failure modes. Five building categories are defined, including underground mines and warehouses. For ELT types, three classes are introduced: "person", "object", and "robot". The "person" type includes various mobility modes, such as backward walking and sidestepping, which are incorporated into 14 proposed test scenarios. Notably, the current version lacks scenarios for the "robot" class. Failure modes address potential pitfalls in typical components of indoor localization technology, guiding testing under challenging conditions. In addition to the calculated performance metrics, the reporting requirements cover numerous aspects describing the ILS, its set-up, and the specific testing conditions.

The publication of the *ISO/IEC 18305* has stimulated scientific discourse. In response, Potortì *et al.* [15] published a critical comment at the IPIN conference 2018, which was reciprocated by Moayeri [103] one of the authors of *ISO/IEC 18305* at the IPIN conference 2021. Potortì *et al.* [15] analyzed the *ISO/IEC 18305* concerning "vocabulary, overview, failure modes", "ELT and building types, scenarios, mobility", and "metrics". For example, Potortì *et al.* criticize that a standard-compliant system test would require an extensive number of test cases. Moayeri replies that it is not mandatory to carry out all the proposed scenarios. Nonetheless, the author acknowledges the possibility of employing fewer scenarios and suggests the revision of the scenarios to alleviate the substantial testing effort. The authors deliberate on the quantity and selection of metrics, the benefits for and focus on users, developers, and testers, as well as the integration of performance requirements. An exhaustive discussion of every aspect of this debate lies beyond the scope of this section. The essential points of discussion are picked up in the appropriate sections as the work progresses. Finally, Moayeri encourages participation in the further development of the *ISO/IEC 18305*.

The *ISO/IEC 18305* embodies a notable effort to establish an international standard for T&E. However, a consensus has not yet been achieved. Its documented usage is limited to the *PerfLoc competition* [23], which was organized by the authors involved in the standard. As highlighted by Bousdar Ahmed *et al.* [14], the *ISO/IEC 18305* produces valuable results for users when the test scenario closely mirrors the intended use case. Nevertheless, for system developers, the benefits are often outweighed by the effort for building-wide testing, whereby the non-repeatable testing approach hinders a systematic investigation of influencing factors.

Various other methodologies exist for component-level testing, such as the *ISO/IEC FDIS 24770-62* to evaluate the performance characteristics of the high rate pulse repetition frequency for UWB [104]. However, component-level testing is not within the scope of this work. Alongside the *ISO/IEC 18305*, the *ISO 18646:2024-2* is another standard that outlines performance criteria and test methods for localization performance, however, focusing exclusively on mobile service robots [105]. This standard details how navigation capabilities should be evaluated, including pose accuracy, repeatability, obstacle detection and avoidance, path deviation, maneuvering through narrow passages, and the precision of mapping functions. Furthermore, *evo* represents a noteworthy software tool that implements various methods for analyzing location data [106]. This open-source *Python* package is dedicated to evaluating location data generated through SLAM. Although *evo* is highly beneficial for evaluating robotics localization systems, its usefulness in producing meaningful T&E results for technologies beyond remains limited. In addition, the development lacks a deep discussion of the underlying methods for performance evaluation, the generation of test scenarios, and application-driven evaluation.

2.3 Identification of Research Questions

The preliminary objective of this dissertation has been defined as the development of a T&E methodology for partially controlled test environments, focusing on the application of ILSs in the context of intralogistics. In the following, open research questions that this work aims to answer are identified, resulting in the delineation of the research objective.

Research Question 1 (RQ1) The first research question involves the formulation of test scenarios that produce meaningful results for stakeholders by employing subsequent benchmarking. Test scenarios should consider influences that are relevant to real-world application scenarios, allowing the transferability of results, while aiming at a high comparability of the results from various experiments. The *ISO/IEC 18305* addresses this conflict by providing a rigid selection of scenarios, and the *EVARILOS Benchmarking Handbook* by focusing on RF-based systems. This work takes a different approach by aiming to provide methods that assist stakeholders in defining application-driven scenarios, with an emphasis on the intralogistics domain. Hence, the first research question is defined as follows.

> How should test scenarios be defined that yield comparable and transferable T&E results in intralogistics contexts?

Research Question 2 (RQ2) The seconds research questions deals with the benchmarking aspect of the T&E methodology. To obtain results that are relevant for different applications in the domain of intralogistics, this focuses on system-level, black-box testing. To highlight that only certain influencing factors can be controlled in dedicated test environments, such as test halls, the term partially controlled test environment has previously been introduced. Unlike building-wide testing, these environments do not cover segments of a real application environment, but instead represent a model on a smaller scale. Such environments are expected to be comparably suitable for T&E in the intralogistics context, as test halls mirror the compartments of typical warehouse and production environments. Furthermore, testing in partially controlled environments could yield repeatable results, which allow systematic analysis of influencing factors and provide a basis for replicability and reproducibility. Hence, this work focuses on testing in partially controlled environments, setting this approach apart from the ones existing in the literature. Considering the aspect of repeatability leads to the following definition of RQ2.

> How should benchmarking of ILSs be conducted for repeatable, black-box, system-level testing in partially controlled test environments?

Research Question 3 (RQ3) Another goal of the methodology to be developed in this work is to assist stakeholders in assessing the suitability of ILSs for specific applications, thus allowing an informed system selection. For successful system selection, a comprehensive methodology must encompass an analytical procedure to specify location data requirements and a matching procedure with system performance determined by benchmarking. However, as highlighted in Section 2.1.4, there is a lack of adequate methods to systematically specify the requirements of location data. Thus, the third research question is formulated as follows.

> How should the suitability of ILSs for an intralogistics application be evaluated based on their localization performance?

Based on the considerations of this discussion, the delineated research objective is defined as follows.

Delineated Research Objective
Design of an application-driven T&E methodology, that provides meaningful results to stakeholders in the context of intralogistics, based on repeatable, black-box, system-level testing within partially controlled test environments.

The following chapter presents the design of the research artifact in alignment with the research objective, providing answers to the research questions at hand.

3 Design of Test and Evaluation Methodology

This chapter presents the design of the *T&E 4Log Framework*, systematically addressing the posed research questions by encapsulating the responses within an overarching methodology. To ensure informed decision-making within a design process a solid foundation is essential. In this work, this foundation is established by the research objective, the practical motivation, and the presented knowledge base. Furthermore, as shown in Figure 3.1, the design process of the framework is guided by a set of design guidelines and the demands of the stakeholders involved.

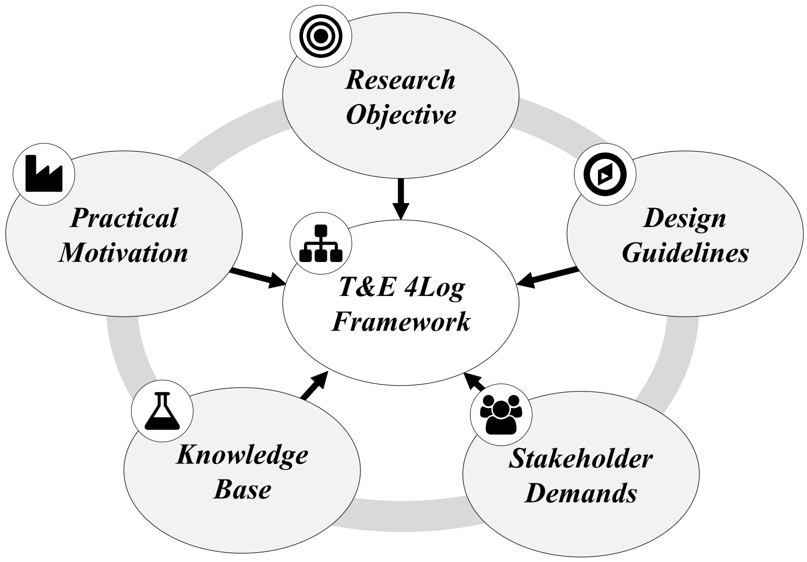


Figure 3.1: Foundational components for the design of the *T&E 4Log Framework*

The design guidelines are presented in the following section, serving as a valuable reference within the design process. Next, a stakeholder analysis is presented in Section 3.2 to derive stakeholder demands. Finally, the chapter delves into a detailed description of the *T&E 4Log Framework* in Section 3.3, starting with a clarification of the scope based on practical considerations and constraints in Section 3.3.1, followed by an overview of its architecture in Section 3.3.2, and an exploration of its components from Section 3.3.3 to 3.3.9.

Earlier versions of the *T&E 4Log Framework* and its components have been presented in previous publications. Initially, a work-in-progress version of the *T&E 4Log Framework* has been presented at the IPIN 2021 [107]. Subsequently, the methodology was refined and generalized, leading to its publication in *MDPI Sensors* [87]. Additionally, an early version of the *Requirement Specification* (Section 3.3.8) was published as a method to systematically specify location data requirements for intralogistics applications [83].

3.1 Design Guidelines

To guide consequent design decisions in the development of the *T&E 4Log Framework*, clear and purposeful design guidelines are crucial. Thus, the following six guidelines are established.

1. **Domain-specificity:** The methodology under development is primarily focused on the intralogistics domain, which encompasses particular environments, influencing factors, and requirement parameters, as well as commonly applied indoor localization technologies. By focusing on a specific application domain, the complexity of a methodology can be significantly reduced while providing a more tailored approach offering increased guidance. Yet, it is worth noting that many of the frameworks' components and the underlying ideas can potentially be generalized and transferred to various other domains.

2. **Application-driven approach:** The practical motivation of the *T&E 4Log Framework* is to help stakeholders make informed decisions about system selection for applications. These applications entail distinct influencing factors and system performance requirements. Thus, the methodology to be developed should permit application-driven T&E, influencing both the testing and evaluation procedure. This is described by Seltzer *et al.* [108] as a hybrid approach of application-driven T&E. The methodology must also accommodate the diversity of relevant applications of indoor localization in intralogistics as presented in Section 2.1.3. However, to increase the practicality and sharpen the focus of the *T&E 4Log Framework*, niche cases in intralogistics are not necessarily covered.

3. **Technology openness:** As long as the application requirements are met, system users are not concerned with the internal workings of a system. Consequently, the *T&E 4Log Framework*, with its application-driven focus, should be flexible to address various types of indoor localization technologies, provided that they are commonly used within the intralogistics domain. An overview of commonly applied technologies has been provided in Section 2.1.2.

4. **Stakeholder focus:** The success of a methodology is significantly dependent on the benefits it offers to the stakeholders involved. By incorporating the demands of different stakeholders, it fosters collaboration and synergies. Thus, stakeholders and their specific demands are of paramount importance for the design of the *T&E 4Log Framework*. An analysis of the stakeholders is provided in the following section.

5. **Modularity:** Given the diverse interests and roles of stakeholders, a modular design of the methodology to be developed is beneficial, allowing adaptations of the T&E process according to the needs of the stakeholders. This will enhance usability and comprehensibility, ensuring guidance and comparability of results while increasing flexibility in their use. Modularity is achieved by providing distinct components with clearly defined interfaces.

6. **Process orientation:** T&E can fundamentally be described as a sequential process following various steps. To optimize its usability and comprehensibility, this should be reflected in the design of the *T&E 4Log Framework*.

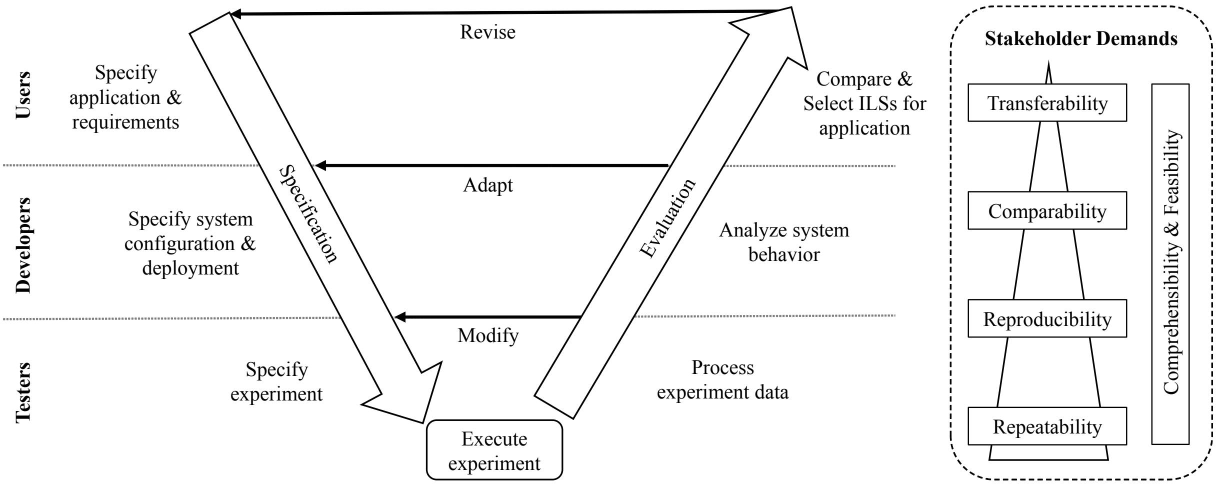


Figure 3.2: The *V-Model for T&E* – illustration of the application-driven T&E process with the involved stakeholders, their functions, and demands

These design guidelines serve as a compass for the development of the *T&E 4Log Framework*, each one emphasizing a critical aspect: domain-specificity reduces the complexity and increases guidance of the methodology for intralogistics, application-driven design ensures practical utility, technology openness provides flexibility regarding system types, stakeholder focus emphasizes the meaningfulness of the results, modularity offers adaptability, and process orientation enhances usability and comprehensibility.

3.2 Stakeholder Analysis

The research objective emphasizes that the methodology to be developed needs to deliver meaningful results to stakeholders within the field of intralogistics. To realize this, it is significant to understand the stakeholders, their function within the T&E process, and their specific demands. Hence, this section analyzes the stakeholders engaged in the process of application-driven, system-level T&E with a focus on the intralogistics domain.

In the context of T&E, the term stakeholders refers to individuals, groups, or organizations engaged in the process or with an interest in the results. For the analysis of stakeholder demands, these are divided into three idealized groups: "users", "developers", and "testers". In practice, the roles of stakeholders often overlap, and their specific interests are naturally case-specific. Nonetheless, an analysis of these idealized stakeholder roles is conducted, structured by the *V-Model for T&E* (Figure 3.2). This model, presented in earlier work [87, 107], has its origins in the V-Model [109] as a frequently employed tool to manage software development and testing. The model has been adapted to describe the process of application-driven T&E, thereby emphasizing the roles and demands of the stakeholders involved at various stages. Every stakeholder contributes to the specification and evaluation aspects of the process. The evaluation results could guide future T&E efforts, as indicated by the horizontal arrows pointing back to a prior phase. In the following, the stakeholder groups are individually discussed. This section closes with a summary of their primary demands.

Users System users are the ultimate beneficiaries of the T&E results, as they are the ones employing the ILSs in real-world application scenarios. In this work, users are further differentiated into end-users and system integrators. System integrators embed ILSs into higher-level products, while end-users, such as warehouse operators, directly employ ILSs, for example, to track goods or assets within the warehouse. They acquire ILSs either directly from the system developer or integrated into a larger product from a system integrator. They are not concerned with each performance metric or the system's performance under conditions apart from their specific application, as long as the requirements of the application are met. Hence, their primary interest lies in the ILSs's suitability for a specific application. However, this necessitates the provision of a clear application description and, ideally, a requirement specification. Users demand that the results generated by T&E are both comparable to others and transferable to real-world scenarios [13]. If no suitable system is identified, the T&E process may need to be repeated, possibly considering other systems or technologies.

Developers System developers, including researchers, are responsible for designing ILSs by developing basic ideas and concepts and transforming them into functional systems for practical applications. The realization of this goal requires the design and examination of prototypes. For developers, T&E is of utmost importance as it enables them to analyze the behavior of the system, leading to a better understanding of the system or its components and, consequently, identifying areas for improvement. To obtain significant results, developers are tasked with specifying the system configuration and deployment while considering real-world applications and requirements provided by system users. This ensures the focus on practically relevant system features. Based on the outcomes of previous experiments, the T&E process can be adjusted to meet the needs of stakeholders. To facilitate significant interpretation of results, developers require comparability with results obtained from other systems, configurations, deployments, and environments. Lastly, developers emphasize the reproducibility of results to validate concepts and findings obtained by T&E endeavors of other stakeholders.

Testers Lastly, testers play a crucial role in the T&E process, since they are responsible for specifying experiments and setting up the test environment to achieve significant results. Testers perform the experiments, process the experiment data, and validate the integrity of the results systematically. To strive for reproducibility, comparability, and transferability of the results, the repeatability of experiments is crucial. If the repeatability or the validity of the results is not satisfactory, testers can modify experiments accordingly.

The stakeholder demands discussed so far, namely repeatability, reproducibility, comparability, and transferability of results, are identical to the limitations of T&E results in scientific research previously outlined in Section 2.2.2. This is not surprising, as a limitation is established by its contribution to the significance for stakeholders. In addition, it is crucial for all stakeholders that the T&E process and its results are comprehensible at every stage. Moreover, feasibility is considered a relevant stakeholder demand, referring to the practicality of conducting the T&E process, considering the required resources, time, and technical limitations.

As discussed previously for the limitations of T&E results in scientific research (Section 2.2.2), the stakeholder demands partially conflict with each other. Feasibility, encompassing cost-related considerations, introduces another significant element to this equation. It is imperative that expenses incurred in the T&E process be justified by its benefits. Therefore, cost-effectiveness is a critical criterion for the design of the T&E methodology.

In summary, the identified stakeholder demands for T&E are repeatability, reproducibility, comparability, transferability, comprehensibility, and feasibility. Understanding the stakeholder roles, recognizing their demands for T&E, and addressing them appropriately is crucial for the design of an effective methodology.

3.3 The T&E 4Log Framework

In this section, the *T&E 4Log Framework* is thoroughly described. To clarify the scope of the framework, the practical considerations and constraints of the *T&E 4Log Framework* are initially elaborated in Section 3.3.1. Ultimately, the *T&E 4Log Framework* is a composition of novel and existing concepts, methods, tools, and guidelines for T&E, integrated into various procedures, each serving a distinct purpose. An overview of the framework's architecture is provided in Section 3.3.2, briefly introducing these procedures and their interrelationships. Sections 3.3.3 to 3.3.9 then delve into a detailed description of each procedure.

3.3.1 Practical Considerations and Constraints

Practical considerations and related constraints significantly influence the applicability of the framework and the significance of the T&E results. Hence, practical considerations and constraints inherent to the *T&E 4Log Framework* are elaborated as follows.

Localization Functions The main purpose of location data is to determine the presence or absence of an entity within a multidimensional interest space. In this work, this is referred to as localization functions. To achieve reliable localization functions, the quality of the location data is crucial. This fundamental aspect forms the foundation for specifying requirements and evaluating the suitability of an ILS for an application. By analyzing the spatial interrelationships of location data parameters to ensure reliable localization functions, concrete guidance is provided for specifying location data requirements.

Location Data Requirements The term location data requirement has been introduced earlier in Section 2.1.4, referring to user requirements related to the location data output of ILSs. This work specifically focuses on location data parameters that are ultimately relevant to achieving reliable localization functions. It is important to note that while they are undoubtedly significant, other user requirements, such as those related to costs or security, are not incorporated in the *T&E 4Log Framework*.

The considered parameters include data output, localization accuracy, update rate (for ILSs with constant update rate), system latency, and availability. Additionally, this work considers localization repeatability, which describes the agreement between location estimates at the same true location [35].

To illustrate the distinction between localization accuracy and localization repeatability, Figure 3.3 provides a visual representation, highlighting that high repeatability can be achieved even with low accuracy. The significance of localization repeatability for ensuring localization functions will be emphasized in Section 3.3.8, which focuses on the specification of location data requirements. For a complete overview, Table 3.1 summarizes the location data requirement parameters that are considered within the *T&E 4Log Framework*.

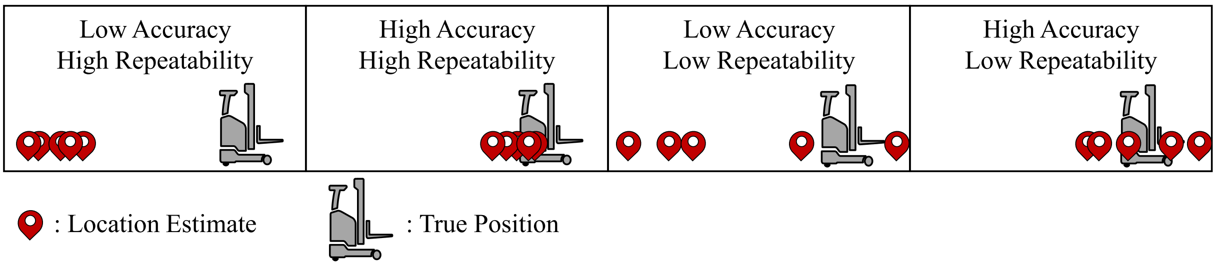


Figure 3.3: Localization accuracy and localization repeatability

It is worth noting that there may exist other requirements related to location data that serve purposes beyond the ones addressed by localization functions. For example, previous work has demonstrated the systematic derivation of requirements for various combinations of update rates and scatter values to facilitate semantic trajectory segmentation using different filter algorithms [4]. However, in such particular cases, the specification of requirements becomes highly dependent on the unique implementation of the data processing algorithms and cannot be meaningfully addressed within a universally applicable methodology. Therefore, the primary focus of the *T&E 4Log Framework* remains on location data parameters that are specifically required for achieving reliable localization functions.

Table 3.1: Overview of location data requirement parameters considered within the *T&E 4Log Framework*

Parameter	**Description**	**Unit**
Data output	DoFs of provided location data	-
Localization accuracy	Level of agreement between the true and estimated location of an entity	[m]
Localization repeatability	Level of conformity among location estimates at the same true position	[m]
Update rate	Frequency of the provision of location estimates	[s^{-1}]
System latency	Time delay between the measurement and provision of location data	[s]
Availability	Proportion of time during which the location data is provided according to the given requirements	-

Absolute Localization The *T&E 4Log Framework* is designed specifically for T&E of systems that provide absolute location data. While various concepts and methods presented within this research may have potential applicability to relative location data, this is beyond the scope of this work. Nevertheless, ILSs designed or employed for relative localization can be evaluated within the framework by establishing a fixed reference frame within the test environment. However, for systems subjected to drift, different procedures and metrics are typically relevant [28].

Indoor Localization Technology Besides limiting the type of location data used, the *T&E 4Log Framework* exhibits a high degree of flexibility in terms of the type of indoor localization technology, following the emphasis of the design guideline on openness of technology. This applies not only to measurement techniques, localization methods, or sensor technologies utilized by the system, but also to other categories introduced. Consequently, the framework is suitable for T&E of device-based, device-free, marker-based, as well as infrastructure-free and infrastructure-based systems. Moreover, it accommodates the positioning and locating capabilities of ILSs. The framework accommodates different types of data provision, including event-driven, trigger-based, or constant update rate, whereby an emphasis has been placed on systems that offer location data with a constant update rate.

Consideration of DoFs The *T&E 4Log Framework* aims to support T&E of ILSs providing location data with up to six DoFs. Although many intralogistics scenarios require fewer DoFs, the framework strives for generic applicability by considering up to six DoFs. This approach avoids the need for isolated solutions for different combinations of DoFs and allows for a consistent methodology across various scenarios. While the mathematical notations presented in the following typically assume six DoFs, they remain valid for lower dimensionality, where missing components can be neglected or assumed to have constant values.

Scope of Evaluation The evaluation of results obtained from different experiments can be approached using various methods, depending on the preferences and needs of stakeholders. For example, previous research has demonstrated the use of decision trees to categorize, analyze, and potentially predict system behavior [110]. Alternatively, sensitivity analyzes can be used to quantify the impacts of different influences on system performance [21]. The *T&E 4Log Framework* recognizes the importance of comparing the results of different experiments, but does not prescribe a specific method for this purpose. Hence, the utilization of the methodology is limited to the evaluation of individual experiments. This intentional design choice provides stakeholders with the freedom to apply the most suitable approach for comparing and analyzing T&E results based on different experiments.

3.3.2 Framework Architecture

The architecture of the *T&E 4Log Framework* aims to achieve a modular and process-oriented design in line with the established design guidelines. To achieve this goal, the structure of the framework consists of different procedures, each possessing a well-defined purpose and explicit output information.

This section provides an overview of the architecture that encompasses the seven procedures of the framework, as represented in Figure 3.4. Each procedure is comprehensively elaborated on in the subsequent sections. However, a brief introduction is provided below. Subsequently, relevant terms are defined, and an overview of the relationship between the framework's components and the posed research questions is presented.

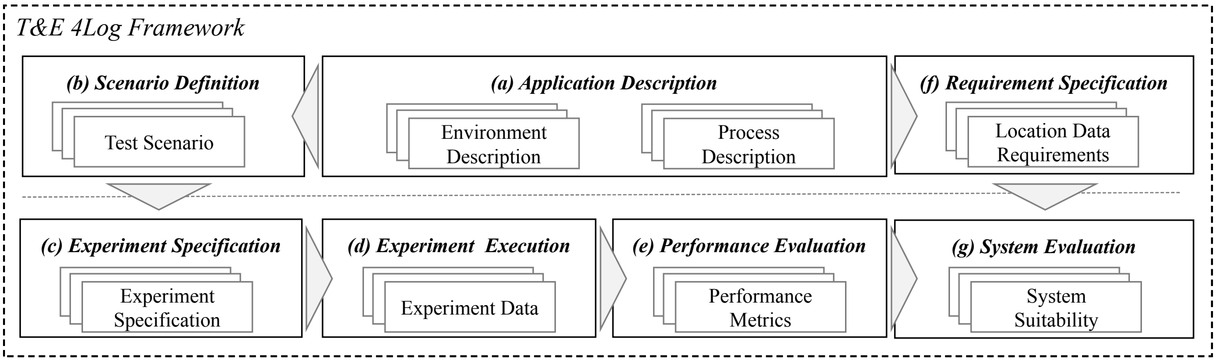


Figure 3.4: Overview of procedures ((*a*) to (*g*)) including output information of the *T&E 4Log Framework*

(*a*) *Application Description*: This initial procedure is crucial for an application-driven T&E process that is aimed at transferability of results. It guides the further process accordingly by outlining the relevant application processes and environment.

(*b*) *Scenario Definition*: Using the process and environment description from the *Application Description*, this procedure provides guidance to systematically define application-driven test scenarios. These scenarios, while independent of the specific test environment, encompass relevant and realistic influencing factors, ultimately supporting the transferability and comparability of performance results.

(*c*) *Experiment Specification*: Taking a test scenario as input, this procedure supports the transformation into a testbed-dependent experiment specification by considering the specific test environment, experiment configuration, and system setup. The comparison of performance metrics from different experiments based on the same experiment specification ultimately yields the experiment's repeatability.

(*d*) *Experiment Execution*: This procedure guides the execution of the experiment according to the specification of the previous procedure, including the experiment setup, resulting in a set of experiment data. In addition, the procedure describes the particularities of a dedicated alignment experiment.

(*e*) *Performance Evaluation*: Processing the experiment data from the *Experiment Execution*, this procedure ultimately leads to the determination of performance metrics, reflecting the relevant location data requirement parameters.

(*f*) *Requirement Specification*: This procedure deals with the specification of location data requirements based on the provided application processes. They are derived from the analysis of localization functions.

(*g*) *System Evaluation*: This final procedure unites the performance metrics from the *Performance Evaluation* and the location data requirements from the *Requirement Specification*, thus allowing data-driven evaluation of the suitability of an ILS for an application.

The combination of the three integral procedures *Experiment Specification*, *Experiment Execution*, and *Performance Evaluation* is referred to as the *T&E 4Log Benchmarking Procedure*. It can potentially be applied without the need to consider the framework's other procedures. This decoupling of the *T&E 4Log Benchmarking Procedure* from the rest of the methodology increases the versatility of the framework by allowing stakeholders to effectively evaluate and compare the performance of different systems or configurations based on freely designed test scenarios.

To achieve their intended outcomes, the presented procedures incorporate a combination of methods, concepts, tools, and guidelines, serving as the fundamental building blocks of the *T&E 4Log Framework*. To ensure a common understanding within the context of this research, the following definitions are provided.

- A **procedure** is a sequence of actions or instructions designed to accomplish a specific outcome
- An **action** refers to a specific operation to contribute to the completion of a procedure within the T&E process
- A **method** is a systematic approach employed to carry out specific actions within a procedure
- A **concept** refers to an abstract idea that can be employed by methods and procedures
- A **tool** is a practical instrument facilitating the T&E process
- A **guideline** is a rule, principle, or piece of advice that provides directional assistance for actions within the T&E process

With the framework architecture clarified, the research questions posed earlier can be associated with the framework procedures. Accordingly, RQ1, dealing with the definition of test scenarios yielding meaningful results in intralogistics contexts, is addressed by the *Application Description* and the *Scenario Definition*. The *T&E 4Log Benchmarking Procedure* provides an answer to RQ2, related to conducting benchmarking of ILSs for repeatable, black-box, system-level testing in partially controlled test environments. Finally, RQ3, which investigates the suitability of ILSs for intralogistics applications, is addressed by the *Requirement Specification* and *System Evaluation* procedures. The relation to the posed research questions is highlighted in the respective section.

In conclusion, the *T&E 4Log Framework* comprises seven procedures incorporate various concepts, methods, tools, and guidelines to describe actions for T&E. The following sections provide a thorough description of the individual procedures, following the order presented in this section.

3.3.3 Application Description

The purpose of the *Application Description* is to provide a comprehensive description of an intralogistics application that involves indoor localization. This is necessary to effectively influence subsequent procedures of the *T&E 4Log Framework* and thus allow for application-driven T&E. Within this work, the application considered is referred to as Application under Consideration (AuC). This can potentially involve a specific real-world application or a hypothetical application that potentially represents a challenging application or a range of related applications. Following the design guidelines presented, a hybrid approach for application-driven T&E is employed according to Seltzer *et al.* [108], which refers to the *Application Description* serving as the foundation for both the *Scenario Definition* and the *Requirement Specification*.

Figure 3.5 presents the methods associated with this procedure. Accordingly, the application processes are initially described. Next, the application environment should be detailed. The methods are elaborated in the following.

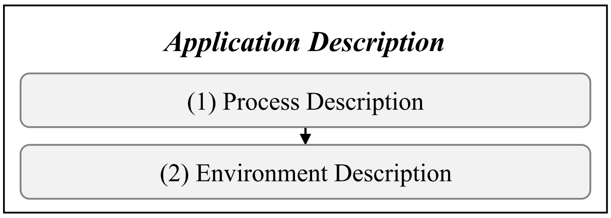


Figure 3.5: Methods of the *Application Description* procedure

(1) Application Processes Description

To provide insights into the AuC and establish a basis for the *Scenario Definition* and *Requirement Specification*, it is imperative to describe the central application processes. This includes describing the involved entities and their primary interactions. For example, in the case of an autonomous forklift truck, key processes might include "driving to shelf", "finding objects", or "pallet pick-up". Naturally, there are different ways to divide and describe these processes. If information from an associated product development or process optimization process is available, many results can be adopted directly or at least utilized to define processes and activities with appropriate granularity.

To describe the application processes linked to localization tasks in a comprehensible manner, concepts from systems engineering can be employed. System engineering is a common approach typically employed for the development of complex systems in large projects [111, pp. 27], whereby processes are described as sequences of activities, that emulate real-world situations. A valuable tool in this context is the activity diagram, a graphical representation of step-wise activities for workflows. Naturally, in practice, the granularity should meet the scope of the T&E endeavor. For a thorough understanding of topics related to systems engineering, the established literature in the field [111] is recommended.

(2) Application Environment Description

Environmental factors have substantial influences on the performance of ILSs. Hence, it is crucial to describe the application environment, serving as a foundation for characterizing relevant influencing factors within the *Scenario Definition*. The application environment encompasses the characteristics of the building, including its basic structure, as well as environmental factors that might influence the localization. As an example, the application environment could be a large warehouse with tall racks and multiple aisles. These factors are important since localization performance is potentially influenced due to factors such as reflection or absorption. Hence, considering these environmental characteristics within the application-driven T&E process is critical to obtain meaningful results.

Additionally, the application environment can be further divided into different spaces, such as “small part storage” or “incoming goods area”, which is reasonable if these spaces possess significantly distinct attributes. In such cases, it is also possible to associate application processes or activities that occur exclusively in specific parts of the environment with the respective spaces.

3.3.4 Scenario Definition

By providing an abstract description of the conditions under which an ILS is tested, the definition of test scenarios plays a crucial role in achieving comparable and transferable T&E results.

Initially, it is briefly recapitulated how the previously presented T&E methodologies address the definition of test scenarios. The EvAAL Framework, while highlighting the significance of conducting “realistic” experiments, lacks detailed information on how realistic scenarios are to be defined. The *ISO/IEC 18305*, on the other hand, offers a set of 14 predefined scenarios focused on the localization of people and objects. These scenarios encompass a description of the entity’s motion and building types for building-wide testing. However, while providing scenarios increases comparability, it severely limits flexibility and, therefore, precludes the definition of scenarios tailored to specific applications. Finally, the *EVARILOS Benchmarking Handbook* presents a flexible approach that allows defining test scenarios suitable for both building-wide and laboratory testing. Nonetheless, it does not provide explicit guidance on how to define comparable and transferable scenarios. The *Scenario Definition* aims to close this research gap, addressed by RQ1. For defining scenarios within this procedure, the following three fundamental principles are introduced.

1. **Independence from testbed:** A test scenario in the *T&E 4Log Framework* is designed to be independent of a specific testbed, encompassing the facility with its infrastructure elements including the deployed GT, allowing it to be applicable across different facilities. Although variations naturally arise in the implementation of a scenario between different facilities and researchers resulting in differences for the T&E results, the methods outlined in the *Scenario Definition* procedure nonetheless serve as a foundation for reproducibility and facilitate a differentiated analysis of results and influencing factors in various studies.

2. **Scenario as a combination of relevant influencing factors:** A test scenario within the *T&E 4Log Framework* is described by conditions that are considered relevant influencing factors. The relevance of influencing factors can be examined from two perspectives. A

factor is relevant if (a) it significantly affects the performance of the considered ILS, and (b) it reflects the conditions of the application scenario considered.

3. **Consideration of application-driven influencing factors:** Influences on ILSs are commonly discussed from a technological point of view, such as signal reflection and the resulting multipath propagation, allowing a systematic analysis of specific influences [112]. However, attempting to quantify the overall "reflectiveness" or similar technical influencing factors of a real-world application scenario is not meaningful. Instead, a more practical approach involves describing the situation, such as the placement of a pallet shelf that leads to obstructions of direct LoS between an emitter and transceiver. Unlike technical influencing factors, this is referred to as application-driven influencing factors. Consequently, a test scenario, encompassing a description of application-driven influencing factors is referred to as an application-driven test scenario.

By leveraging these considerations and the information provided from the *Application Description* this procedure offers a structured approach to systematically identify and characterize relevant application-driven influencing factors. As a result, the *Scenario Definition* facilitates the creation of application-driven test scenarios. The methods associated with this procedure, illustrated in Figure 3.6, are explained as follows. First, an analysis of relevant indoor localization technologies is performed to identify potentially significant application-driven influencing factors. Subsequently, these factors are characterized to ultimately define application-driven test scenarios.

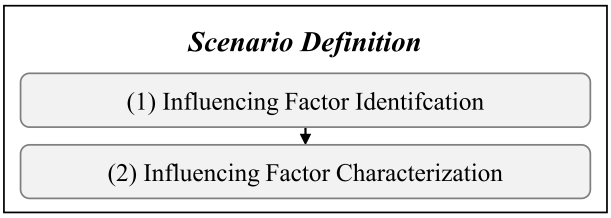


Figure 3.6: Methods of the *Scenario Definition* procedure

(1) Influencing Factor Identification

To identify relevant application-driven influencing factors it is crucial to analyze the indoor localization technology of the ILSs that is to be examined. The discussion of the state of the art in indoor localization technologies in Section 2.1.2 presented a comprehensive taxonomy that classifies indoor localization technologies according to the employed measurement techniques, localization methods, and sensor technologies. For each of these components, the dependencies of technical influencing factors can be linked to application-driven influencing factors. Based on this association, potentially relevant application-driven influencing factors can be determined for an ILS by considering its technological components.

In the following, a reference is provided by associating selected application-driven influencing factors with the previously discussed components of indoor localization technologies. The resulting associations are summarized in Table 3.2. The influencing factors are classified into process influences, environmental influences, and system influences. Process influences refer to significant dependencies of system performance on application processes, while environmental influences relate to dependencies on the application environment. Lastly, system influences are associated with the specific configuration and deployment options of the system. In the following, the application-driven influencing factors for relevant components of indoor localization technologies are discussed, separated by component type.

Table 3.2: Application-driven influencing factors on various components of indoor localization technologie. The bullets mark the potential relevance of an influencing factor on the performance of an ILS, for each technology component utilized

	Application-driven influencing factors											
	Process				Environment					System		
Technology component	ELT	Device/marker/feature loc.	Motion	Path	Static objects	Dynamic entities	Radio interferences	Floor quality	Lighting conditions	Amount of reference nodes	Location of reference nodes	Parameter configuration
Measurement technique												
RSS	●	●		●	●	●					●	●
ToF/ToA	●	●		●	●	●					●	●
TDoA	●	●		●	●	●					●	●
AoA	●	●		●	●	●					●	●
Wheel odometry								●				●
Localization method												
Lateration			●	●						●	●	●
Angulation			●	●						●	●	●
Proximity detection / CoO			●	●						●	●	●
Map matching / fingerprinting			●	●	●	●				●	●	●
Dead reckoning			●	●								●
Sensor technology												
WLAN							●					●
Bluetooth							●					●
RFID							●					●
UWB							●					●
LiDAR	●	●	●	●				●				●
Vision		●							●			●
IMU			●	●				●				●

Influences on Measurement Techniques Firstly, measurement techniques are considered, referring to the methods used to quantify the physical quantities used to determine the location of an entity. Notable examples include RSS, ToF/ToA, TDoA, AoA, and wheel odometry. The accuracy of wheel odometry depends on technical factors, such as wheel slip and deviations between the actual and assumed wheel diameter. These factors depend on application-driven influencing factors, such as floor quality and system configuration, that encompass the assumed diameter of the wheel. In contrast, the efficacy of the other measurement techniques relies on the relative location of emitters and receivers, as well as other technical factors such as obstructions, absorption, reflection, and reduced signal propagation speed. However, in practical applications, the relative location between the emitter and receiver can be described by the path of the ELT, the location of the device or marker on the ELT, and, if existent, the location of the respective reference node. Obstructions, absorption, reflection, and reduced signal propagation speed are furthermore influenced by application-driven influencing factors such as static objects and dynamic entities within the environment of the ELT and by the ELT itself.

Influences on Localization Method Next, the focus is on the introduced localization methods. Localization based on lateration and angulation can be improved for an overconstrained localization problem, which refers to a higher number of reference nodes providing measurements. In addition, the relative location of the nodes to each other and the ELT are crucial aspects, where localization is typically worse if the reference nodes are close to each other and far away from the ELT. Hence, the system's performance is significantly impacted by the number and locations of reference nodes, as well as the path traveled by the ELT. This also holds for the proximity detection/CoO method as well since these parameters reflect the cell size and the reference nodes responsible for localization. The map matching localization method, including fingerprinting, is based on a comparison of sensor data with map data. Certain environments are more suitable than others, whereby an environment is primarily characterized by static objects. In addition, deviations between sensor data and map data can occur due to errors during map recording or changes in the environment after recording. The map data are represented by the amount and location of the reference nodes. Moreover, dead reckoning predicts the entity's location on the basis of an estimation of its past location and kinematic parameters. Hence, the path and motion of the ELT significantly influence dead reckoning algorithms. Additionally, for all localization methods, dependencies on the dynamics of the ELT potentially arise from time stamping and synchronization issues.

Influences on Sensor Technologies The third component of indoor localization technologies according to the introduced taxonomy is the sensor technology. In general, RF-based systems are susceptible to interferences of the radio signal. Due to the typically high accuracy of ILSs based on LiDAR, vibrations of the sensor caused by the type and motion of the ELT and the floor type, as well as the location of the sensor, can be relevant. The same is true for IMUs. Vision-based systems are particularly susceptible to under- and over-exposure of the scene. Furthermore, the type of markers or features and the angle and distance between the marker or feature and the camera are relevant.

Additionally, alongside the already introduced application-driven influencing factors, there typically exist some kind of configuration parameters specific for each of the introduced components. For measurement techniques, examples include the adaptation of the assumed transmission speed for signal propagation or the assumed location of the reference nodes.

Due to the heterogeneity and complexity of indoor localization technology, a multitude of influences can potentially arise. The influences on the performance of an ILS in its entirety can ultimately be selected by attributing the influencing factors to each component of its technology, considering several components within each category when data fusion techniques are used.

To provide an example, the relevant influences of an application of an ILS based on UWB with ToF and lateration, with the additional incorporation of IMU data through dead reckoning should be identified. Using the associations provided in Table 3.2, the application-driven influencing factors relevant to the system's performance are identified as ELT, device location, motion, path, static objects, dynamic entities, radio interferences, quantity and placement of reference nodes, and parameter configurations.

Notably, other components of indoor localization technology and dependencies can also be described by different or additional application-driven influencing factors. Thus, Table 3.2 presents just one of the infinite possible associations of technology components with application-driven influencing factors. Nonetheless, it serves as a tool of reference for stakeholders to facilitate the identification of relevant application-driven influencing factors.

(2) Influencing Factor Characterization

The previous method has focused on the potential significance of application-driven influencing factors regarding the performance of an ILS, thereby providing a template to generate relevant test scenarios. This section, along with the second method of the *Scenario Definition*, involves characterizing the chosen influencing factors to complete the definition of a test scenario. Although it is possible to support this characterization by providing numerical values or ranges related to physical quantities, this information is not mandatory. For instance, lighting conditions can be described by indicating minimum and maximum intensity values or by providing details about the type and approximate density of the lighting sources.

When characterizing application-driven influencing factors for a scenario, the following two primary objectives are taken into account.

(a) A scenario best resembles an application when the application-driven influencing factors ideally align with those of the AuC, resulting in fully transferable T&E results. Although achieving this within a test environment is not feasible in practice, efforts should still be made to achieve a degree of transferability.

(b) A scenario should ideally be independent of a specific test facility to allow comparison of T&E results obtained from different studies. Therefore, application-driven influencing factors must be characterized in a manner that allows their implementation within a partially controlled test environment at reasonable costs.

These two objectives represent a conflict when specifying test scenarios. Ideally, resembling an application requires specifying the influencing factors in full detail. However, fully replicating real-world conditions is not feasible, especially for various independent test environments. Thus, a suitable compromise must be found based on the demands of the stakeholders. If an end user's primary concern is the system's performance for the described application, achieving resemblance is the dominant objective. However, if a system developer aims to compare the system's behavior, maximizing reproducibility and repeatability becomes more important. For example, it can be reasonable to employ a standard robot that is guided by the GT to increase the consistency of the path and the motion, thus reducing the deviations between experiments.

To derive a scenario that resembles the AuC, the process and environment descriptions provided from the *Application Description* procedure are considered to characterize each of the previously identified application-driven influencing factors. Factors categorized as process influences are derived from the process description, while factors categorized as environmental influences are derived by considering the environment description. This process is depicted in Figure 3.7. In addition, the characterization of the system influences depends on the specific ILS and cannot be determined solely by considering the AuC but by considering the installation instructions and recommendations of the system provider. The defined test scenario is the result of this characterization process. Notably, a scenario can be related to an AuC as a whole or a relevant combination between the application process and the environment.

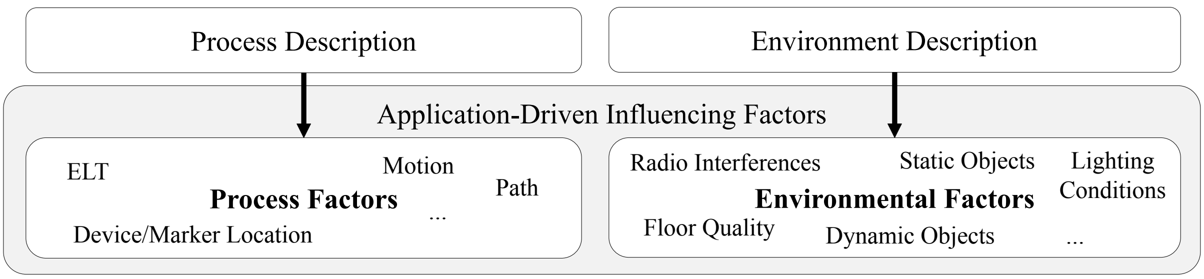


Figure 3.7: Characterizing application-driven influencing factors

The provided test scenarios are subsequently utilized by the *T&E 4Log Benchmarking Procedure*, encompassing the procedures *Experiment Specification*, *Experiment Execution*, and *Performance Evaluation* to ultimately determine performance metrics. These are elaborated on in the forthcoming sections.

3.3.5 Experiment Specification

The purpose of the *Experiment Specification* procedure is to translate a test scenario into a concrete experiment specification for a particular testbed. To enhance the reproducibility and comparability of the results in different studies using common or similar test scenarios, the objective of the *Experiment Specification* is to provide a rigid procedure that produces a detailed description of the experiment to be conducted. This description should provide minimal flexibility to support the repeatability and comprehensibility of experiment results. Consequently, following the resulting experiment specification, an independent researcher should ideally be able to replicate the experiment within the same test facility. As illustrated in Figure 3.8, the methods within the *Experiment Specification* procedure involve (1) the description of the testbed, (2) the specification of the experiment process, environment, and system setup, as well as (3) the specification of evaluation poses. The methods are presented in the following.

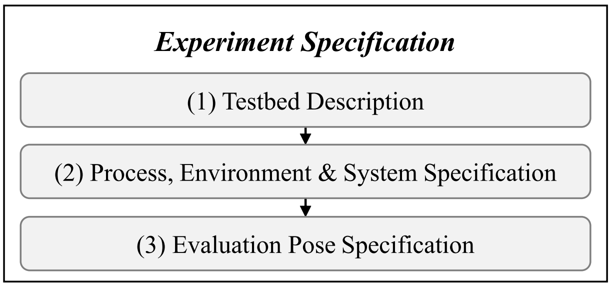


Figure 3.8: Methods of the *Experiment Specification* procedure

(1) Testbed Description

The first step is to describe the testbed encompassing the test environment, test volume, and the provided GT. Each of these components is subject to specific requirements and considerations within the *T&E 4Log Benchmarking Procedure* that are described below.

- **Test Environment:** Denotes the spatial environment in which the test volume is contained. It is described by outlining the dimensions of structural elements, ideally providing a floor plan and photos.

- **Test Volume:** Situated within the test environment, the test volume refers to the space in which experiments will be performed. The test area describes the horizontal footprint of the three-dimensional test volume. It is typically limited by the mobility of the ELT and the GT coverage. As fixed objects within the test volume can significantly influence the performance of ILSs, a free space that can be configured according to the test scenario is preferred. In addition, the size and shape of the test volume limit the path of the ELT during experimentation and thus influence the T&E results. A rectangular shape of the test area or a cuboid shape of the test volume is suggested, significantly larger than the typical room size. Based on practical experiences at the ITL, the minimum dimensions of 6 m x 6 m are suggested for the test area to facilitate the guidance of the ELT, object placement, and evaluation of location-dependent errors.

- **Ground Truth:** The provision of a GT is absolutely crucial for T&E. Only for DoFs that are covered by the GT, the absolute localization accuracy of an ILS can be examined. In addition, the GT-accuracy should be significantly higher than that of the SuT. As suggested by the *ISO/IEC 18305*, the GT-accuracy should be one order of magnitude higher than the one of the SuT, for each relevant DoF. In practice, precise indications of the GT-accuracy are often difficult as T&E of the GT suffers similar challenges to the ones addressed within this research. Nevertheless, testers must be aware of this issue and indicate resulting uncertainties appropriately. Furthermore, utilizing offline surveyed points as GT comes with certain limitations, especially for measuring dynamic parameters, as exact timestamps of the location data are not available. Hence, while many aspects are nonetheless applicable, the *T&E 4Log Benchmarking Procedure* is focused on the usage of a reference ILS as a GT. The description of the GT encompasses the type and name of the reference ILS as well as the DoF and, ideally, accuracy of the location data output.

In conclusion, providing a comprehensive description of the testbed, encompassing the test environment, test volume, and GT, is crucial to ensure the comparability and comprehensibility of T&E results. The subsequent step involves specifying the process, environment, and system setup to provide detailed information to conduct experiments.

(2) Process, Environment, System Specification

The application-driven characterizations of the influencing factors related to process influences, environmental influences, and system influences have been elaborated within the *Scenario Definition* procedure. Now, the focus shifts to translating these characterizations into a detailed specification. This is achieved by considering each relevant application-driven influencing factor in the test scenario and providing the necessary information to implement them. If the test scenario cannot be implemented due to the limitations of available ELTs or controllable environment parameters, the scenario definition might need to be revisited. In the following, the specification of selected application-driven influencing factors is discussed.

First, the process specification is considered. A testbed typically provides various options for ELTs that allow the attachment of localization devices, manual or automated guidance along an experimental path, and the provision of corresponding GT data. The selection of the ELT should align with the test scenario provided. Specifying the selected ELT is necessary, describing the type or model of the ELT and its approximate dimensions. Photos are recommended to indicate the location of any attached localization devices or markers. Within the *T&E 4Log Framework*, a path is partially predetermined by evaluation poses traversed by the ELT in sequential order. The concept of evaluation poses and the suggested approach to select them appropriately is elaborated in the following section. The path and motion can be specified by providing additional information, such as the planned trajectory or velocity profile. For automated test execution, the parameters to define the path and motion can typically be set in configuration files.

Next, the specification of the environment configuration is considered according to the suggested influencing factors. Similarly to the process specification, the environment should be set up according to the provided test scenario. Static objects and dynamic entities are further specified by describing the entities themselves, as well as their locations or movements within the test environment during experimentation. It is suggested to indicate the location of static objects within floor plans.

Finally, the system specification includes providing information about the type and model of the system, as well as the placement of relevant hardware components and the configuration of software parameters. The locations of the components can be indicated alongside the ones of static objects within the floor plans of the test environment.

The resulting specification of the process, environment, and system configuration should ideally enable researchers to repeat the same experiment using the same testbed and equipment, thereby obtaining repeatable experiment results.

(3) Evaluation Pose Specification

Within the existing T&E methodologies discussed in Section 2.2.4, the evaluation of the accuracy of the system is based on measurements of the GT and the SuT at specified evaluation points.

The application of predetermined evaluation points, contrary to utilizing the full location data provided by the SuT offers several benefits.

1. Increased repeatability, as the path may vary between experiment executions while evaluation points remain consistent

2. Enhanced comparability, as the influence of different update rates among ILSs is minimized

3. Systematic analysis of location-dependent accuracy and localization repeatability, as the locations of the relevant location estimates can be controlled

However, using only the measurements corresponding to the evaluation points reduces the number of total measurements used, leading to an increase in the uncertainty of the results. Poorter *et al.* [113] examined the relationship between the number of evaluation points and the resulting uncertainty and found that beyond a certain threshold, the increase in uncertainty is of minor relevance. Based on this study, the *T&E 4Log Benchmarking Procedure* suggests a minimum number of 30 evaluation points, as a compromise between testing effort and uncertainty of results. Therefore, the issue of increased uncertainty is mitigated even when a subset of the total measurements provided by the SuT is used.

In contrast to existing methodologies, the *T&E 4Log Benchmarking Procedure* introduces evaluation poses that incorporate orientation components in addition to the position. This is crucial since orientation components can significantly affect localization. A set of evaluation poses E_{poses} for an experiment, encompassing six DoFs, can be described as follows, whereby l represents the total number of evaluation poses:

$$E_{poses} = [x_k, y_k, z_k, \phi_k, \theta_k, \psi_k]_{k=1}^{l} \tag{3.1}$$

The coordinates are to be defined within the global coordinate frame of the GT, which is independent of the SuT and therefore serves as the reference coordinate frame for the testbed.

A straightforward grid-based sampling approach is employed to establish consistency in determining evaluation poses, similar to the one described in the *EVARILOS Benchmarking Handbook* [21]. Evaluation poses are sampled on a grid with a certain grid size g. In addition, a fixed set of orientation options, such as multiples of 90°, should be utilized. Tolerance values are defined as the minimum proximity to reach an evaluation pose. Setting the position tolerance b to equal or less than half the grid size is recommended to avoid overlapping tolerance spaces.

For grid-based sampling, the selection of evaluation poses must align with the specified path. For example, if the focus of the T&E effort is on horizontal movements, this should ideally be reflected in the evaluation poses. However, it must be taken into account that the choice of evaluation pose is limited by the mobility of the ELT and the placement of static objects within the test volume. Figure 3.9 illustrates exemplary results of grid-based pose sampling with six evaluation poses, each containing three DoFs (x, y, ψ), on a rectangular test area.

Furthermore, computing performance metrics related to localization repeatability within the *T&E 4Log Benchmarking Procedure* requires approaching the same evaluation pose at least twice for each experiment. The process of computing localization repeatability will be discussed in detail as part of the *Performance Evaluation* procedure. Nevertheless, this condition must already be considered.

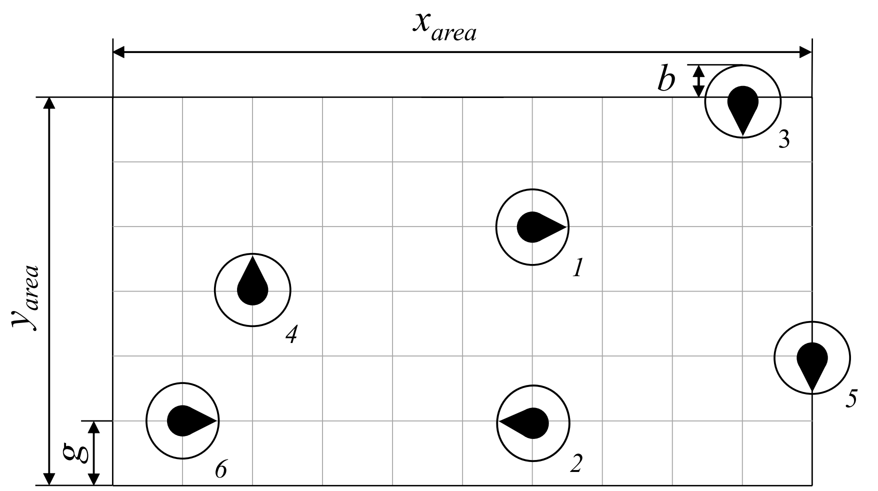


Figure 3.9: Evaluation poses $k \in [1, 6]$ with position tolerance b, grid size g, and arrows indicating the heading direction of each pose as a result of grid-based pose sampling

3.3.6 Experiment Execution

The *Experiment Execution* procedure aims to generate repeatable experiment data according to the experiment specification, consisting of three primary methods, as shown in Figure 3.10.

Firstly, the experiment is set up as specified. Next, the experiment data are recorded; the ELT is moved through the test volume, traversing the specified evaluation poses. Finally, an experiment is performed to record alignment data, which is necessary to determine the alignment of the associated coordinate frames of the SuT and the GT in the subsequent procedure.

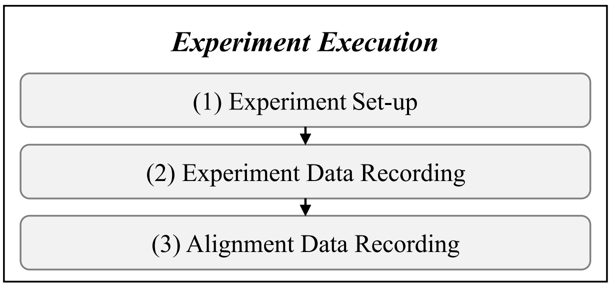


Figure 3.10: Methods of the *Experiment Execution* procedure

(1) Experiment Set-up

Setting up an experiment involves implementing the experiment specification. The test environment is configured accordingly, and, if necessary, the configuration is verified by taking measurements of the environment parameters. The SuT is deployed and configured to match the system specifications. For example, for RF-based ILSs, this includes placing reference nodes, while for LiDAR-based ILSs employing map matching, a contour map of the environment must be recorded. Additional information on the system setup can be provided. Furthermore, internal system calibrations may be required. In addition, the ELT is set up to enable localization by the SuT and the GT.

Finally, accurate synchronization of the clocks of the SuT and the reference system is necessary to allow the determination of the system latency and establish the correspondence between location and reference data. Synchronization errors, represented as $t_{err,sync}$, should be kept to a minimum. For this purpose, the utilization of *Precision Time Protocol* (PTP) is recommended [114]. By multiplying $t_{err,sync}$ with the maximum velocity $\vec{v}_{max}$ of the ELT during the experiment, an estimation of the resulting measurement error can be obtained.

(2) Experiment Data Recording

The experiment is then executed by recording the location data from the SuT and the reference system as the ELT traverses the evaluation poses. The data collected from the experiment includes time-stamped location data from the SuT and the GT. In addition to the timestamps associated with each SuT measurement, a second timestamp t_{lat} is recorded at the moment of data provision, which is necessary to determine system latency in the *Performance Evaluation* procedure.

The location data acquired from the SuT, denoted as E_{SuT}, can be represented for six DoFs as follows:

$$E_{SuT} = [t_{i,SuT}, t_{i,lat}, [x_{i,SuT}, y_{i,SuT}, z_{i,SuT}, \phi_{i,SuT}, \theta_{i,SuT}, \psi_{i,SuT}]]_{i=1}^{n} \tag{3.2}$$

Here, n corresponds to the number of measurements. The timestamp $t_{i,SuT}$ is provided by the SuT, and $t_{i,lat}$ represents the moment of data provision.

Similarly, the data acquired from the GT, denoted as E_{GT} can be described as follows, with m representing the total number of GT measurements:

$$E_{GT} = [t_{j,GT}, [x_{j,GT}, y_{j,GT}, z_{j,GT}, \phi_{j,GT}, \theta_{j,GT}, \psi_{j,GT}]]_{j=1}^{m} \tag{3.3}$$

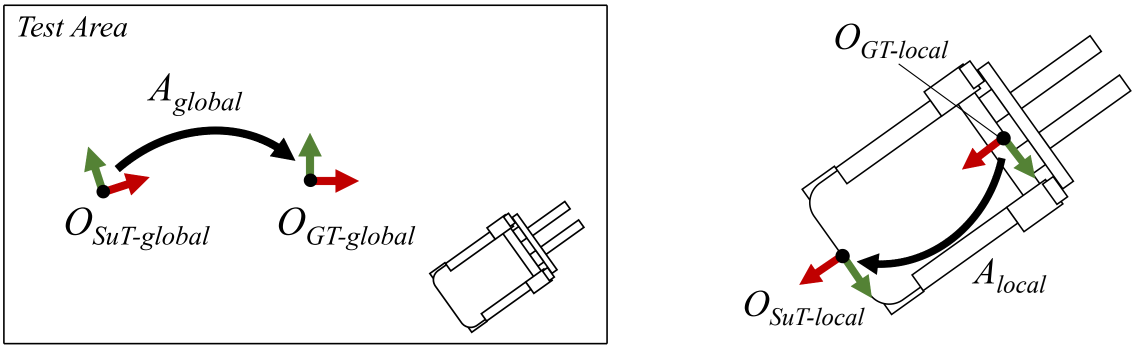
The number m and timestamps $t_{j,GT}$ of the measurements from the GT are typically distinct from the number n with the respective timestamps of the SuT. In the *Performance Evaluation* procedure, the correspondence is achieved through the timestamp and spatial alignment.

When testing different ILSs within similar scenarios, it is beneficial to record the location data from multiple systems during the same experiment to increase comparability between the results. Furthermore, in instances of automated ELT guidance, it is optimal for the robot to obtain location data for navigation from the reference system, thereby ensuring high accuracy when approaching the evaluation poses and increasing the repeatability of the experimental path.

(3) Alignment Data Recording

Finally, within this method, the focus is on recording the data necessary to align the coordinate frames associated with the SuT and the GT. This involves considering both the global alignment, which represents the transformation between the global coordinate frames of the SuT ($O_{SuT-global}$) and the GT ($O_{GT-global}$), and the local alignment, which pertains to the transformation of the local coordinate frames associated with the ELT ($O_{SuT-local}$ and $O_{GT-local}$). Local coordinate frames are typically established by using attached devices or markers or by using specific features of the ELT. The alignments do not have to be determined for each experiment but for each new configuration of the coordinate frames.

The global and local coordinate frames are illustrated in Figure 3.11 along with their transformations. It is crucial to account for deviations that arise from arbitrary or incorrectly aligned coordinate frames, as these significantly affect the quality of the performance evaluation [115, p. 67]. Manual determination and correction of alignments are prone to errors, particularly when dealing with orientation components and high-accuracy ILSs. To overcome these challenges, the *T&E 4Log Benchmarking Procedure* proposes a data-driven approach based on a dedicated alignment experiment. In this alignment experiment, the focus is on recording the alignment data under specific conditions, which are described in this procedure. The subsequent *Performance Evaluation* procedure will delve into the determination and use of global and local alignments, where the alignment results will be applied to the previously recorded experiment data E_{SuT} and E_{GT}. The current method deals with recording the alignment data.



(a) (b)

Figure 3.11: Coordinate frames and alignments for (a) global alignment and (b) local alignment

An alignment experiment is specifically designed to mitigate biases in location data, allowing for the isolation of localization errors resulting from global and local misalignments. Hence, an alignment experiment implements different measures to diminish the following systematic error components.

- **Position-dependent errors:** Evaluation poses should be evenly distributed around the center of the test volume to reduce biases in location estimates within specific regions
- **Orientation-dependent errors:** Evaluation poses should be approached from opposing directions, effectively averaging out orientation-dependent errors and mitigating systematic influences of positional misalignments between the coordinate frames associated with the ELT
- **Velocity-dependent errors:** Localization errors should be determined when the ELT is stationary at each evaluation pose, minimizing the influence of velocity on individual location estimates
- **Errors due to environmental influencing factors:** To minimize the impact of significant static objects, dynamic entities, radio interferences, or challenging lighting conditions, they should be avoided during the alignment experiment

Similarly to the experiment data recorded before, the alignment data comprises the recorded and timestamped location data provided by the SuT ($E_{SuT,align}$) and the GT ($E_{GT,align}$). By implementing the measures presented, systematic errors in alignment data can be effectively mitigated, resulting in a more accurate and reliable determination of alignment transformations within the *Performance Evaluation* procedure.

3.3.7 Performance Evaluation

The objective of the *Performance Evaluation* procedure is to generate performance metrics based on the provided experiment data. These metrics facilitate the systematic analysis of an ILS's performance and the evaluation of the system's suitability within the *System Evaluation* procedure. Consequently, performance metrics must reflect the location data requirements presented specified within the *Requirement Specification* procedure.

Similar to the experiment results, the performance metrics correspond to a specific test scenario and are associated with an experiment specification. This correspondence allows for a systematic analysis of the factors influencing system performance. However, as previously noted, the analysis of these influencing factors falls outside of the scope of the *T&E 4Log Framework*.

The *Performance Evaluation* procedure consists of three primary methods illustrated in Figure 3.12. Firstly, the experiment data undergo two computational processing steps to achieve timestamp alignment and association with evaluation poses. Subsequently, alignment transformations are computed and applied to the experiment data, resulting in two sets of evaluation data. Finally, the evaluation data are utilized to determine the performance metrics.

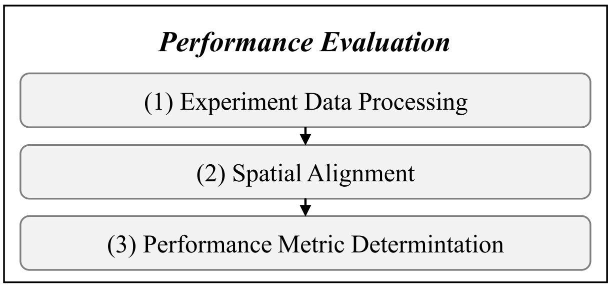


Figure 3.12: Methods of the *Performance Evaluation* procedure

(1) Experiment Data Processing

The experiment data and the alignment data from the previous procedure consist of GT data and SuT data provided by the respective systems. It is necessary to perform initial processing steps to utilize these data to determine alignment transformations and performance metrics with the subsequent methods. These involve the timestamp alignment to establish the correspondence of the GT data (E_{GT}, $E_{GT,align}$) with the respective SuT data (E_{SuT}, $E_{SuT,align}$). Next, the data must be associated with the evaluation poses (E_{poses}) specified in the *Experiment Specification* procedure. These processing steps are elaborated on below.

Timestamp Alignment As mentioned above, the timestamps of the GT and SuT data points do typically not match. To find corresponding data points, the simplest approach is to select the GT data point $E_{j,GT}$ with the timestamp $t_{j,GT}$ closest to the timestamp of the SuT data $t_{i,SuT}$. However, this approach may induce localization errors depending on the time difference between consecutive GT data points and the ELT's velocity. Depending on the accuracy of the SuT, this error may or may not be tolerated.

To achieve more accurate results, interpolation techniques can be employed to the GT data. Linear interpolation is typically sufficient, but more advanced techniques can be employed if considered necessary. The following equation demonstrates linear interpolation for the x-coordinate of the GT data:

$$x_{i,GT} = x_{j,GT} + (t_{i,SuT} - t_{j,GT}) \cdot \frac{x_{j+1,GT} - x_{j,GT}}{t_{j+1,GT} - t_{j,GT}} \tag{3.4}$$

For successful application of the equation to the rotational components, due to their circular nature, it must be ensured that $\phi, \theta, \psi \in] - \pi, \pi]$.

Merging of the corresponding data points leads to the timestamp-aligned experiment data $E_{SuT,GT}$:

$$\begin{aligned} E_{SuT,GT} = [t_{i,SuT}, [&x_{i,SuT}, y_{i,SuT}, z_{i,SuT}, \phi_{i,SuT}, \theta_{i,SuT}, \psi_{i,SuT}], \\ &[x_{i,GT}, y_{i,GT}, z_{i,GT}, \phi_{i,GT}, \theta_{i,GT}, \psi_{i,GT}]]_{i=1}^{m} \end{aligned} \tag{3.5}$$

The timestamp alignment must also be applied for the latency timestamps ($t_{i,lat}$) of the SuT data and for the alignment data ($E_{GT,align}$), resulting in a second ($E_{SuT,GT,lat}$) and third ($E_{SuT,GT,align}$) set of timestamp-aligned experiment data.

Evaluation Pose Association In this step, the three sets of timestamp-aligned data points are associated with the evaluation poses. Figure 3.13 exemplifies the horizontal positions of the

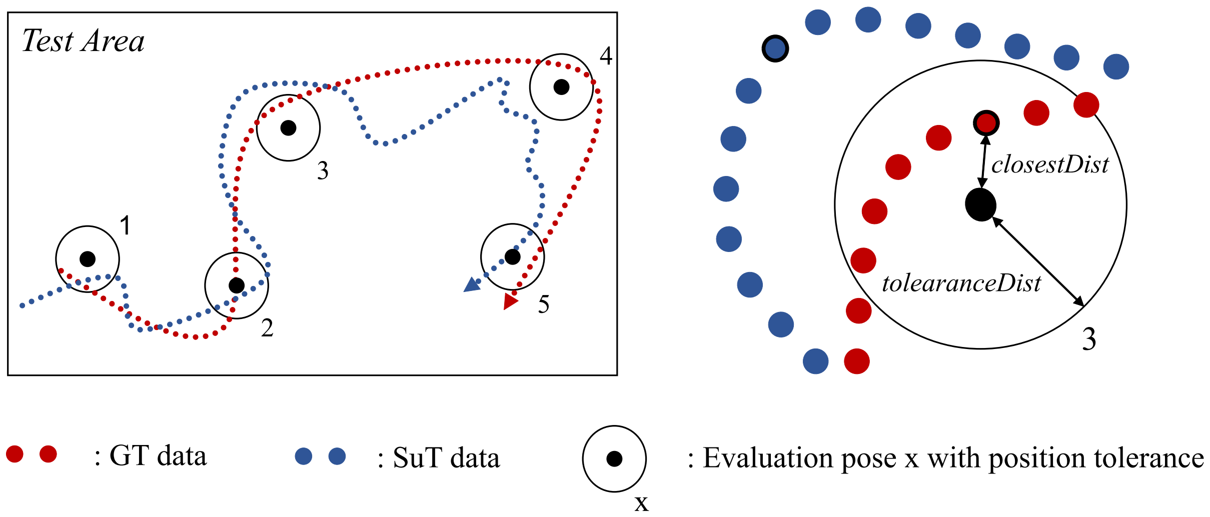


Figure 3.13: Timestamp-aligned GT and SuT data for evaluation pose association

evaluation poses and the timestamp-aligned GT and SuT data. The left side presents a top view of the test area, depicting the evaluation poses ($k \in [1, 5]$) for both the GT and the SuT data. On the right, a closer look at evaluation pose $k = 3$ reveals the horizontal position tolerance, referred to as *toleranceDist*. It also highlights the nearest GT data point to the evaluation pose, and the corresponding SuT data point, which represents the data point ultimately relevant to the determination of the accuracy metrics.

To associate the evaluation poses with the data points, corresponding to the smallest distance between the GT, under the conditions of maintaining the order of the poses and entering the tolerance space, the *ClosestPoseAssociation* algorithm is proposed (Algorithm 1). It takes the list of evaluation poses (*evaluationPoses*), the list of timestamp-aligned experiment data (*experimentData*), and a potentially multi-dimensional tolerance distance (*toleranceDist*) as inputs and returns a filtered list of experiment data (*experimentDataFiltered*), where each evaluation pose is associated with a data point. The algorithm comprises the following key operations, with the numbers in brackets indicating the corresponding lines within the provided pseudocode.

1. **Initialization (1-2):** Sets up the empty list *experimentDataFiltered* to store the filtered data. It also prepares *idxEnterToleranceList* to track the indices where the tolerance of each evaluation pose is met.

2. **Tolerance Entry Points Identification (3-15):** The algorithm iterates over each pose in *evaluationPoses* to identify the index in *experimentData* where the tolerance conditions (position and angular) for each pose are first met since having entered the previous evaluation pose. These indices are stored in *idxEnterToleranceList*. This step ensures that searching for the closest data point is limited to the relevant data segment for each evaluation pose.

3. **Closest Data Point Search (16-28):** The algorithm searches for the closest data point within its corresponding data segment for each evaluation pose. The segments are defined by the indices in *idxEnterToleranceList*. Within each segment, the algorithm computes the distance between the evaluation pose and each data point of the segment. If a data point is within the defined tolerances (position and angular) and is the closest found so far, it updates *closestDist* and *closestData*. This segmented approach ensures that the chronological order is respected and that the closest match is found within the context of each evaluation pose's segment.

4. **Post-Processing (29-31):** Appends the *closestData* to *experimentDataFiltered* for each evaluation pose if a suitable data point is found within its segment. This process is repeated for each evaluation pose, with the algorithm progressing through the evaluation poses and their corresponding segments, ensuring that all evaluation poses are processed and associated with the most appropriate data points.

If the tolerance space of an evaluation pose is not entered by the timestamp-aligned GT data, the result is considered invalid for the given configuration. Hence, either the tolerance distance should be adjusted, the list of evaluation poses should be revised, or the experiment execution should be repeated.

Algorithm 1 *ClosestPoseAssociation*

Require: List of evaluation poses (E_{poses}): *evaluationPoses*, List of timestamp-aligned experiment data ($E_{SuT,GT}$): *experimentData*, Position tolerance: *positionTolerance*, Angular tolerance: *angularTolerance*

Ensure: Filtered list of experiment data associated with evaluation poses ($\overline{E}_{SuT,GT}$): *experimentDataFiltered*

```
Initialize experimentDataFiltered as empty list
Initialize idxEnterToleranceList as empty list to store index where tolerance is first entered
startIndex ← 0
for evalIndex from 0 to size(evaluationPoses) - 1 do
    toleranceEntered ← False
    while startIndex < size(experimentData) and not toleranceEntered do
        Compute distance and angular distance between evaluationPoses[evalIndex] and experimentData[startIndex]
        if distance ≤ positionTolerance and angular distance ≤ angularTolerance then
            Add startIndex to idxEnterToleranceList
            toleranceEntered ← True
        else
            startIndex ← startIndex + 1
        end if
    end while
end for
for evalIndex from 0 to size(idxEnterToleranceList) - 1 do
    Set searchStartIndex to idxEnterToleranceList[evalIndex]
    Set searchEndIndex to idxEnterToleranceList[evalIndex + 1] or size(experimentData) if evalIndex is last
    Initialize closestDist as infinity and closestData as None
    for dataIndex from searchStartIndex to searchEndIndex - 1 do
        Compute distance and angular distance between evaluationPoses[evalIndex] and experimentData[dataIndex]
        if distance ≤ positionTolerance and angular distance ≤ angularTolerance then
            if distance < closestDist then
                Update closestDist as distance
                Update closestData as experimentData[dataIndex]
            end if
        end if
    end for
    if closestData is not None then
        Append closestData to experimentDataFiltered
    end if
end for
    return experimentDataFiltered
```

In addition, thresholds for the velocity can be incorporated according to the experiment specification. If static measurements are required, this threshold can be set to 0 or close to 0. The algorithm output is a filtered list of timestamp-aligned experiment data associated with the specified evaluation poses, denoted $\overline{E}_{SuT,GT}$. It can be expressed as follows:

$$\begin{aligned}\overline{E}_{SuT,GT} = [k, t_{k,SuT}, [&x_{k,SuT}, y_{k,SuT}, z_{k,SuT}, \phi_{k,SuT}, \theta_{k,SuT}, \psi_{k,SuT}], \\ &[x_{k,GT}, y_{k,GT}, z_{k,GT}, \phi_{k,GT}, \theta_{k,GT}, \psi_{k,GT}]]_{k=1}^{l}\end{aligned} \tag{3.6}$$

Likewise, the algorithm is applied to the timestamp-aligned experiment data associated with the latency timestamp and to the alignment data. As a result, three lists of timestamp-aligned data sets associated with the evaluation poses E_{poses} are obtained: $\overline{E}_{SuT,GT}$, $\overline{E}_{SuT,GT,lat}$, and $\overline{E}_{SuT,GT,align}$.

(2) Spatial Alignment

This method involves the determination and application of spatial alignment transformations. The alignment transformations are derived from the data list $\overline{E}_{SuT,GT,align}$, which is the result of processing both GT and SuT data collected during the alignment experiment. This method involves (a) the determination and application of the global alignment transformation, followed by (b) the determination and application of the local alignment transformation. To facilitate understanding and introduce relevant notations for both steps, the transformation tree depicted in Figure 3.14 is considered. The foundations of coordinate transformations are thoroughly elaborated by Craig [116, pp. 19].

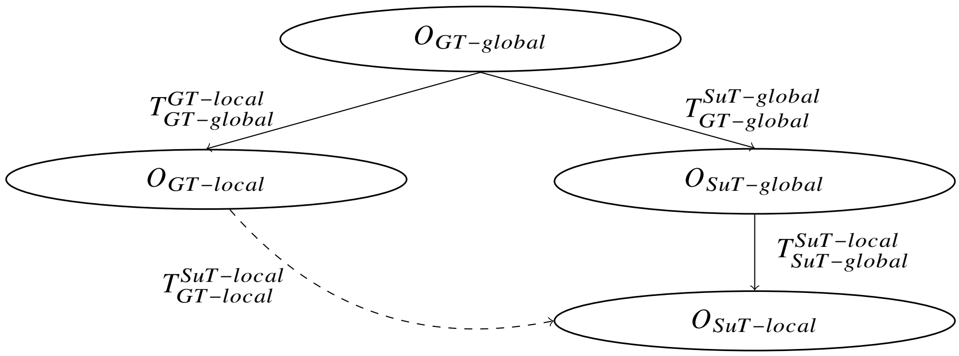


Figure 3.14: Transformation tree describing spatial relationships between coordinate frames via transformations

The transformation tree illustrates the spatial relationships between the coordinate frames, relevant for T&E in the context of this work. As previously described, $O_{GT-global}$ and $O_{SuT-global}$ represent the global coordinate frames of the GT and SuT, respectively, while $O_{GT-local}$ and $O_{SuT-local}$ refer to the local coordinate frames. An illustrative example has been previously provided in Figure 3.11.

Spatial relationships are described by transformations T_x^y, referring to the transformation of locations from coordinate frame O_x to O_y. Transformations are typically represented as 4×4 (or 3×3 for 2D) homogeneous matrices, consisting of a rotation matrix R_x^y and a translation vector $\vec{t}_x^{\,y}$:

$$T_x^y = \begin{bmatrix} R_x^y & \vec{t}_x^y \\ 0 & 1 \end{bmatrix} \tag{3.7}$$

Transformation matrices can be employed to describe the location of a coordinate frame. Therefore, each data point in the experiment data can be represented by a transformation matrix. For example, based on Equation 3.6, $\overline{E}_{SuT,GT}$ can also be expressed as follows:

$$\overline{E}_{SuT,GT} = [k, t_{k,SuT}, T_{E,k,SuT}, T_{E,k,GT}]_{k=1}^{l} \tag{3.8}$$

Whereby, $T_{E,k,SuT}$ represents the location $[x_{k,SuT}, y_{k,SuT}, z_{k,SuT}, \phi_{k,SuT}, \theta_{k,SuT}, \psi_{k,SuT}]$, and $T_{E,k,GT}$ represents the location $[x_{k,GT}, y_{k,GT}, z_{k,GT}, \phi_{k,GT}, \theta_{k,GT}, \psi_{k,GT}]$.

The transformations depicted in the presented transformation tree are described below.

- $T_{GT-global}^{GT-local}$ refers to the location of the GT device/marker/feature within the global coordinate frame, derived directly from the output of location data from the reference system
- $T_{GT-global}^{SuT-global}$ pertains to the deviations between the global GT and SuT frames. It is given by the global alignment transformation A_{global}
- $T_{SuT-global}^{SuT-local}$ signifies the location of the SuT device/marker/feature within the global coordinate frame of the SuT. The location data provided by the SuT estimates this transformation but deviates due to inherent measurement errors of the SuT
- $T_{GT-local}^{SuT-local}$ refers to the location of the local SuT coordinate frame within the local GT frame. The transformation is given by the local alignment A_{local}

Based on these considerations, the determination and application of the global and local alignments A_{global} and A_{local} are elaborated in the following.

Global Alignment Figure 3.15 illustrates the exemplary horizontal positions of the alignment data associated with the evaluation poses ($\overline{E}_{SuT,GT}$). The Umeyama-technique (without scaling) [117] can be utilized to determine the optimal global alignment transformation A_{global} based on the data obtained from the alignment experiment. It does this by minimizing the mean squared error between the two sets of associated data points. This ultimately results in the global alignment transformation (A_{global}) based on the data obtained from the alignment experiment.

To apply the global alignment transformation, a fundamental characteristic of transformation matrices is employed, which allows them to be multiplied to represent chains of transformations, as follows:

$$T_1^3 = T_2^3 \cdot T_1^2 \tag{3.9}$$

Therefore, to express the location data provided by the SuT within the global GT coordinate frame $O_{GT-global}$, A_{global} must be applied. For example, when applying the global alignment

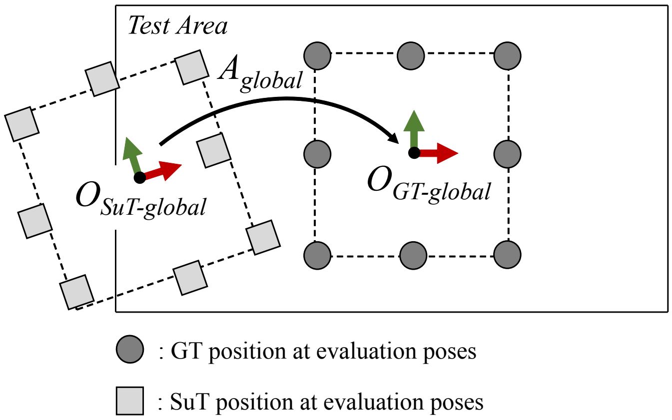


Figure 3.15: Umeyama-technique for determining the global alignment transformation A_{global} [117]

to the location data of the SuT, expressed as a transformation matrix, the following equation can be used to determine the spatially aligned experiment data ($\widehat{E}_{SuT}$):

$$\widehat{E}_{SuT} = [t_{i,SuT}, t_{i,lat}, T_{E,i,SuT} \cdot A_{global}]_{i=1}^{n} \tag{3.10}$$

Local Alignment After determining the global alignment transformation, the local alignment A_{local} can be determined on the basis of the same measurement data captured in the alignment experiment. Referring to the transformation tree in Figure 3.14, following the two chains of transformations leads to the following equation:

$$T_{GT-local}^{SuT-local} \cdot T_{GT-global}^{GT-local} = T_{SuT-global}^{SuT-local} \cdot T_{GT-global}^{SuT-global} \tag{3.11}$$

Using the notation of the alignment matrices A_{local} and A_{global}, it follows:

$$A_{local} \cdot T_{GT-global}^{GT-local} = T_{SuT-global}^{SuT-local} \cdot A_{global} \tag{3.12}$$

At each evaluation pose $k \in [1, l]$, a location provided by the GT can be represented as the transformation $T_{GT-global,k}^{GT-local}$. Furthermore, the location provided by the SuT estimates the transformation $T_{SuT-global}^{SuT-local}$. Hence, it follows:

$$[A_{local,k} \cdot T_{GT-global,k}^{GT-local}]_{k=1}^{l} = [T_{SuT-global}^{SuT-local} \cdot A_{global,k}]_{k=1}^{l} \tag{3.13}$$

Since A_{global} represents a known static transformation, an estimate of the local alignment can be obtained by solving for $A_{local,k}$:

$$[A_{local,k}]_{k=1}^{l} = [T_{SuT-global}^{SuT-local} \cdot A_{global,k} \cdot (T_{GT-global,k}^{GT-local})^{-1}]_{k=1}^{l} \tag{3.14}$$

As both A_{local} and A_{global} represent static transformations, A_{local} can be computed by expressing each matrix $A_{local,k}$ as a pose, minimizing the mean square error between the poses, and finally converting the result back into a transformation matrix.

Errors in the alignment data will affect both the global and local alignment transformations. However, the impact of random errors is mitigated by using multiple data points, each corresponding to an evaluation pose, and minimizing the mean squared error. Additionally, systematic influences in the alignment data are mitigated during data recording. Therefore, it can be assumed that the resulting errors within the determined alignments are comparably low. Nevertheless, when the assumed errors in the alignment data are substantial, manual alignment determination may be the preferable option.

Finally, to spatially align the location data provided by the GT and the SuT, both the global and local alignment transformations are applied. The local alignment is applied to the GT location data, yielding a GT location of the local SuT frame. Likewise, the global alignment is applied to the SuT location data, resulting in a location estimate in the same coordinate frame. When applied to the timestamp-aligned experiment data associated with the specified evaluation poses $\overline{E}_{SuT,GT}$ from Equation 3.8, the following data are obtained:

$$\widehat{\overline{E}}_{SuT,GT} = [k, t_{k,SuT}, A_{local} \cdot T_{E,k,SuT}, T_{E,k,GT} \cdot A_{global}]]_{k=1}^{l} \tag{3.15}$$

Similarly, for each location data point $i \in [1, m]$ provided by the SuT, it follows:

$$\overline{E}_{SuT,GT} = [t_{i,SuT}, A_{local} \cdot T_{E,i,SuT}, T_{E,i,GT} \cdot A_{global}]]_{i=1}^{m} \tag{3.16}$$

Furthermore, these considerations apply to the data related to the latency timestamp, resulting in $\widehat{\overline{E}}_{SuT,GT,lat}$ and $\overline{E}_{SuT,GT,lat}$. Together with $\widehat{\overline{E}}_{SuT,GT}$ and $\overline{E}_{SuT,GT}$, they are called evaluation data. Hence, the evaluation data consist of two sets of location data provided by the SuT with corresponding GT data that are timestamp and spatially aligned. For $\widehat{\overline{E}}_{SuT,GT}$ and $\widehat{\overline{E}}_{SuT,GT,lat}$, the data points have been additionally associated with the evaluation poses using the proposed *ClosestPoseAssociation* algorithm. In the subsequent method, these evaluation data sets are finally utilized to compute the performance metrics.

(3) Performance Metric Determination

This method focuses on determining performance metrics based on the evaluation data sets obtained from the previous method. The metrics correspond to the location data requirement parameters introduced earlier, namely the data output, localization accuracy, localization repeatability, update rate, system latency, and availability. However, since availability describes the portion of time during which location data is provided under the given requirements, it is not considered an independent performance metric but in conjunction with other performance characteristics by setting tolerances for the limits of error distributions. The determination of the data output is straightforward, as it involves describing the DoFs of the location data provided by the SuT, thus, referring to the dimensions of E_{SuT}. In the following, the determination of the performance metrics for the remaining location data requirement parameters is elaborated.

Localization Accuracy Localization accuracy is characterized by the distribution of the deviations between the timestamp as well as spatially aligned GT and SuT data. The *T&E 4Log Benchmarking Procedure* utilizes data points corresponding to the specified evaluation poses ($\widehat{\overline{E}}_{SuT,GT}$) to assess this parameter. The deviation in each DoF, as well as combinations represented by Euclidean distances, follows a distribution that can be characterized using various metrics. The calculation of accuracy metrics for localization accuracy from sets of corresponding locations has been presented in *ISO/IEC 18305*. Nonetheless, for the sake of clarity and coherence, a concise description is presented below.

First, the localization error data $\widehat{\overline{\epsilon}}$ is computed, which describes the component-wise differences between the SuT and the GT data:

$$\widehat{\overline{\epsilon}} = \widehat{\overline{E}}_{SuT} - \widehat{\overline{E}}_{GT} \tag{3.17}$$

$$\Leftrightarrow \widehat{\overline{\epsilon}} = [k, t_{k,SuT}, [\widehat{\overline{\epsilon}}_{x,k}, \widehat{\overline{\epsilon}}_{y,k}, \widehat{\overline{\epsilon}}_{z,k}, \widehat{\overline{\epsilon}}_{\phi,k}, \widehat{\overline{\epsilon}}_{\theta,k}, \widehat{\overline{\epsilon}}_{\psi,k}]]_{k=1}^{l} \tag{3.18}$$

Furthermore, the absolute error data are computed as the absolute value of each error component:

$$|\widehat{\overline{\epsilon}}| = [k, t_{k,SuT}, [|\widehat{\overline{\epsilon}}_{x,k}|, |\widehat{\overline{\epsilon}}_{y,k}|, |\widehat{\overline{\epsilon}}_{z,k}|, |\widehat{\overline{\epsilon}}_{\phi,k}|, |\widehat{\overline{\epsilon}}_{\theta,k}|, |\widehat{\overline{\epsilon}}_{\psi,k}|]]_{k=1}^{l} \tag{3.19}$$

Next, a column is added to the absolute error data to describe the horizontal position error $\widehat{\overline{\epsilon_h}}$, which represents the Euclidean distance between the horizontal position estimate of the SuT and the GT at each evaluation pose, referring to the error most commonly considered [88]:

$$\widehat{\overline{\epsilon_h}} = [\sqrt{\widehat{\overline{\epsilon}}_{x,k}^2 + \widehat{\overline{\epsilon}}_{y,k}^2}]_{k=1}^{l} \tag{3.20}$$

To provide a comprehensive understanding of the distribution of error data, the following statistical metrics should be computed for each error component of the $\widehat{\overline{\epsilon}}$ and the absolute error data $|\widehat{\overline{\epsilon}}|$.

- Mean error
- Standard deviation
- Median error
- 95th percentile

The *ISO/IEC 18305* recommends using the 95th percentile as a widely adopted metric in practice, which is why it is also incorporated in the *T&E 4Log Benchmarking Procedure*. Percentiles are crucial in assessing system suitability as they indicate the proportion of location data points that meet specified location data requirements. Therefore, a percentile should ideally align with the required level of availability (A). However, the *Performance Evaluation* seeks to present metrics independent of specified requirements. As a solution, this work suggests providing a set of percentiles corresponding to defined availability levels. To establish reasonable thresholds, the *T&E 4Log Framework* aligns with the Six Sigma (6σ) methodology, a well-established tool for quality management, aimed at improving process quality to a level where only 0.00034 % are defective [118]. Consequently, availability levels are given as follows.

- Very low (2σ): $A > 69$ %
- Low (3σ): $A > 93.3$ %
- Moderate (4σ): $A > 99.38$ %
- High (5σ): $A > 99.977$ %
- Very high (6σ): $A > 99.9997$ %

Hence, in addition to the metrics for localization accuracy suggested by *ISO/IEC 18305*, it is recommended to calculate the percentiles corresponding to these specified availability levels.

Localization Repeatability Localization repeatability is a relevant capability of ILSs for many applications in intralogistics, as will be highlighted within the *Requirement Specification*. Nevertheless, the determination of localization repeatability has not been addressed in the prevalent T&E methodologies. In this work, the term has been previously defined as the degree of agreement between location estimates at the same true location. In practice, it is sometimes mistakenly equated with the precision of localization accuracy characterized by the standard deviation, as discussed by Stephan *et al.* [119]. However, while the standard deviation of localization accuracy is undoubtedly significant, it does not capture this capability.

An intuitive approach to characterizing localization repeatability would be approaching the same evaluation pose multiple times and measuring the deviations between the location estimates. However, this approach is not practical since the localization repeatability of an ILS may surpass the accuracy with which this evaluation pose can be approached. Alternatively, the deviations between the error components can be determined using the GT data for each approximate approach to the same evaluation pose. Since this approach is expected to yield more accurate and consistent results, it is employed by the *T&E 4Log Benchmarking Procedure*.

Figure 3.16 illustrates this concept by providing a top view of an evaluation pose with a tolerance area. The figure shows the GT and SuT evaluation data for a repeated approach to this evaluation pose. The goal is to determine the repetition error $\widehat{\widehat{\epsilon}}_{rep,p}$ for each repetition $p \in [1, r]$ where r is the total number of repetitions. Although the GT location $\widehat{\widehat{E}}_{idx_1}$ for the first approach is not exactly the same as the GT location $\widehat{\widehat{E}}_{idx_2}$ for the second approach, the repetition error $\widehat{\widehat{\epsilon}}_{rep,p}$ can be computed as the component-wise difference between the two corresponding location errors $\widehat{\widehat{\epsilon}}_{idx_1}$ and $\widehat{\widehat{\epsilon}}_{idx_2}$.

The pseudocode for the *ComputeRepErrorData* algorithm is presented to help stakeholders determine the introduced repetition errors (Algorithm 2). The algorithm iterates over each evaluation pose and checks within the second *for* loop whether the evaluation pose has been reached before by iterating over all previous poses. If the evaluation pose has been repeated, the differences between the error components of the error vectors $\widehat{\widehat{\epsilon}}_{idx_2}$ and $\widehat{\widehat{\epsilon}}_{idx_1}$ are computed. The

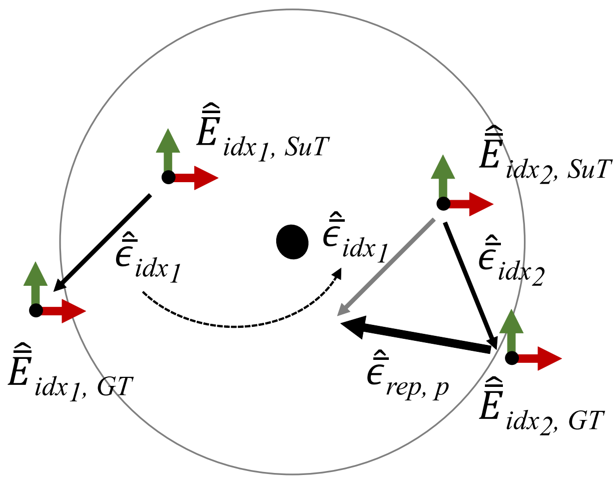


Figure 3.16: Determination of repetition error $\widehat{\widehat{\epsilon}}_{rep,p}$ for repetition p at evaluation pose with tolerance area

difference calculated along with a new index p is added to the *errorDataRep* list. The resulting data list *errorDataRep* can be represented as follows:

$$\begin{aligned}\widehat{\widehat{\epsilon}}_{rep} &= [p, [\widehat{\widehat{\epsilon}}_{x,idx_2} - \widehat{\widehat{\epsilon}}_{x,idx_1}, \widehat{\widehat{\epsilon}}_{y,idx_2} - \widehat{\widehat{\epsilon}}_{y,idx_1}, \widehat{\widehat{\epsilon}}_{z,idx_2} - \widehat{\widehat{\epsilon}}_{z,idx_1}, \\ &\qquad \widehat{\widehat{\epsilon}}_{\phi,idx_2} - \widehat{\widehat{\epsilon}}_{\phi,idx_1}, \widehat{\widehat{\epsilon}}_{\theta,idx_2} - \widehat{\widehat{\epsilon}}_{\theta,idx_1}, \widehat{\widehat{\epsilon}}_{\psi,idx_2} - \widehat{\widehat{\epsilon}}_{\psi,idx_1}]]_{p=1}^{r} \end{aligned} \tag{3.21}$$

$$\Leftrightarrow \widehat{\widehat{\epsilon}}_{rep} = [p, [\widehat{\widehat{\epsilon}}_{x,rep}, \widehat{\widehat{\epsilon}}_{y,rep}, \widehat{\widehat{\epsilon}}_{z,rep}, \widehat{\widehat{\epsilon}}_{\phi,rep}, \widehat{\widehat{\epsilon}}_{\theta,rep}, \widehat{\widehat{\epsilon}}_{\psi,rep}]]_{p=1}^{r} \tag{3.22}$$

The absolute repetition error data $|\widehat{\widehat{\epsilon}}_{rep}|$ can also be computed and the horizontal component added. Next, the localization repeatability for the experiment is characterized by applying

Algorithm 2 *ComputeRepErrorData*

Require: List of evaluation poses (E_{poses}): *evaluationPoses*, List of localization error data for each DoF of the location estimate ($\widehat{\widehat{\epsilon}}$): *errorData*
Ensure: List of repeated Poses with attached error distances ($\widehat{\widehat{\epsilon}}_{rep}$): *errorDataRep*

```
Initialize errorDataRep as empty list and idx_new ← 0
for idx_2 from 0 to size(evaluationPoses) - 1 do
    for idx_1 from idx_2 to 0 do
        if evaluationPoses[idx_2] equals evaluationPoses[idx_1] then
            Initialize errorDist as component-wise difference of location data from errorData[idx_2] and errorData[idx_1]
            idx_new ← idx_new + 1
            Add tuple of (idx_new, errorDist) to errorDataRep
            break
        end if
    end for
end forreturn errorDataRep
```

various statistical metrics to the data set, similar to the determination of metrics for localization accuracy from the error data $\widehat{\overline{\epsilon}}$, including the percentiles referring to the previously introduced sigma levels.

The *ISO 18646-2* standard, which focuses on assessing the navigation capabilities of mobile service robots, describes a similar approach. However, according to this standard, repeatability is determined by calculating the deviations for each repeated evaluation pose individually. Subsequently, the overall localization repeatability across all poses is determined by computing the mean value. This approach contrasts with that used in the *T&E 4Log Framework* because it necessitates a significantly higher number of path repetitions, leading to substantially increasing the testing effort.

The approach proposed by the *T&E 4Log Benchmarking Procedure* to determine localization repeatability does not rely on highly accurate ELT control and many approaches to each evaluation pose, thus increasing accuracy while minimizing testing effort. In addition, an algorithm has been provided to generate a set of repetition errors. Ultimately, the distribution of localization repeatability is characterized by employing various metrics, similar to the ones applied for localization accuracy.

System Latency In this work, system latency refers to the time delay inherent to the SuT between the actual measurement and the provision of location data by the system. This definition excludes any additional time delays caused by transmission times related to network traffic, as they are not inherent to the SuT, contrasting with the definition of the *EVARILOS Benchmarking Handbook* according to which latency is determined as the time from request to provision of location data. Hence, the approach employed within the *T&E 4Log Framework* determines the "push" latency according to the *ISO/IEC 18305*, significant for ILSs providing location data with constant update rate.

For synchronized clocks, a straightforward method for computing latency in the case of synchronized clocks is to determine the difference between the timestamps $t_{i,SuT}$ and $t_{i,lat}$ for $i \in [1, m]$. However, this approach is based on the trustworthiness of the timestamps provided by the SuT. For independent black-box testing, the *T&E 4Log Benchmarking Procedure* incorporates an alternative approach to determine system latency based on the analysis of velocity-dependent errors inherent to the error data $\overline{\epsilon}_{lat}$ corresponding to the latency timestamps.

By utilizing the location estimates associated with the timestamp for data provision, the determined localization error incorporates an additional error component due to the change in position of the ELT. Assuming that no other significant errors are induced by velocity, the relationship between the velocity-dependent error $\vec{\epsilon}_v$, the latency t_{lat}, and the velocity $\vec{v}$ is as follows:

$$\vec{\epsilon}_v = -\vec{v} \cdot t_{lat} \tag{3.23}$$

This equation holds for each DoF of $\vec{\epsilon}_v$. However, to isolate the velocity-dependent error the absolute velocity and the corresponding error in the same direction are used. Therefore, if the velocity-dependent error and the absolute velocity are known, the latency t_{lat} can be determined as follows:

$$t_{lat} = \frac{\epsilon_v}{v} \tag{3.24}$$

For highly accurate reference systems, the velocity is commonly provided by the system or can be computed as the time derivative of the GT location data. The velocity-dependent error can be derived based on the aligned evaluation data ($\overline{E}_{SuT,GT}$) and the corresponding localization errors ($\overline{\epsilon}_{SuT,GT}$). To isolate the velocity-dependent localization error from other systematic and random errors in the location estimates, the following approach, illustrated in Figure 3.17, is proposed.

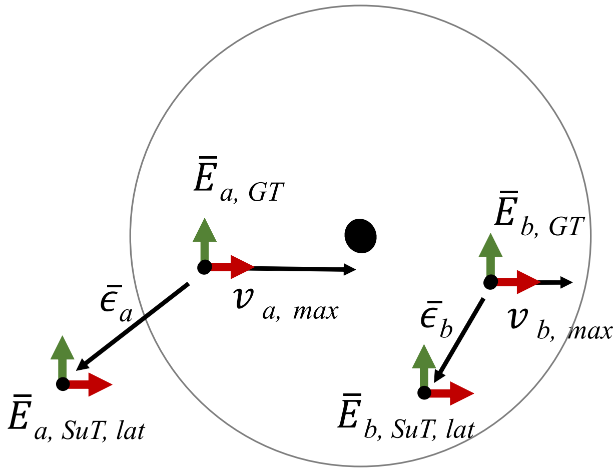


Figure 3.17: Determination of velocity-dependent localization error at an evaluation pose with a tolerance area based on the location estimates a and b

For each entry into the tolerance space of an evaluation pose k, two corresponding pairs ($\overline{E}_{a,SuT}$, $\overline{E}_{a,GT}$ and $\overline{E}_{b,SuT}$, $\overline{E}_{b,GT}$) of data points are considered, where $a, b \in [0, m]$. Assuming that, besides the velocity-dependent error, all other systematic error components remain constant for all location estimates within the tolerance space, the velocity-dependent error component associated with evaluation pose k can be determined as the difference between the localization errors ϵ_a and ϵ_b. Hence, it follows:

$$t_{lat,k} = \frac{|\epsilon_a - \epsilon_b|}{|v_a - v_b|} \tag{3.25}$$

To maximize the effect of the velocity-dependent error relative to random errors, a and b are chosen to correspond to the maximum and minimum velocities:

$$t_{lat,k} = \frac{|\epsilon_a - \epsilon_b|}{|v_{a,max} - v_{b,min}|} \tag{3.26}$$

To mitigate the effect of random and non-constant systematic errors on the accuracy of the computed latency value for the evaluation pose k, the median of the latency values $t_{lat,k}$ for the evaluation poses $k \in [1, m]$ is calculated:

$$t_{lat} = \text{median}\{[t_{lat,k}]_{k=1}^{m}\} \tag{3.27}$$

This approach is suitable for evaluating systems with relatively high latency compared to random localization errors. Furthermore, in contrast to the approach presented by Bouvet *et al.* [120] for determining GPS latency, which relies on multiple abrupt changes of direction on a straight line, the approach incorporated within the *T&E 4Log Benchmarking Procedure* allows the determination of the system latency based on the same experiment data as used for determining the other metrics.

Update Rate Computing the update rate is only meaningful for ILSs that provide location data with a constant update rate. Typically, the update rate can be configured to some extent as part of the system's parameter configuration. However, computing the update rate based on experiment data allows for checking how much it deviates from the configured or specified value. Furthermore, adjusting the update rate potentially impacts the accuracy of the provided location data. Therefore, the update rate must be determined together with other performance metrics.

Assuming a constant update rate f_{update}, the value is obtained by computing the inverse of the median difference between consecutive timestamps from the SuT measurements. The data points associated with the evaluation poses are used to provide better control and consistency with the other metrics. The median value is utilized rather than the mean because it entails greater robustness against outliers:

$$f_{update} = (\text{median}\{[t_{k+1,SuT} - t_{k,SuT}]_{k=1}^{m}\})^{-1} \tag{3.28}$$

Where m represents the total number of location measurements the SuT provides.

Overall, the comprehensive set of metrics provides valuable insights into system performance based on the experiment data of a given experiment. The literature has extensively discussed the usefulness of different metrics and the appropriate number of metrics to consider [11, 121]. The underlying conflict of this discussion arises from the fact that system developers or testers may utilize multiple different metrics to analyze system behavior, which could overwhelm system users. To address this conflict, the *T&E 4Log Framework* incorporates the computation of various metrics commonly used in the literature during the *Performance Evaluation* procedure but only provides the relevant metrics for each specific application to determine system suitability during the *System Evaluation* procedure. This approach ensures that the complexity of the information output is reduced for system users without losing important information for system developers. System testers rely on the whole experiment and evaluation data to verify results, considering the various challenges such as data recording, clock synchronization, interpolation, alignment, and metric computation.

The *Performance Evaluation* is the final procedure within the *T&E 4Log Benchmarking Procedure*. It provides a structured approach for T&E by conducting empirical experiments in partially controlled test environments, thus addressing RQ2. By comparing performance metrics from multiple experiments, repeatability can be assessed based on the same experiment specification, while replicability can be evaluated based on the same test scenario. Furthermore, comparing experiments based on varying specifications or scenarios enables analysis of the impact of application-driven influencing factors. However, the comparison of different experiments falls outside the scope of the *T&E 4Log Framework*. This chapter continues by discussing the specification of location data requirements, which are ultimately matched with the performance metrics determined from an experiment to assess the system's suitability.

3.3.8 Requirement Specification

The objective of the *Requirement Specification* procedure is to specify location data requirements. These specified requirements are then used in the *System Evaluation* procedure to assess the suitability of an ILS based on the previously established performance metrics. As such, the *Requirement Specification* and *System Evaluation* provide an answer to RQ3. An earlier version of the location data requirements specification has been published and presented at the CPSL 2023 [83].

As previously discussed, the *T&E 4Log Framework* focuses on the specification of location data requirements to ensure reliable localization functions. This is based on two key concepts, each represented by a derived parameter. The first derived parameter is the effective localization error, which represents the error ultimately relevant to an application process or activity, based on the systems' performance parameters and the application processes. The second parameter is the requirement margin, which sets the maximum acceptable effective localization error to ensure the reliable association of an entity with an interest space.

The *Requirement Specification* procedure consists of three main methods, as depicted in Figure 3.18. These involve (1) identifying the relevant localization functions, (2) specifying the corresponding requirement margin, and (3) specifying the effective localization error. The following discussion delves into these methods in detail, providing a deeper understanding of the introduced concepts.

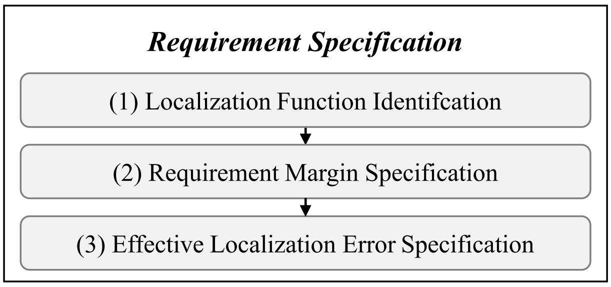


Figure 3.18: Methods of the *Requirement Specification* procedure

(1) Localization Function Identification

First, the operations, referred to as localization functions of the AuC must be identified that require particular localization capabilities to reliably identify the presence or absence of an ELT within a multidimensional interest space. For instance, within a typical intralogistics scenario, a localization function could involve the detection of a pallet within a storage compartment to notify the Warehouse Management System (WMS) or the detection of a forklift truck in front of a certain shelf. In these cases, the pallet and the forklift truck represent ELTs, whereby the storage compartment and “in front of the shelf” define the interest spaces. When the interest space is entered or exited, the higher-level AuC is notified, enabling it to process this information, ultimately generating value for the end user. Typically, a localization function can be associated with an application process or even with a specific activity of this process. An application can incorporate one or many localization functions, that can be based on location data provided from different ILSs and entities, as depicted in Figure 3.19.

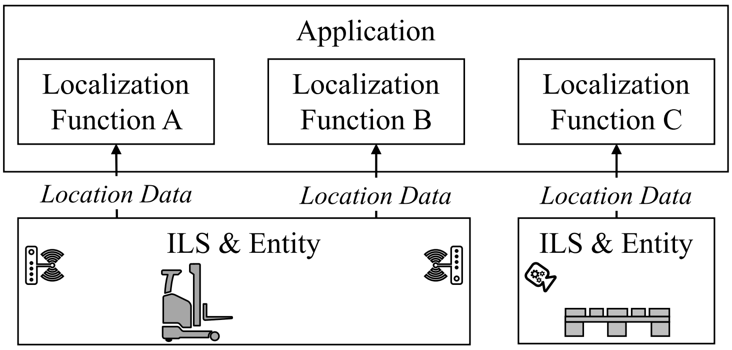


Figure 3.19: Localization functions incorporated within an application based on the location of different ILSs and ELTs

For each localization function, distinct location data requirements must be derived. In cases where an AuC encompasses multiple localization functions, the application requirements are given as the union of requirements of the localization functions. In the following, the specification of the requirement margin for a particular localization function is presented.

(2) Requirement Margin Specification

As previously outlined, the requirement margin establishes a maximum threshold for the effective localization error by analyzing the identified localization functions, thereby ensuring reliable detection of an entity’s presence or absence within a specified interest space. Alongside the interest space, the concept of a motion space is central, which denotes a multi-dimensional space wherein the entity can move without providing false associations with an interest space. As an optional measure, a safety margin can be considered within the requirement margin, which is particularly relevant for safety-critical applications.

To illustrate the specification of the requirement margin based on the consideration of the motion space, the interest space, and the a safety margin an example of a forklift truck in a warehouse aisle is considered, as depicted in Figure 3.20. The specific localization function under consideration is the reliable identification of the forklift truck’s presence within the correct warehouse aisle. Consequently, the interest space is delineated by the boundaries to adjacent aisles, while the motion space is constrained by the open aisle area and the forklift’s dimensions. Hence, the motion space is given by the open aisle area minus half the forklift’s width (w). In the case where the forklift truck is situated on the edge of the motion space, it should still be reliably identified within the interest space. This gives rise to the definition of the requirement margin defining the maximum allowable error as the minimal distance between the edge of the motion space and the interest space.

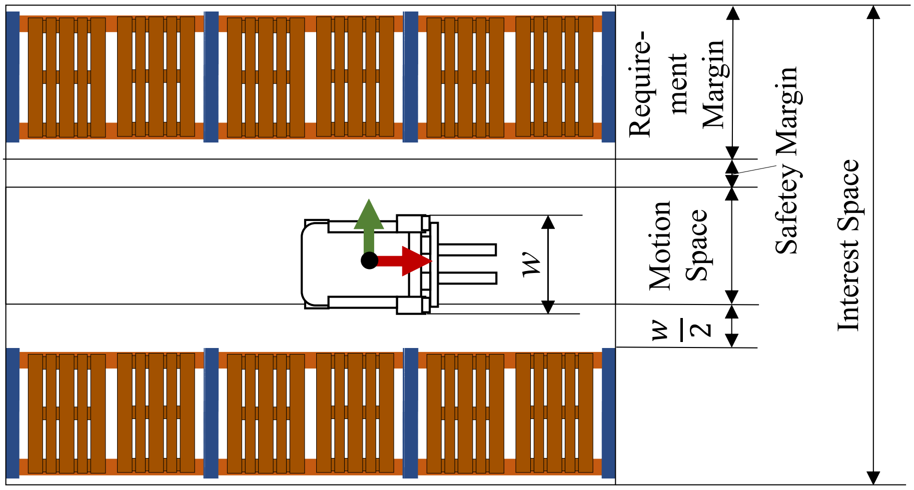


Figure 3.20: Top view of forklift within warehouse aisle, indicating the associated spaces for asserting reliable localization function

The considerations outlined in this illustrative example can be generalized to accommodate generic localization functions for intralogistics, encompassing arbitrary multi-dimensional interest spaces. The interest space can typically encompass up to six DoFs or any combination. For instance, if the side of the warehouse facing the forklift holds relevance, only the heading angle may be considered. Hence, the requirement margin can be expressed as a vector, with each component representing a DoF of the interest space. To indicate the dependency on the specified interest space I, the motion space M, and the safety margin S, the requirement margin is denoted as the vector $\vec{R}(I,M,S)$. It contains the minimal distance of each component of a relevant dimension of the interest space. In cases where the absence of an entity needs to be detected, the motion space is situated outside of the interest space.

Typically, intralogistic environments such as warehouses and production halls can be abstracted well by rectangular or cuboid spaces, such as warehouse aisles or storage compartments. Leveraging these abstractions and resultant symmetries can aid in determining the requirement margins in practical scenarios. Furthermore, in most intralogistics scenarios, it is satisfactory to consider a maximum of four DoFs, disregarding roll and pitch. Moreover, distinguishing between horizontal error components is often neither relevant from an application perspective nor meaningful from a technological perspective. Finally, in the case of ground-based entities, the vertical component (z) can be disregarded.

In summary, the requirement margin establishes an upper limit for the effective localization error by considering the dimensions of the interest space, motion space, and safety margin for a localization function. In the subsequent section, the contribution of each location data requirement parameter to the effective localization error is discussed.

(3) Effective Localization Error Specification

This method involves the specification of the effective localization error, which denotes the error ultimately relevant to an application process or activity. The following three main components of location data parameters contribute to the effective localization error.

- **Relevant localization error ($\vec{\epsilon}_{rel}$):** The relevant localization error arises either from a system's localization accuracy or from its localization repeatability, depending on the interest space. If the interest space is defined within the true coordinates of an environment, localization accuracy takes precedence. On the other hand, if the interest space is defined using the coordinates provided by the ILS itself, the relevant criterion is localization repeatability. For example, within the process of an autonomous forklift truck picking up a pallet from a location where the same truck had previously placed it, localization repeatability would be the decisive factor. This also applies to another truck employing the same reference coordinate system. The relevant localization error is assumed to be independent of the entity's motion.

- **Time gap error ($\vec{\epsilon}_{gap}$):** A location estimate obtained from an ILS is associated with a specific point in time. However, a localization function may process this estimate at a later time, by which the entity's location might have already changed. The resulting deviation in the location is considered the time gap error. It is significant for dynamic processes when the localization function strictly dictates the point in time for data processing. In safety-critical real-time applications, this is typically the case.

- **Time delay error ($\vec{\epsilon}_{del}$):** The time delay error arises from the possible change in the entity's location that may have occurred before the location data is made available to the localization function. It is significant for dynamic processes involving localization functions in which the entity's current location must be known [120].

The effective localization error denoted as $\vec{\epsilon}_{eff}$ is given as the sum of the three introduced components:

$$\vec{\epsilon}_{eff} = \vec{\epsilon}_{rel} + \vec{\epsilon}_{gap} + \vec{\epsilon}_{del} \tag{3.29}$$

Analogous to the considerations of the SuT's coordinate frames during experimentation, an ILS provides location estimates of the local coordinate frame established by a localization device, markers, or features on the ELT. In the following, this coordinate frame is referred to as the entity's localization frame (O_{loc}). However, a different coordinate frame on the ELT might be of interest for the localization function, which is referred to as the entity's interest frame (O_{int}). The transformation from the localization frame to the interest frame is denoted as T_{loc}^{int}, which is typically measured or estimated.

Figure 3.21 serves as an illustrative example, depicting the top view of a forklift truck. The coordinate frame of interest (O_{int}) is positioned between the fork ends. The location data, however, are provided in the entity's localization frame O_{loc}. Inaccuracies associated with the rotational components of provided location estimates transfer to additional inaccuracies of the translational components within the interest frame, depending on T_{loc}^{int}. If the rotation of the entity is unknown, the distance between the coordinate frames fully contributes to the effective localization error.

Distinguishing between the entity's localization frame and the entity's interest frame holds paramount importance in practical applications. When determining the suitability of a system it is therefore crucial to consider the potential locations of the entity's localization frame. While

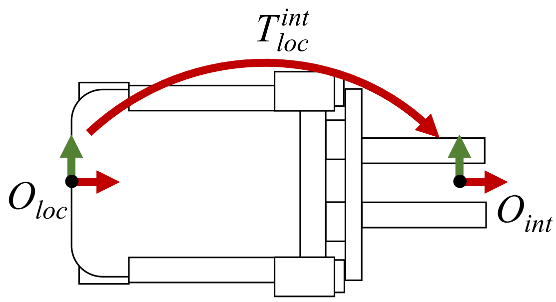


Figure 3.21: Top view of transformation T_{loc}^{int} between the entity's localization frame O_{loc} and the entity's interest frame O_{int}

this issue can be mitigated if the entity's localization frame is situated within or near the interest frame, due to spatial constraints, such optimal placement is often impractical in real-world scenarios.

To emphasize that the introduced error components from Equation 3.29 must refer to the entity's interest frame, they are annotated with a subscript *int*:

$$\vec{\epsilon}_{eff} = \vec{\epsilon}_{rel,int} + \vec{\epsilon}_{gap,int} + \vec{\epsilon}_{del,int} \tag{3.30}$$

The individual components in Equation 3.30 can be expressed through performance metrics and process parameters. As explained earlier, the relevant localization error is determined either by the localization accuracy or the localization repeatability, which is then transformed into the entity's interest frame. Both parameters are effectively characterized by percentiles, according to the specified availability levels A of the localization function. The corresponding percentile of the localization accuracy or localization repeatability, expressed in the entity's interest frame, is represented as $\vec{\epsilon}_{A,Acc/Rep,int}$.

To quantify the time gap error, the change in location over time ($\vec{v}_{int}(t)$) within the entity's interest frame is considered for the time Δt between the provision and utilization of the location data. The time gap error is then expressed as an integral over the velocity as follows:

$$\vec{\epsilon}_{gap,int} = \int_{t}^{t+\Delta t} \vec{v}_{int}(t) \tag{3.31}$$

Similarly, the time delay error $\vec{\epsilon}_{del,int}$ is determined by integrating the entity's velocity $\vec{v}_{int}(t)$ over the time delay t_{del}. In practice, the time delay includes the system's latency t_{lat} and, if the data is processed remotely, network transmission times t_{net} for providing the location data to the localization function [120]:

$$\vec{\epsilon}_{del,int} = \int_{t}^{t+t_{del}} \vec{v}_{int}(t) \tag{3.32}$$

$$\Leftrightarrow \vec{\epsilon}_{del,int} = \int_{t}^{t+t_{lat}+t_{net}} \vec{v}_{int}(t) \tag{3.33}$$

Inserting Equation 3.31 and 3.33 into Equation 3.30, it follows:

$$\vec{\epsilon}_{eff} = \vec{\epsilon}_{A,Acc/Rep,int} + \int_{t}^{t+\Delta t} \vec{v}_{int}(t) + \int_{t}^{t+t_{del}} \vec{v}_{int}(t) \tag{3.34}$$

To provide an estimation of location data requirements, an upper bound can be set for both time delay and time gap error by assuming a maximum velocity and expressing Δt as the time between two consecutive location updates t_{gap}. Hence, for systems with a constant update rate f_{update}, it follows:

$$\vec{\epsilon}_{eff} = \vec{\epsilon}_{A,Acc/Rep,int} + \vec{v}_{max,int} \cdot (\frac{1}{f_{update}} + t_{lat} + t_{net}) \tag{3.35}$$

Finally, Equation 3.34 or Equation 3.35 can be employed to provide a conservative estimation of the effective localization error. Depending on the specific AuC and the knowledge of the velocity profile, considering the provided time integrals, maximum, or even average velocities might be preferred.

The reliable detection of an entity within an interest space requires the effective localization error to be equal to or lower than the specified requirement margin. This condition can be expressed as follows:

$$\vec{R}(I,M,S) \geq \vec{\epsilon}_{eff} \tag{3.36}$$

$$\Leftrightarrow \vec{R}(I,M,S) \geq \vec{\epsilon}_{A,Acc/Rep,int} + \int_{t}^{t+\Delta t} \vec{v}_{int}(t) + \int_{t}^{t+t_{del}} \vec{v}_{int}(t) \tag{3.37}$$

$$\Rightarrow \vec{R}(I,M,S) \geq \vec{\epsilon}_{A,Acc/Rep,int} + \vec{v}_{max,int}(\frac{1}{f_{update}} + t_{lat} + t_{net}) \tag{3.38}$$

These Equations provide a conservative assessment by assuming the ELT to be located at the boundary of the motion space, that refers to the minimal distance to the interest space.

In addition to the proposed approach centering on asserting a reliable localization function, the *T&E 4Log Framework* allows for the manual specification of requirements for each performance characteristic going beyond the scope of localization functions.

3.3.9 System Evaluation

The purpose of the *System Evaluation* procedure is to assess the suitability of an ILS for an application. This task within the final procedure of the *T&E 4Log Framework* is primarily aimed at system users to support informed system selections. However, it also serves the purpose of assisting system developers in understanding the performance and limitations of the system for their targeted applications.

The assessment of system suitability is based on the location data requirements, expressed through Equations 3.37 or Equation 3.38, taking into account the process-dependent relationships between the location data parameters. The performance metrics obtained from the *T&E 4Log Benchmarking Procedure*, such as localization accuracy, localization repeatability, system latency, and update rate, are inserted into Equation 3.37 or Equation 3.38, under consideration of the potentially relevant transformation into the entity's interest. If the condition holds, reliable operation is ensured and the system is deemed suitable for the localization function.

The condition must be checked for each localization function to assess suitability for the AuC as a whole. On the other hand, a set of performance metrics is associated with an experiment specification, which is associated with a test scenario typically designed by considering a real-world application scenario. Consequently, an application can be associated with multiple sets of performance metrics, resulting from more or less challenging experimental conditions. Therefore, by utilizing different sets of performance metrics for different requirements of localization functions, it can be analyzed under which conditions a localization system is expected to ensure the reliable operation of which localization functions.

The *System Evaluation* procedure determines the suitability of an ILS for a considered application by evaluating whether the stated condition for achieving reliable localization functions is met, based on the performance metrics obtained from an experiment. It is important to note that the given condition provides a conservative estimate, and the system is, thus, not automatically deemed unsuitable if it is not met.

Overall, The *T&E 4Log Framework* shows considerable potential in facilitating the reliable comparison of system performance and assisting stakeholders in making informed decisions regarding system selection. Nevertheless, the applicability and utility necessitate validation. Hence, the following chapter presents an empirical examination, setting the foundation for a profound discussion in Section 5.

4 Empirical Examination

The empirical base for this work is created by applying the *T&E 4Log Framework*, thus gaining valuable information on its applicability and utility. This chapter is structured into three sections, each presenting an empirical study involving experiments carried out within the facility of the ITL. Firstly, in Section 4.1, the *T&E 4Log Framework* is applied to an exemplary case study to determine the suitability of a LiDAR-based and an UWB-based system for the practical application "Mobile Robot for Material Transport". The remaining two sections present studies aimed primarily at providing a basis for analyzing significant stakeholder demands. As such, Section 4.2 presents the results of a study to examine the repeatability of experiments, while Section 4.3 focuses on the comparability of the results of experiments under varying conditions. The outcomes of this chapter are discussed in Chapter 5 to complete the validation process. In addition to its main purpose of providing an empirical basis for the critical discussion on the validity of the framework and its components, this chapter presents the exemplary usage of the developed methodology and offers comprehensive insights into the performance and characteristics of the examined ILSs as well as into the application "Mobile Robot for Material Transport". The experiment and evaluation data, as well as the performance results, are available in this dissertation's *GitLab* repository [122]. Additionally, the repository hosts the source code for the *T&E 4Log App*. This app facilitates T&E in alignment with the *T&E 4Log Benchmarking Procedure*. An overview of the usage of the *T&E 4Log App* is provided in the Appendix.

4.1 Exemplary Case Study: Mobile Robots for Material Transport

In this section, a practical case study is presented that involves the data-driven assessment of the suitability of two ILSs for the application of "Mobile Robots for Material Transport". The section begins with a brief motivation for this specific case study and then proceeds to present the sequential application of each procedure of the *T&E 4Log Framework* for both ILSs.

The specific application of mobile robots for material transport is chosen for this case study due to its high relevance and potential impact on logistics. As discussed in the state of the art in Section 2.1.3, mobile robots have emerged as a high-impact solution in intralogistics, offering various potential use-cases and benefits [79]. The high complexity of robotics, particularly with respect to location data requirements, requires a systematic and data-driven approach, as proposed by the *T&E 4Log Framework*. Furthermore, the absence of a single dominant indoor localization technology for mobile robots further motivates the examination of different types of ILSs in this case study. Although qualitative discussions on different sensor technologies for mobile robots exist, comparative quantitative results are lacking [123].

The two selected localization systems for the empirical examination are the *SICK LOCU Localization Solution* (LOCU), which is an UWB-based system utilizing TDoA measurements [124], and *SICK LiDAR-LOC* (LLS), which is a LiDAR-based localization system utilizing the *SICK microScan3* (ms3), a multi-layer safety LiDAR scanner, for pose estimation [125, 126]. LOCU is an infrastructure-based system, relying on RF signals, while LLS primarily utilizes map matching for self-localization. The systems also differ in their data output, with the LOCU system providing horizontal position and potentially vertical position data and the LLS system providing heading information in addition to horizontal position data. Both provide absolute location data with a constant update rate, making them suitable for evaluation within the *T&E 4Log Framework*. The systems are chosen due to the relevance of the utilized technologies for intralogistics and their distinct working principles and characteristics, which enables the examination of the framework's versatility and the comparability of results across distinct technologies. Furthermore, both technologies are commonly considered for the application in mobile robotics [123], however, the methodological comparison based on T&E results remains seldom.

With the application of mobile robots for material transport and the selected localization systems motivated, the subsequent parts describe the sequential application of the procedures within the *T&E 4Log Framework*.

4.1.1 Application Description

The *Application Description* procedure serves to describe and detail the selected intralogistics application of "Mobile Robot for Material Transport". This AuC leverages fleets of floor-bound mobile robots, potentially including AGVs as well as AMRs, to execute material transport tasks. In this case study, the focus is on automated low-lift pallet trucks that transport pallets between different points within a warehouse environment. The AuC potentially involves multiple application processes. For consideration within this exemplary case study, the following three key processes have been identified. Figure 4.1 shows the resulting activity diagram for the considered application "mobile robot for material transportation".

(1) Pallet Pick-Up

The robot is directed to a specific area where it approaches the pallet to be picked up. Next, a fine alignment step is executed to prepare the subsequent steps of inserting the fork into the pallet, and finally lifting the pallet. Absolute localization plays a crucial role in approaching the pallet, while the fine alignment step is typically based on the relative localization of the pallet from the robot.

(2) Pallet Transport

Upon successful pick-up, the robot transports the pallet to a designated drop-off point. This could be a different storage place, buffer area, or processing station. The robot needs to follow an efficient path while avoiding both static and dynamic obstacles. The activities involved may include: "plan path", "navigate through facility", "monitor obstacles", and "adjust path" as needed. Here, localization assists in real-time global path planning as part of the activities "plan path" or "adjust path". The monitoring of obstacles is typically achieved by relative localization.

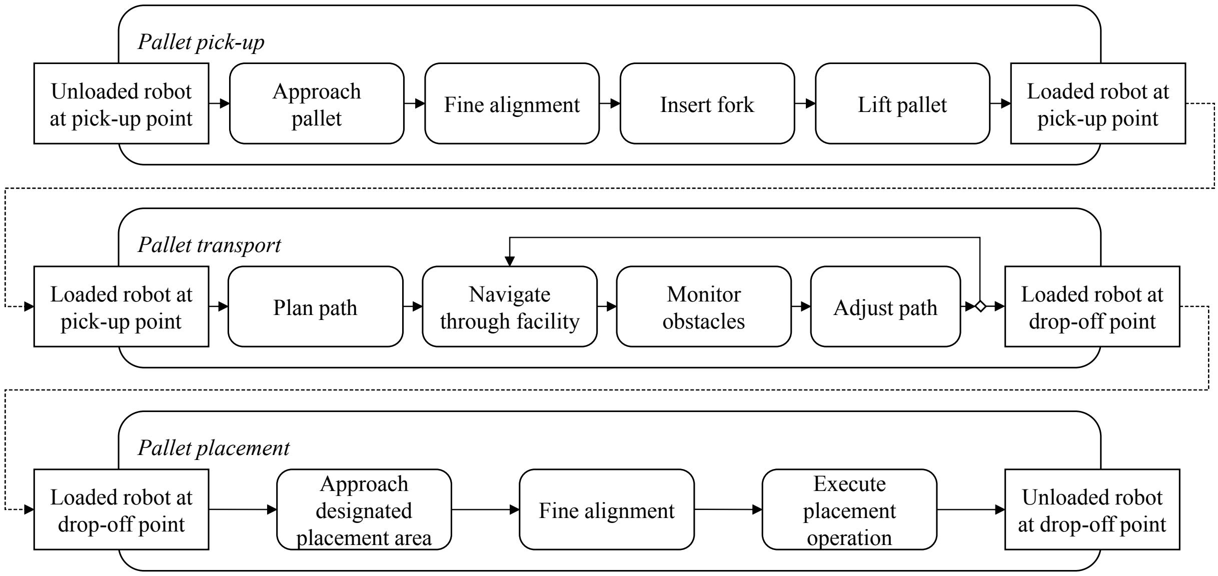


Figure 4.1: Exemplary activity diagram for the AuC "Mobile Robot for Material Transport"

(3) Pallet Placement

At the destination of the pallet, the robot places the pallet within the designated area. Similarly to the pick-up process, this also requires precise maneuvering. Here, localization helps to achieve accurate pallet placement. Activities may include: "approach designated placement area", "fine alignment", and "execute placement operation".

As an application environment, an industrial warehouse is considered, characterized by large open spaces and storage racks. Other elements could include areas for loading and unloading goods, processing areas, and possibly human-operated workstations. The environment may pose challenges to localization due to factors such as reflections or occlusions caused by storage racks, stocked items, concrete floors, and walls. The environment could be segmented into different spaces such as "bulk storage", "packing area", "pallet shelf area", or "loading docks", each entailing unique characteristics. For example, some processes or activities may be specific to certain spaces, such as "pallet placement" occurring in "loading docks". However, in this case study an open space within an industrial warehouse is considered as the application environment.

4.1.2 Scenario Definition

Based on the described AuC as well as the briefly introduced ILSs, an application-driven test scenario can now be defined according to the *Scenario Definition* procedure. First, relevant influencing factors must be identified for the technologies considered, which are subsequently characterized depending on the AuC.

To identify relevant influencing factors, the technology components of the localization systems are considered according to the taxonomy presented in Section 2.1.2. The first system (LOCU) is a UWB-based system employing the TDoA measurement technique in combination with lateration. In addition, an EKF is employed for dead reckoning. The LLS system is also employing an EKF for dead reckoning additionally supported by IMU data. However, the localization of the LLS system is primarily based on map matching, using a particle filter based on a point cloud resulting from distance and angle measurements of a LiDAR scanner. From the table presented in Chapter 3.3.4, linking common components of indoor localization technology with application-driven influencing factors, follows the specified Table 4.1. The bullets indicate the relevance of a particular influencing factor for the LOCU system in red and for the LLS system in blue, as a consequence of the utilized technology components. The last two rows summarize the application-driven influencing factors for each of the two respective systems.

The next step according to the *Scenario Definition* is the characterization of the influencing factors under consideration of the AuC. For both systems, the ELT is represented by a mobile robot, equipped with the localization devices mounted on top. This robot navigates at a maximum velocity similar to walking speed, following a pre-defined route composed of linear paths and curves on a horizontal plane. The spatial environment should resemble an "open space" within

Table 4.1: Application-driven influencing factors for LLS in blue and LOCU and red. The bullets mark the potential relevance of an influencing factor on the performance of the respective system, due to each component utilized. The final two rows summarize the potentially relevant influencing factors for each of the two systems.

	Application-driven influencing factors											
	Process				Environment					System		
Technology component	ELT	Device/marker/feature loc.	Motion	Path	Static objects	Dynamic entities	Radio interferences	Floor quality	Lighting conditions	Amount of reference nodes	Location of reference nodes	Parameter configuration
Measurement technique												
TDoA	🔴	🔴		🔴	🔴	🔴					🔴	🔴
ToF/ToA	🔵	🔵		🔵	🔵	🔵					🔵	🔵
AoA	🔵	🔵		🔵	🔵	🔵					🔵	🔵
Localization method												
Lateration			🔴	🔴						🔴	🔴	🔴
Map matching / fingerprinting			🔵	🔵	🔵	🔵				🔵	🔵	🔵
Dead reckoning			🔵 🔴	🔵 🔴								🔵 🔴
Sensor technology												
UWB							🔴					🔴
LiDAR	🔵	🔵	🔵	🔵				🔵				🔵
IMU			🔵	🔵				🔵				🔵
Localization system												
LOCU	🔴	🔴	🔴	🔴	🔴	🔴	🔴			🔴	🔴	🔴
LLS	🔵	🔵	🔵	🔵	🔵	🔵		🔵		🔵	🔵	🔵

a typical warehouse setting. To achieve this, the placement of static objects within the test area is avoided while allowing for the placement of shelves and other logistic-related equipment in peripheral areas without introducing dynamic entities. The floor, relevant for the LLS system is characterized as a typical flat concrete floor and radio interferences as commonly available signals in industrial environments such as Bluetooth and WLAN. For both ILSs, the system influences can not be characterized solely by considering the AuC. For the LLS system, the number and placement of reference nodes correspond to the environment map used for map matching. Therefore, for both systems, the number and location of the reference nodes, along with the configuration of parameters, should be defined following the installation instructions for the operational modes.

The resulting test scenario, consisting of the aggregation of characterized influencing factors, is summarized in Table 4.2. In the scope of this case study, the established test scenario lays the foundation for the *Experiment Specification* procedure as the subsequent step within the *T&E 4Log Framework*.

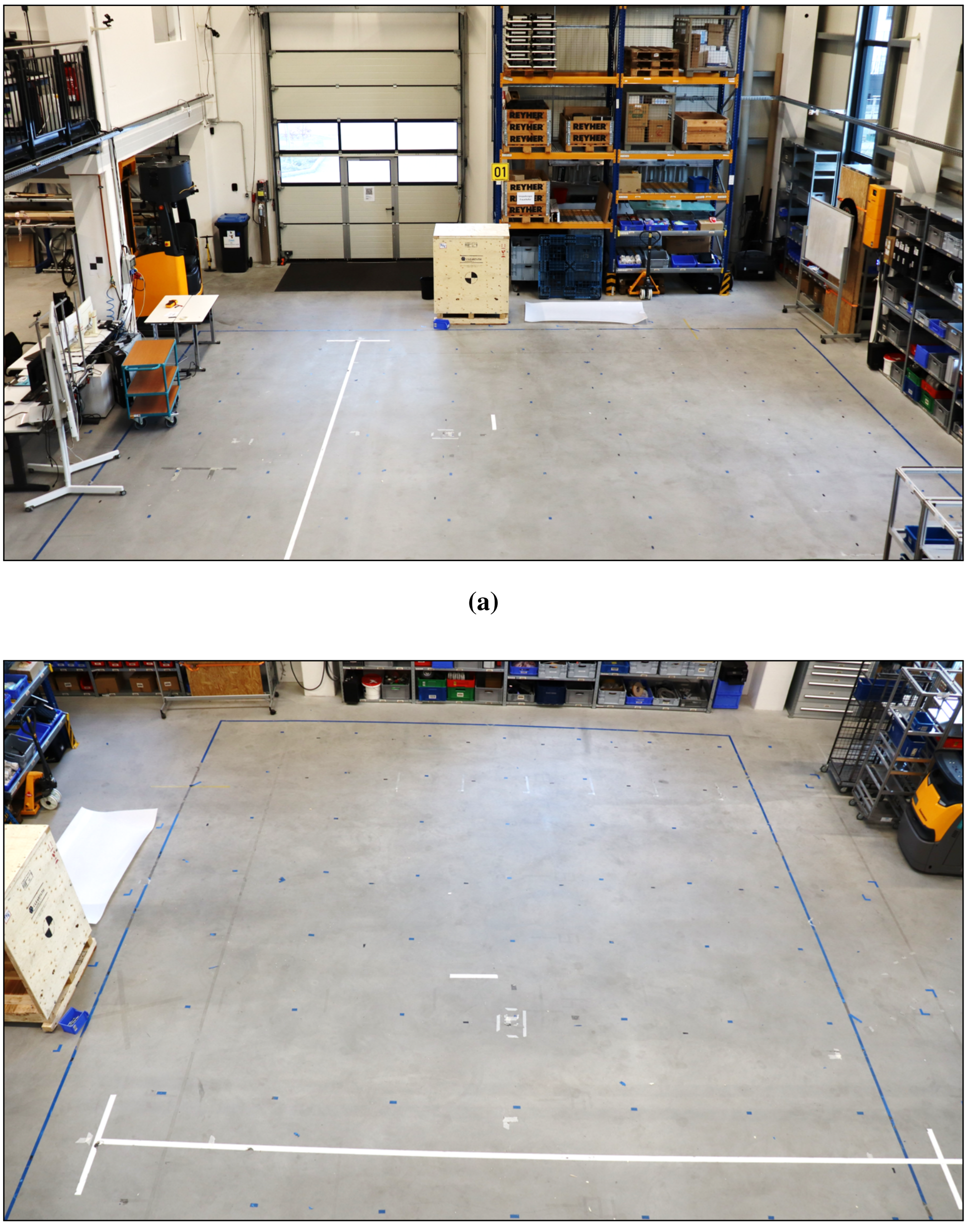
Table 4.2: Characterization of application-driven influencing factors

Factor	Characterization
ELT	Mobile robot
Device/ marker location	Fixed on top of robot
Motion	Maximum velocity limited by typical walking speed
Path	Straights and curves within horizontal plane
Static objects	Open space
Dynamic entities	None
Floor quality	Typical concrete warehouse floor
Radio interference	Bluetooth and WLAN
Amount of reference nodes	Installation instructions
Location of reference nodes	Installation instructions
Parameter configuration	Installation instructions

4.1.3 Experiment Specification

Using the defined test scenario, specific experiments are now outlined following the *Experiment Specification* procedure, beginning with a description of the testbed used for experimentation, encompassing aspects such as the test environment, test volume, and the GT system. Subsequently, the SuTs, the process and the environment influences, as well as the evaluation poses, are specified. Firstly, the testbed components are described as follows.

- The **test environment** for conducting the experiments is the test hall at the ITL of TUHH (Figure 4.2). The hall is characterized by an open space area of approximately 12 m by 20 m with a vertical clearance of 8 m. The area is equipped with various logistic elements, such as pallet shelves, industrial trucks, and a conveyor belt. Figure 4.3 marks



(a)

(b)

Figure 4.2: Test environment at ITL

the locations of key logistic objects within this environment, alongside supplementary elements addressed throughout this section.

- The **test volume**, encompassing a rectangular test area that measures 7 m by 9 m with a height of 3 m, is located within the test environment. This volume is given as the calibrated capture volume of the reference system. Figure 4.3 illustrates its approximate location within the broader environment.

- Concerning the **GT**, the testbed features an optical passive Motion Capture (MoCap)

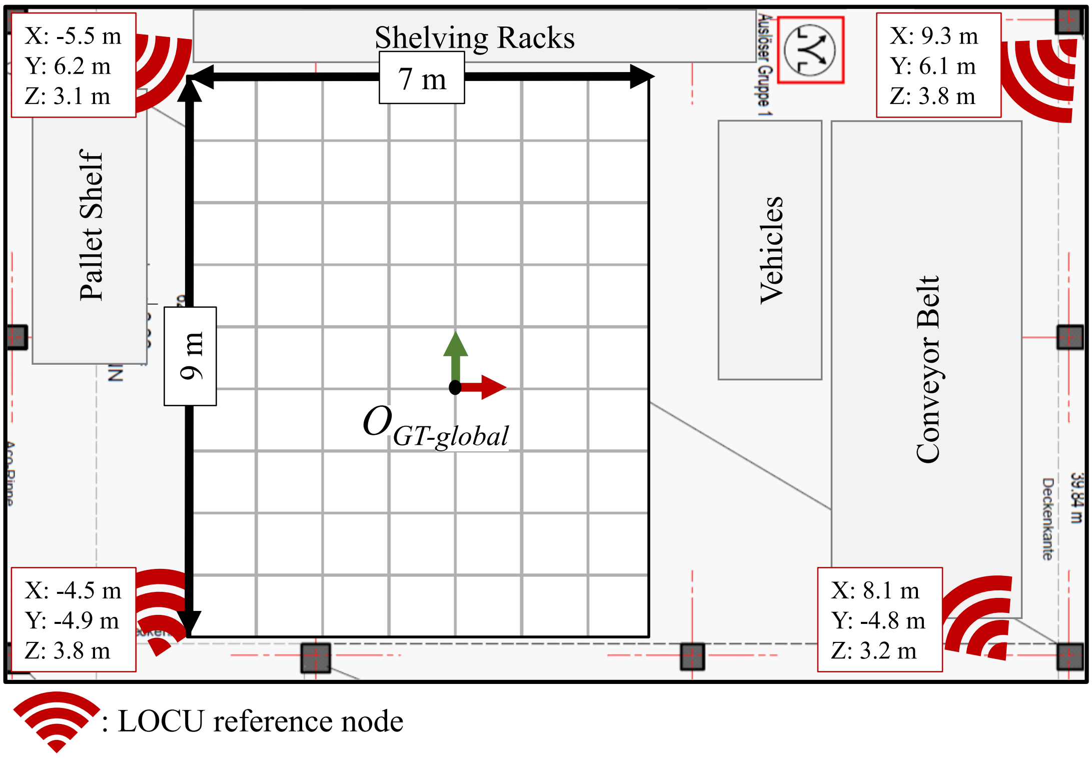


Figure 4.3: Illustration of experiment specification

system, provided by *Qualisys AG*. This system is composed of eight *Miqus M5* infrared cameras, strategically positioned around the test volume at heights ranging from 3 m to 5 m. The deployment and utilization of this system at the ITL is documented in a customer story of *Qualisys AG* and can be accessed online [127]. The MoCap system provides six degrees of freedom (DoF) reference pose and velocity data, noted for its low latency and high absolute positional accuracy within the millimeter range. Its accuracy has been confirmed through various experiments, as well as other scientific studies on similar systems [128, 129]. Figure 4.3 indicates the coordinate origin $O_{GT-global}$ of the calibrated system.

Next, the experiments are further specified under consideration of the test scenarios provided from the *Scenario Definition* procedure. This involves detailing the SuTs and the implementation of application-driven influencing factors as follows.

The LOCU system comprises four reference nodes and a localization tag. The reference nodes are strategically distributed around the test area to ensure effective coverage while avoiding symmetrical layouts. The system self-estimates the positions of these reference nodes through an auto-calibration procedure. These estimated positions are indicated in the floor plan shown in Figure 4.3. During experimentation, these locations serve as references for the system to calculate the horizontal position of the localization tag. Furthermore, the system is configured with an update rate of 8 Hz, while an EKF is employed for data filtering.

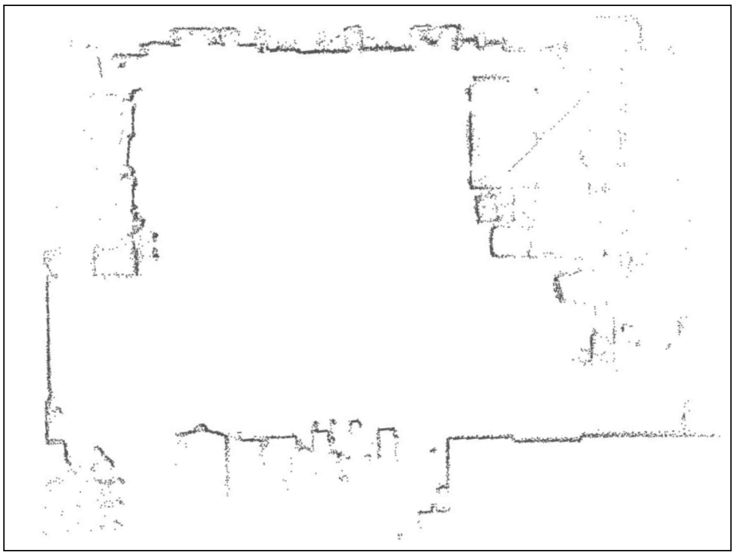


Figure 4.4: Contour map recorded by the LLS system

As for the LLS system, a critical aspect is the existence and quality of a pre-recorded environmental map. This map, crucial for the system's operation, is recorded in advance using the same equipment as for the experiments, while the recording process adheres to the manufacturer's guidelines for the system and is done utilizing the software provided. The resulting map is shown in Figure 4.4. Furthermore, the system uses its entire Field of View (FoV) of 270° and is configured to send location updates with a frequency of 25 Hz.

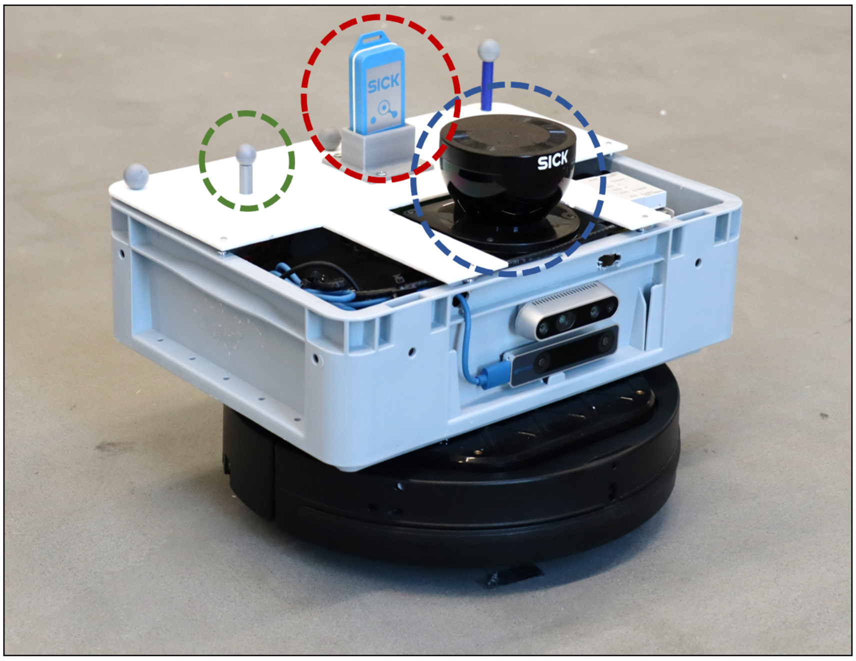


Figure 4.5: TurtleBot2 robotic platform with microScan3 (blue) sensor for the LLS system, localization tag for the LOCU system (red), and MoCap reflectors for GT (green)

Regarding process and environment specifications, the MoCap reflectors and the localization devices of the SuTs are mounted on a TurtleBot2 robotic platform [130] (Figure 4.5). This platform serves as an ELT, representing a mobile robot for the experiment. The robot is programmed for automatic navigation using MoCap data and is limited to a maximum velocity of 1.1 m/s. The periphery of the test area includes various logistic equipment, but the area itself is kept free from static objects. The floor specification and radio interference conditions align with those outlined in the test scenario, where prevalent Bluetooth and WLAN signals are acknowledged but not provoked in particular. The robot's trajectory during the experiment comprises straight lines and curves on a horizontal plane. However, the precise trajectory is determined by the robot's configuration and the list of evaluation poses, detailed in the following. The path is repeated twice to assess localization repeatability, adhering to the *T&E 4Log Benchmarking Procedure*.

The experiment entails 85 evaluation poses sequentially arranged along the predetermined path. As the path lies in a horizontal plane, the evaluation poses are defined with three degrees of freedom (x, y, ψ). A chosen grid length of 0.5 m sets the horizontal position tolerance at 0.25 m. The heading tolerance is specified at 10°, and the orientation of the evaluation poses aligned at multiples of 90°. In the alignment experiment, 36 evaluation poses surround the global origin $O_{GT-global}$, where nine distinct positions are approached four times each from varied directions. Figure 4.6 displays both sets of evaluation poses.

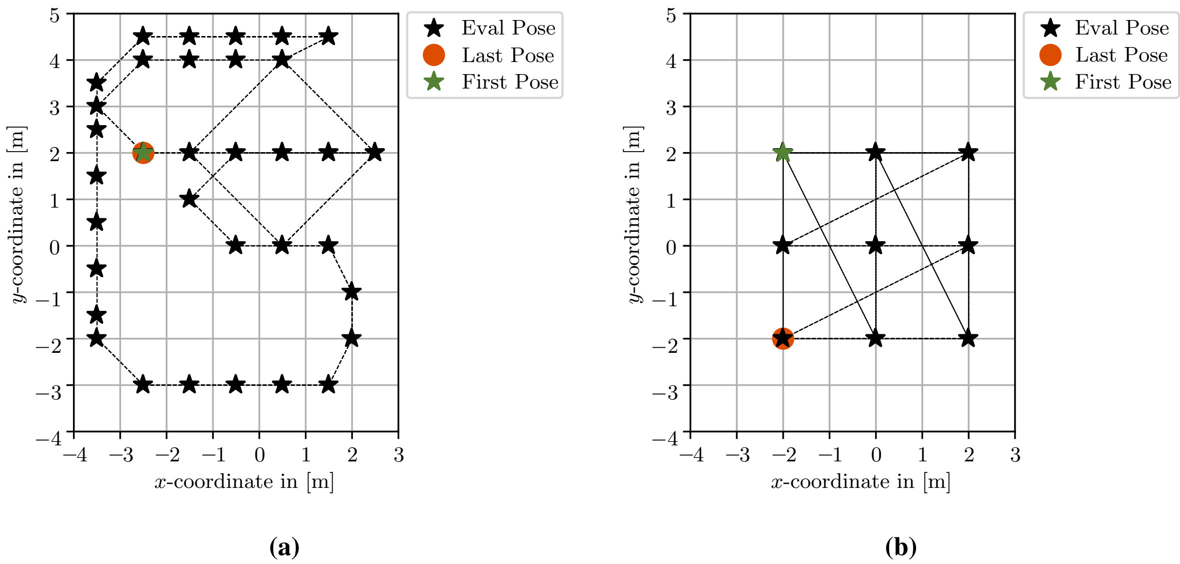


Figure 4.6: Horizontal positions of evaluation poses for (a) experiment and (b) alignment

4.1.4 Experiment Execution

Further following the methods of the *T&E 4Log Benchmarking Procedure*, the experiment is conducted as specified. After calibration and configuration of the SuTs as well as the reference system, and manual alignment of local coordinate frames, time synchronization is achieved through an implementation of PTP [114]. For the LLS system, time synchronization with

MoCap is done via WLAN, achieving a time offset of less than 0.5 ms. Given the maximum velocity of the robot of 1.1 m/s, this translates to an acceptable maximal measurement error of 0.55 mm for the horizontal error components.

Subsequently, the experiments were executed. Data from the SuTs and the reference system are simultaneously collected using the *T&E 4Log App* leveraging Robot Operating Systems (ROS). The ELT automatically navigates along the predefined evaluation poses, ensuring entry into the tolerance space before progressing to the next pose.

Figure 4.7 (a) presents the experiment data (E_{GT} and E_{SuT}), while Figure 4.7 (b) displays the alignment data ($E_{GT,align}$ and $E_{SuT,align}$) for both SuTs. A notable observation is the displacement between the MoCap and the LLS trajectories, indicating a suboptimal global alignment between the system's coordinate frames. Conversely, the LOCU system exhibits better alignment, attributable to its configuration process. These datasets, combined with the evaluation poses, form the basis for calculating performance metrics as part of the subsequent *Performance Evaluation*.

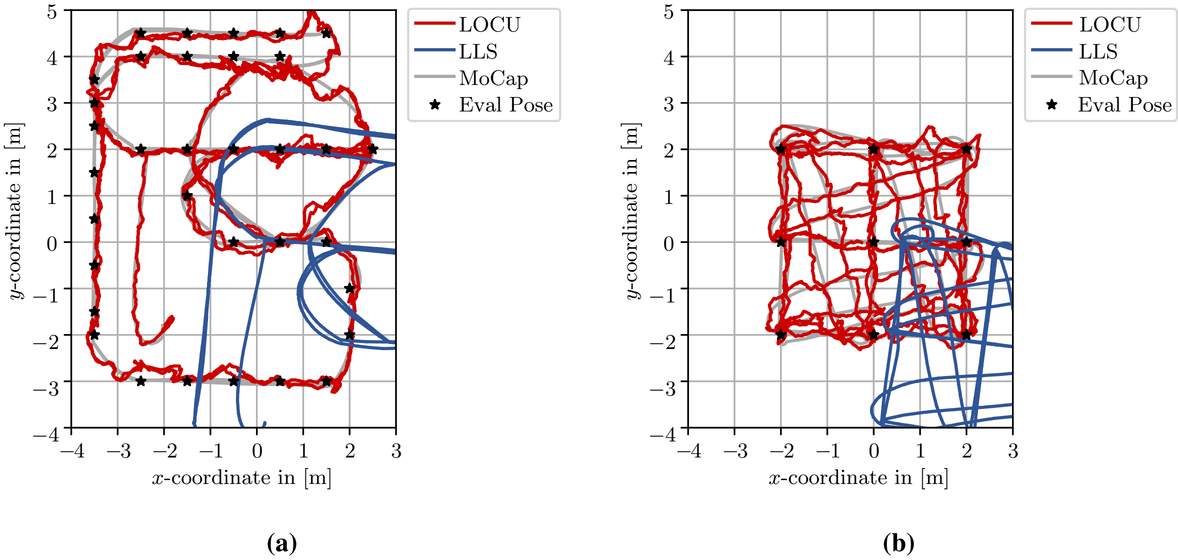


Figure 4.7: Horizontal position of (a) experiment data and (b) alignment data (LLS and LOCU)

4.1.5 Performance Evaluation

The focus of the *Performance Evaluation* procedure lies in computing various metrics to assess the systems' performance in terms of localization accuracy, repeatability, latency, and update rate. The methods described in Section 3.3.7 are applied to the experiment and alignment data for both systems obtained by the experiments in the previous step. Since the location data from both systems are limited to the horizontal plane, the evaluation focuses on pose data encompassing three DoFs (x, y, ψ).

The initial phase of experiment data processing involves linear interpolation for timestamp alignment, followed by the application of the *closestPoseAssociation* algorithm to associate reference and location data points from each dataset with corresponding evaluation poses. Figure 4.8 shows the resulting data points for the LLS alignment data. This dataset is subsequently utilized to compute both global and local alignments.

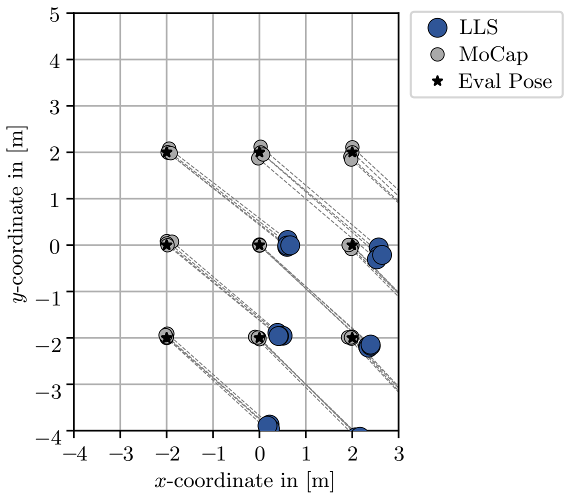


Figure 4.8: Data points of alignment experiment associated with evaluation poses (LLS)

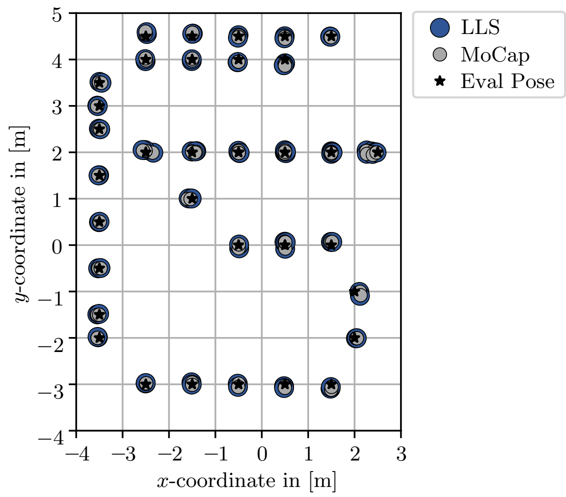


Figure 4.9: Spatially aligned experiment data points associated with evaluation poses (LLS)

The alignment computation is performed as described in Section 3.3.7, resulting in global alignment transformations ($A_{global,LLS}$, $A_{global,LOCU}$) and local alignment transformations ($A_{local,LLS}$, $A_{local,LOCU}$). For the LOCU system, which provides only horizontal position data, the identity matrix is used to neglect local rotations. The alignments are represented as homogeneous transformation matrices with the translations provided in [m] as follows:

$$A_{global,LLS} = \begin{bmatrix} 0.996 & -0.094 & -2.564 \\ 0.094 & 0.996 & 1.939 \\ 0 & 0 & 1 \end{bmatrix}; A_{global,LOCU} = \begin{bmatrix} 0.999 & 0.003 & 0.015 \\ -0.003 & 0.999 & -0.005 \\ 0 & 0 & 1 \end{bmatrix} \tag{4.1}$$

$$A_{local,LLS} = \begin{bmatrix} 0.999 & 0.029 & -0.026 \\ -0.029 & 0.999 & 0.004 \\ 0 & 0 & 1 \end{bmatrix}; A_{local,LOCU} = \begin{bmatrix} 1.000 & 0.000 & 0.024 \\ 0.000 & 1.000 & 0.027 \\ 0 & 0 & 1 \end{bmatrix} \tag{4.2}$$

The rotational and translational components can be extracted from these transformation matrices. For instance, the local coordinate transformation of the LLS system represents a rotation of approximately 2.56° around the z-axis, followed by a translation of -0.026 m along the x-axis and -0.004 m along the y-axis, as observed from the ELT's viewpoint. Hence, these values offer a systematic estimation of the deviations resulting from the manual alignment of the local coordinate frames.

Next, local and global alignment transformations are applied to the experiment data. Figure 4.9 presents the spatially aligned experiment data points associated with evaluation poses for the LLS system, indicating the successful determination and application of the global alignment. In addition, Figure 4.10 (a) and (b) illustrate the position error in the x and y directions, before

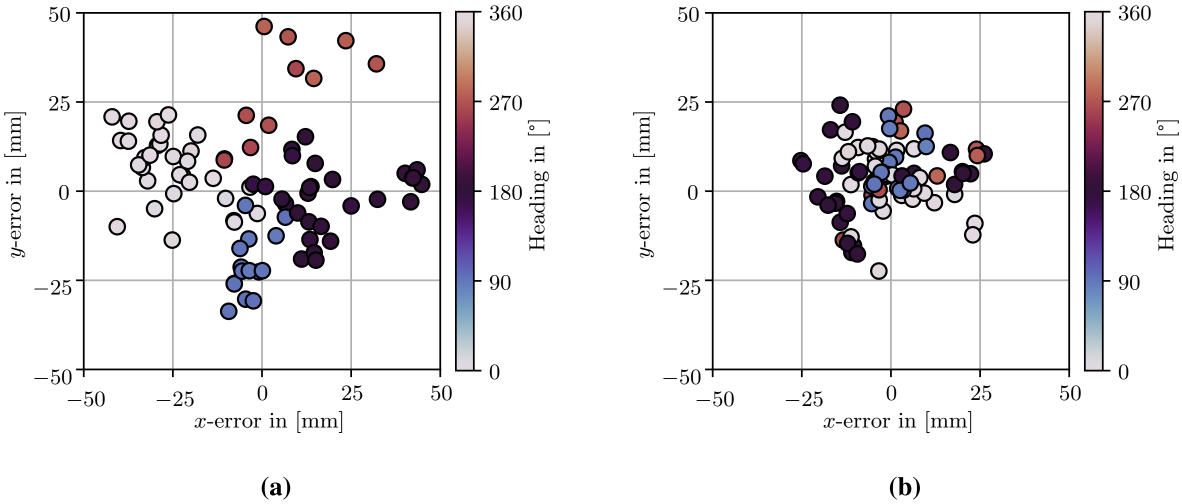


Figure 4.10: Horizontal position error plot (a) before local alignment and (b) after local alignment (LLS)

and after applying the local alignment transformations. The more uniform distribution of data points observed after local alignment demonstrates the effective mitigation of systematic errors, related to the ELT's heading direction, achieved through local alignment.

Finally, performance metrics are determined. To compute metrics related to the systems' accuracy, the horizontal position errors, and for the LLS system, also the absolute heading errors are calculated for each evaluation pose. For the LLS system, these are shown in Figure 4.11. Similarly, the horizontal localization repeatability and absolute heading repeatability are computed for each repeatedly approached evaluation pose, utilizing the *ComputeRepErrorData* algorithm. Furthermore, the outlined methods are employed to calculate the time delay and time gap for each evaluation pose. The resulting Comulative Distribution Functions (CDFs) are depicted in Figure 4.12. This commonly employed representation facilitates a comparative analysis while allowing for an understanding of the data distribution patterns.

The data is used to calculate the proposed performance metrics with the results summarized in Table 4.3. The percentiles (P_{95} to $P_{6\sigma}$) can be extracted from the CDFs by interpolation. From these results, several observations can be made regarding the systems' performance. For example, the computed horizontal localization accuracy and repeatability of the LLS system

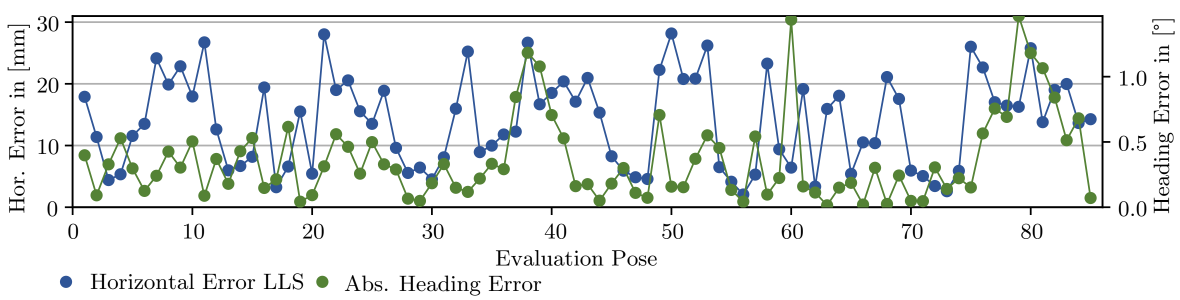


Figure 4.11: Horizontal position error and absolute heading error corresponding to evaluation pose (LLS)

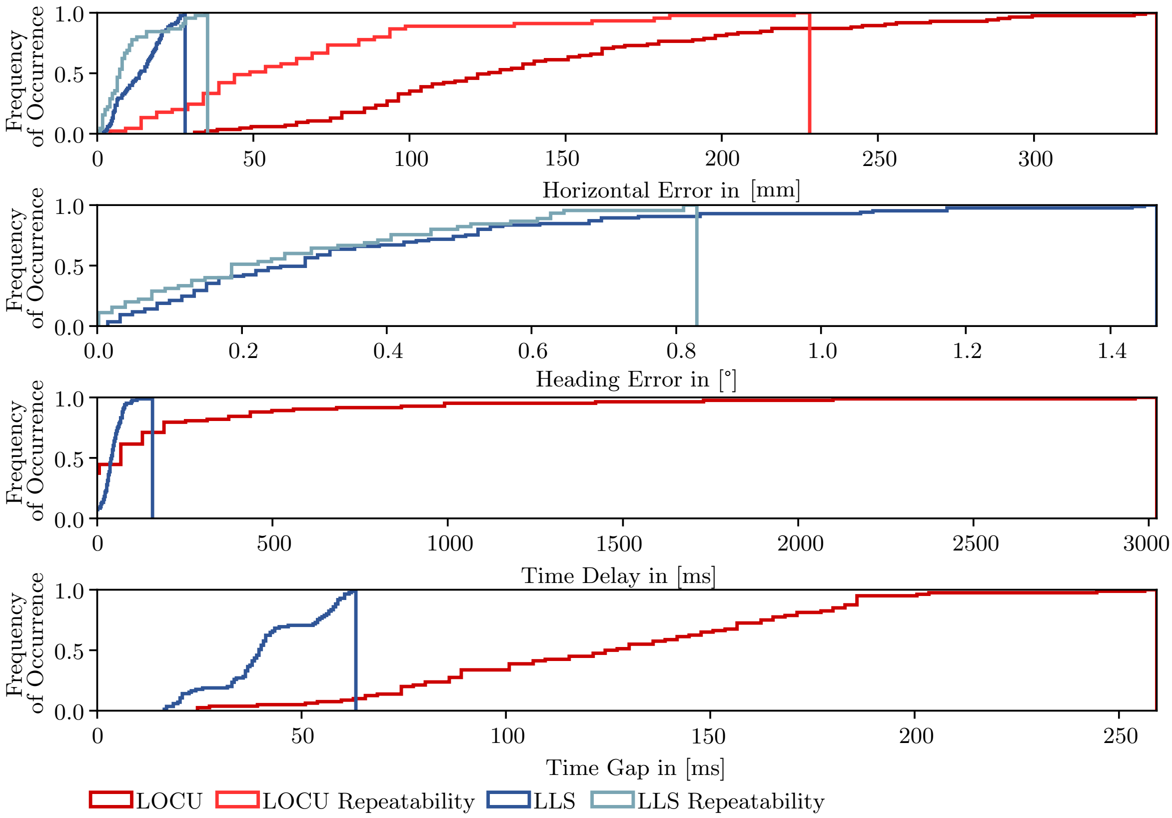


Figure 4.12: CDFs characterizing localization accuracy, repeatability, time gap, and time delay (LLS and LOCU)

Table 4.3: Summary of performance metrics. The bold numbers are referred to in the text

	Mean	σ	**Median**	P_{95}	$P_{2\sigma}$	$P_{3\sigma}$	$P_{4\sigma}$	$P_{5\sigma}$	$P_{6\sigma}$
				LLS					
Horizontal acc. in [mm]	14	7	14	**26**	19	26	28	28	28
Horizontal rep. in [mm]	10	9	7	**29**	10	29	34	35	35
Abs. heading acc. in [°]	0.36	0.32	0.29	**1.08**	0.43	0.93	1.45	1.46	1.46
Abs. heading rep. in [°]	0.28	0.23	0.20	**0.65**	0.39	0.64	0.83	0.83	0.83
Time delay in [ms]	44	30	**42**	84	54	80	136	157	158
Time gap in [ms]	41	**13**	**40**	61	45	60	63	63	63
				LOCU					
Horizontal acc. in [mm]	144	70	128	**290**	163	276	337	339	339
Horizontal rep. in [mm]	65	48	50	**176**	75	163	217	228	228
Time delay in [ms]	141	**605**	**93**	1042	150	943	2561	3006	3023
Time gap in [ms]	138	**70**	**132**	297	162	252	367	410	412

are in the low centimeter range, with a P_{95} of 26 mm and 29 mm, respectively. In contrast, the respective characteristics for the LOCU system are in the lower decimeter range, with P_{95} values of 290 mm and 176 mm. Furthermore, with few exceptions, the plots indicate that the localization repeatability of each system is lower than that of its associated localization accuracy component. This is also true for the absolute heading accuracy and repeatability of the LLS system, which is characterized by P_{95} values of 1.08° and 0.65°, respectively. This outcome is anticipated, as location-specific systematic errors of the systems are mitigated when determining localization repeatability.

Regarding the measured time delay, the LLS data being mostly in the positive range indicates that an effect of the system's latency on the measured position error is measured. This effect is assumed to be constant, characterizing the systematic influence of the system's latency. Since it is more robust to outliers than the mean, the median value is used to characterize the system's latency. The computed median of 0.0416 s is close to the minimal configurable update time of 0.04 s, which is limited by the time that the system needs to compute the pose, indicating the successful computation of the system latency. Regarding the LOCU system, a positive time delay, characterized by a median of 0.093 s is computed with a standard deviation of 0.6 s, indicating a strong scattering of the data.

As for the measured time gap, the LLS system's median value of 40 ms aligns precisely with the pre-configured update rate of 25 Hz. The variability of the measured data is characterized by a standard deviation of 13 ms. The observed time gap of the LOCU system shows a slight deviation from the configured value of 8 Hz, with a median of 132 ms and a standard deviation of 70 ms.

Although the measured time delay and time gap for each of both systems are not constant, for further evaluation, they are assumed to be constant, represented by their respective median values. This assumption, despite observed variability, simplifies the analysis, but requires a careful consideration of its impact on the overall evaluation of the system.

Together, the overview of performance metrics in Table 4.3 and the provided CDFs in Figure 4.12 offer comprehensive insight into the performance of the systems. Moreover, the experiment data and evaluation results presented here will be utilized, along with the findings from subsequent experimental studies, to discuss the validity of the framework in Chapter 5.

4.1.6 Requirement Specification

The application processes "pallet pick-up", "pallet transport", and "pallet placement", previously outlined in the *Application Description*, form the basis for applying the methods outlined within the *Requirement Specification* procedure. In the following, each of these application processes is further examined to derive specific location data requirements, expressed as an individual equation. Table 4.4 provides an overview of the results.

Pallet Pick-up In the "pallet pick-up" process, a key activity involves the robot approaching the pallet's location. Crucially, the robot must be accurately positioned in front of the correct pallet to ensure its detection within the sensors' FoV. Hence, the localization function is to confirm the robot's presence within this interest space. The interest space measures 2 m by 1.6 m, with a 20° orientation range (ψ). The entity's interest frame aligns with its localization

Table 4.4: Outcome of the *Requirement Specification* procedure for different localization functions of identified application processes

Process	**"Pallet pick-up"**	**"Pallet transport"**	**"Pallet placement"**
Process step	Approach pallet	Plan path	Align with drop-off point
Localization function	Confirm presence in front of pallet	Identify aisle and heading direction	Ensure pallet is dropped within designated area
Interest space	Space in front of the pallet with dimensions given by (2 m, 1.6 m, 20°). Location is given in reference to the localization system's coordinates	Positional component of 5.5 m limited by the distance between adjacent aisles; orientation component is given as 180° to determine heading direction; given in absolute coordinates	Dimensions given by (0.4 m, 0.8 m, 10°); positional components limited by placement area of (1.6 m x 1.6 m) and pallet dimensions (1.2 m x 0.8 m); given in absolute coordinates
Entity's interest frame	Equivalent to entity's localization frame	Equivalent to entity's localization frame	Given by point between the robot's fork that reflects the center of the pallet
Motion space	Robot's inaccuracy in approaching a given coordinate, given by (0.2 m, 0.2 m, 4°)	Positional component of 1.2 m, limited by the distance between shelves of 2.4 m and the robot's width of 1.2 m. Orientation component given by 160°	Robot's inaccuracy in approaching a given coordinate, given by dimensions (0.2 m, 0.3 m, 4°)
Safety margin	0.1 m in horizontal directions and 3° heading direction	None	0.05 m in horizontal directions and 1° heading direction
Velocity	0.2 m/s linear velocity and 10 °/s rotational velocity	1 m/s maximum linear velocity and 30° rotational velocity	0.2 m/s in x; 0.4 m/s in y; 10 °/s rotational velocity
Real-time capability	yes	yes	yes
Network time	Not relevant	0.1 s	Not relevant
Confidence level	High (5σ)	Moderate (4σ)	High (5σ)

frame in the robot's center. Furthermore, it is assumed that the pallet's coordinates, and thereby indirectly the interest space coordinates, have been set during "pallet placement" by the same or another robot referring to the same reference coordinate system as the localization system. Hence, localization repeatability is considered the decisive criterion.

Figure 4.13 illustrates the positional components of the relevant spaces. The motion space, signifying the robot's accuracy in approaching a specific coordinate, is situated at the center of the interest space. This space encompasses a horizontal dimension of 0.2 m and a heading orientation of 4°. Furthermore, a safety margin of 0.1 m for horizontal movements and 3° for orientation is considered. Consequently, the requirement margin $\vec{R}_{lf1}$, given as the minimal distance between the boundary of the motion space to the boundary of the interest space, is established as (0.8 m, 0.6 m, 3°).

The determination of the effective localization error necessitates additional process parameters. Hence, it is assumed that while approaching the pallet, the robot's velocity is restricted to a maximum of 0.2 m/s linearly and 10 °/s rotationally. Real-time response capabilities are essential. However, network transmission times are negligible since data processing is performed on board the robot. A high confidence level of 5σ is selected to mitigate the substantial

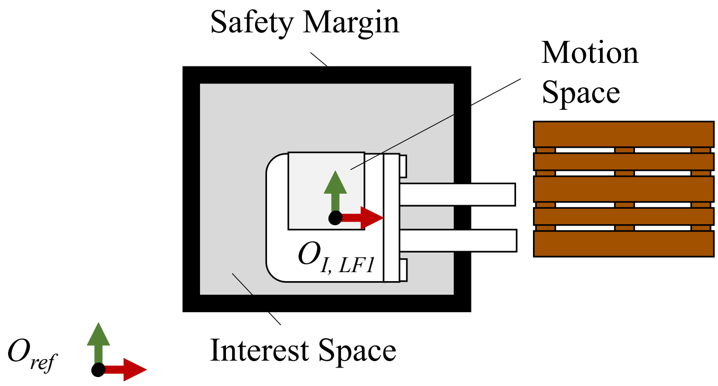


Figure 4.13: Illustration of horizontal spaces considered for process step "approach pallet"

follow-up costs associated with picking up an incorrect pallet. In conclusion, the location data requirements for the process "pallet pick-up" are given by inserting the discussed values into equation 3.38:

$$\begin{pmatrix} 0.8\,\mathrm{m} \\ 0.6\,\mathrm{m} \\ 3° \end{pmatrix} \geq \vec{\epsilon}_{5\sigma,Rep,loc} + \begin{pmatrix} 0.2\,\mathrm{m/s} \\ 0.2\,\mathrm{m/s} \\ 10\,°/\mathrm{s} \end{pmatrix} \left(\frac{1}{f_{update}} + t_{lat}\right) \tag{4.3}$$

Pallet Transport For the application process "pallet transport", the focus is on the activity of global path planning within a warehouse aisle. The aim of finding the fastest path to a desired goal requires the robot's awareness of its current aisle and orientation. This particular localization function was exemplified in Section 3.3.8, with Figure 3.20 illustrating the determination of the requirement margin. However, here the discussion delves into a concrete example by discussing and quantifying each element required to specify the location data requirements for the considered case study.

For this process, the positional requirements in the y-direction, perpendicular to the aisle, are determined by the 5.5 m distance between adjacent storage aisles. The entity's interest and localization frames are assumed to be centered within the robot, with the interest space defined in absolute coordinates based on the positions of the shelves. The positional requirement in y-direction is considered critical. As such, the x-direction is neglected. Furthermore, the heading direction is critical for the interest space, set at 180° to establish the robot's facing direction. The motion space in y-direction is limited to 1.2 m due to the aisle width of 2.4 m and the robot's width of 1.2 m, with a heading component of 160° allowing for a 80° angle either way during which the robot is tolerated to be guided in the opposite aisle direction. Consequently, the motion space is defined by the dimensions (1.2 m, 160°). The safety margin is neglected, resulting in a requirement margin $\vec{R}_{lf2}$ of (2.15 m, 10°).

Regarding the process parameters to estimate the effective localization error, the robot's maximum velocity is set at 1 m/s linearly and 30 °/s rotationally around the z-axis. Path planning, involving multiple entities, is processed centrally. Hence, since real-time capabilities are required, network transmission times are considered at 0.1 s. Considering the less critical nature of follow-up costs in global path planning, a moderate confidence level of 4σ is considered appropriate. These considerations collectively lead to the following condition that a localization system must meet to be considered suitable for this particular localization function:

$$\begin{pmatrix} 2.15\,\mathrm{m} \\ 10° \end{pmatrix} \geq \vec{\epsilon}_{4\sigma,Acc,loc} + \begin{pmatrix} 1\,\mathrm{m/s} \\ 30\,°/\mathrm{s} \end{pmatrix} \left(\frac{1}{f_{update}} + t_{lat} + 0.1\,\mathrm{s}\right) \tag{4.4}$$

Pallet Placement The fine alignment step in the “pallet placement” process focuses on ensuring the accurate placement of the pallet within a designated area with established absolute coordinates. The entity’s interest frame is pinpointed at the horizontal midpoint between the robot’s forks, mirroring the pallet’s center when loaded. The distance from the entity’s localization frame to its interest frame is approximated to be 1.2 m along the robot’s heading direction. To ensure that the entire pallet is correctly placed within the designated area, the positional components of the interest space are determined by deducting the pallet’s dimensions from the placement area’s dimensions, as illustrated in Figure 4.14. With a placement area measuring (1.6 m x 1.6 m) and pallet dimensions of (1.2 m x 0.8 m), the resultant positional components of the interest space are (0.4 m, 0.8 m). Additionally, a rotational tolerance of 5° around the z-axis is accounted for in either direction.

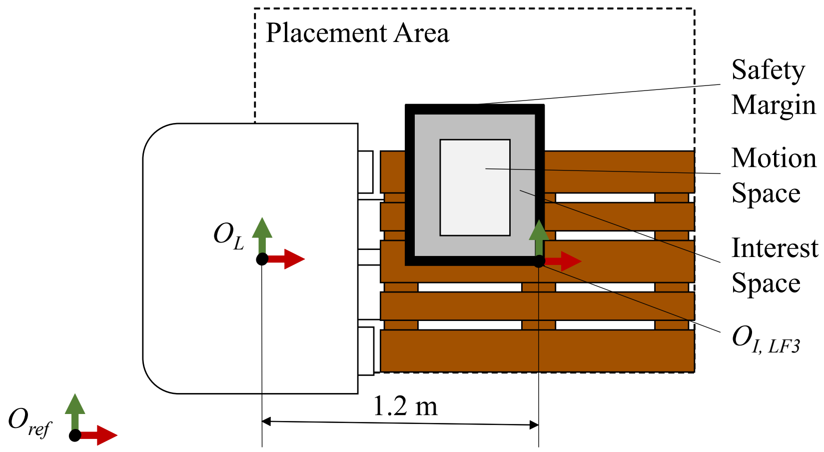


Figure 4.14: Illustration of horizontal spaces considered for process step “fine alignment”

The motion space for this step is defined by the robot’s maneuvering ability, similar to the “pallet pick-up” process. However, due to rotational inaccuracies in the entity’s localization frame, the motion space, measuring (0.2 m, 0.3 m, 4°), is comparably larger in the y-direction. An additional safety margin of (0.05 m, 0.05 m, 1°) is factored in, yielding a requirement margin $\vec{R}_{lf3}$ of (0.05 m, 0.2 m, 1°).

The assumed velocity within the localization frame mirrors the one of the “approach pallet” process. However, considering the velocity in the entity’s interest frame, the rotational velocity at the localization frame induces additional linear velocity. Therefore, the maximum linear velocity is estimated at 0.5 m/s in y-direction. Network transmission times are neglected, as location information is processed on the robot. Finally, a high confidence level of 5σ is taken into account to ensure reliable operation, resulting in the following equation:

$$\begin{pmatrix} 0.05\,\mathrm{m} \\ 0.2\,\mathrm{m} \\ 1^\circ \end{pmatrix} \geq \vec{\epsilon}_{5\sigma,Acc,int} + \begin{pmatrix} 0.2\,\mathrm{m/s} \\ 0.4\,\mathrm{m/s} \\ 10\,^\circ/\mathrm{s} \end{pmatrix} \left(\frac{1}{f_{update}} + t_{lat}\right) \tag{4.5}$$

Equations 4.3, 4.4, and 4.5, corresponding to the outlined application processes, represent the outcomes of the *Requirement Specification* process. These equations are the criteria for assessing localization system suitability in the upcoming section.

4.1.7 System Evaluation

The final phase in the *T&E 4Log Framework* involves assessing the suitability of the systems. This is accomplished by comparing the location data requirements for each localization function defined in the *Requirement Specification* with the corresponding performance metrics from the *Performance Evaluation*.

The relevant performance metrics for Equations 4.3, 4.4, and 4.5 have been identified as $\vec{\epsilon}_{5\sigma,Rep,loc}$, $\vec{\epsilon}_{4\sigma,Acc,loc}$, $\vec{\epsilon}_{5\sigma,Acc,int}$, t_{lat}, and t_{gap}. The parameters describe the accuracy within the entity's localization frame, thus, allowing the values for the "pallet pick-up" ($\vec{\epsilon}_{5\sigma,Rep,loc}$) and for the "pallet transport" processes ($\vec{\epsilon}_{4\sigma,Acc,loc}$) to be taken directly from the performance metrics provided in Table 4.3 for each of both systems. The percentiles for horizontal position accuracy are utilized for the direction in x and y.

In the "pallet placement" process, the entity's interest frame differs from its localization frame. Therefore, the accuracy within the interest frame is calculated by adding the one caused by rotational errors to the positional accuracy in the entity's localization frame. The vector $\vec{\epsilon}_{5\sigma,\mathrm{acc,int}}$ is determined as follows, with the distance d denoting the linear distance between the frames along the robot's heading direction (depicted in Figure 4.14):

$$\vec{\epsilon}_{5\sigma,Acc,int} = \begin{pmatrix} P_{x,5\sigma,Acc} \\ P_{y,5\sigma,Acc} \\ P_{\psi,5\sigma,Acc} \end{pmatrix} + \begin{pmatrix} (1 - \cos(P_{\psi,5\sigma,Acc})) \cdot d \\ \sin(P_{\psi,5\sigma,Acc}) \cdot d \\ 0^\circ \end{pmatrix} \tag{4.6}$$

Using the performance metrics presented in Table 4.3 and considering 1200 mm for the distance d, the relevant localization error for the LLS system is determined as (28 mm, 59 mm, 1.46°).

For the LOCU system, which does not provide rotational data, the full distance d between the coordinate frames must be considered as an additional error. Table 4.5 presents a summary of the pertinent values, reflecting the systems' performance for the examined localization functions. In addition to the vectors $\vec{\epsilon}_{5\sigma,Rep,loc}$, $\vec{\epsilon}_{4\sigma,Acc,loc}$, and $\vec{\epsilon}_{5\sigma,Acc,int}$, the table includes the median values for the time gap and time delay.

Table 4.5: Relevant performance metrics for determining system suitability

	$\vec{\epsilon}_{5\sigma,\mathbf{rep,loc}}$	$\vec{\epsilon}_{4\sigma,\mathbf{acc,loc}}$	$\vec{\epsilon}_{5\sigma,\mathbf{acc,int}}$	$t_{\mathbf{lat}}$	$t_{\mathbf{gap}}$
LLS	$\begin{pmatrix}35\,\text{mm}\\35\,\text{mm}\\0.83°\end{pmatrix}$	$\begin{pmatrix}28\,\text{mm}\\28\,\text{mm}\\1.45°\end{pmatrix}$	$\begin{pmatrix}28\,\text{mm}\\59\,\text{mm}\\1.46°\end{pmatrix}$	42 ms	40 ms
LOCU	$\begin{pmatrix}228\,\text{mm}\\228\,\text{mm}\\-\end{pmatrix}$	$\begin{pmatrix}337\,\text{mm}\\337\,\text{mm}\\-\end{pmatrix}$	$\begin{pmatrix}1539\,\text{mm}\\1539\,\text{mm}\\-\end{pmatrix}$	93 ms	132 ms

With all the necessary parameters established for Equations 4.3, 4.4, and 4.5, the systems' suitability can finally be assessed. Table 4.6 contains the resulting effective localization errors, highlights the respective requirement margins, and concludes component-wise whether each system meets the necessary conditions for suitability. However, if a system fails to meet the condition for even one component, it is considered not suitable. Accordingly, the LLS system is deemed suitable for the processes "pallet pick-up" as well as "pallet transport". However, for the "pallet placement" process, it does not meet the stringent requirements for heading accuracy. As the LOCU system lacks rotational data, it is considered not suitable for all of the assessed localization functions.

Table 4.6: Overview of requirement margin, and effective localization errors including system suitability (green: requirement met; red: requirement not met) for each considered application process and SuT

Process	$\vec{R}(I,M,S)$	**LLS** $\vec{\epsilon}_{eff}$	**LOCU** $\vec{\epsilon}_{eff}$
"Pallet pick-up"	$\begin{pmatrix}800\,\text{mm}\\600\,\text{mm}\\3°\end{pmatrix}$	$\begin{pmatrix}51\,\text{mm}\\51\,\text{mm}\\1.65°\end{pmatrix}$	$\begin{pmatrix}273\,\text{mm}\\273\,\text{mm}\\\times\end{pmatrix}$
"Pallet transport"	$\begin{pmatrix}2150\,\text{mm}\\10°\end{pmatrix}$	$\begin{pmatrix}210\,\text{mm}\\6.91°\end{pmatrix}$	$\begin{pmatrix}662\,\text{mm}\\\times\end{pmatrix}$
"Pallet placement"	$\begin{pmatrix}50\,\text{mm}\\200\,\text{mm}\\1°\end{pmatrix}$	$\begin{pmatrix}44\,\text{mm}\\92\,\text{mm}\\2.28°\end{pmatrix}$	$\begin{pmatrix}1584\,\text{mm}\\1629\,\text{mm}\\\times\end{pmatrix}$

Overall, this case study demonstrated the utilization of the *T&E 4Log Framework*, providing an empirical basis for discussions on the validity of the proposed concepts and the applicability and utility of the framework.

4.2 Study of Experiment Repeatability

One of the key stakeholder demands for a T&E methodology, pointed out in Section 3.2, is the experiment repeatability. If this profound characteristic is not given, comparability, replicability, reproducibility, and transferability are compromised. Hence, this section is dedicated to analyzing the experiment repeatability. Initially, the conducted experiments are briefly introduced. The performance results for each experiment are then computed and visualized. This section concludes with an analysis of the performance results.

4.2.1 Experiments

The experiments outlined in the previously presented case study have been carried out a total of ten times for each of the two systems, resulting in a total of twenty sets of experiment data. These repetitions were carried out on the same day. Following the criteria for experiment repeatability, the experiments adhered to the same experiment specification and preparation. For each SuT, The same alignment transformation is used for all experiments. Hence, alignment data are only recorded once.

4.2.2 Performance Results

Performance results are obtained using the methods for the *Performance Evaluation* implemented in the *T&E 4Log App*. This generates ten separate data distributions for each performance characteristic and system. Details of the evaluation process have already been extensively shown in Section 4.1.5. Therefore, this section will focus on presenting the results.

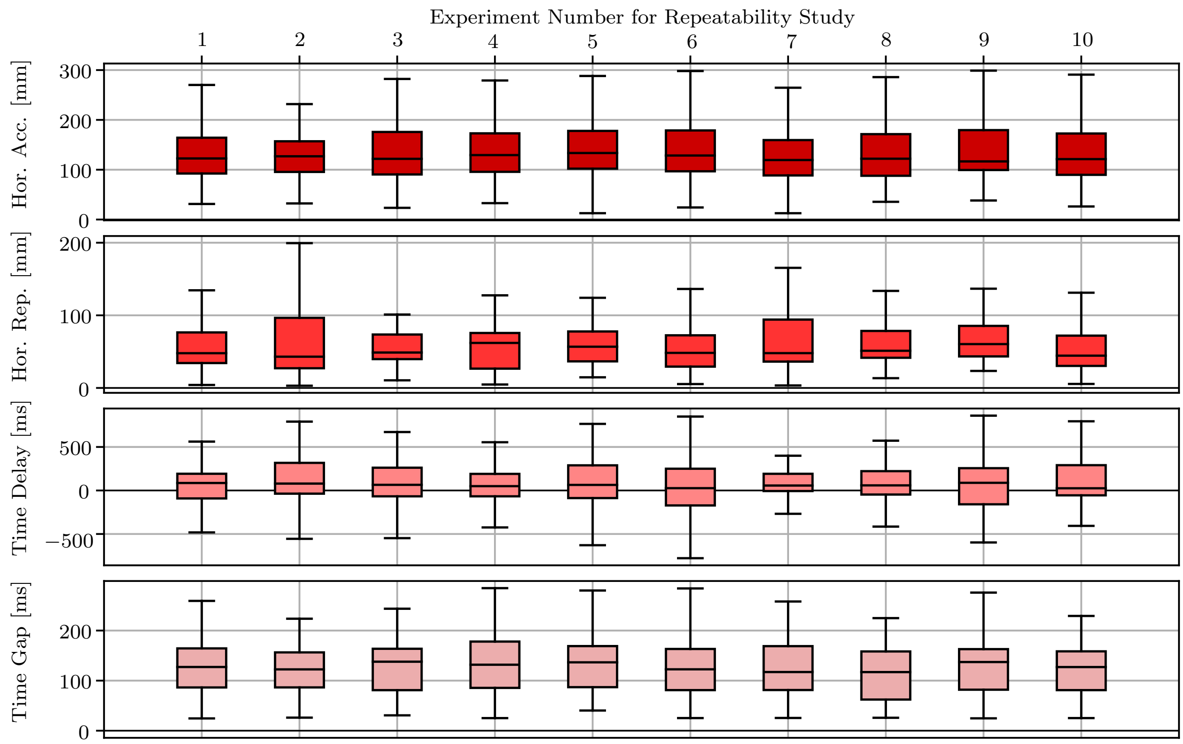


Figure 4.15: Whisker plots with performance characteristics for experiments of the repeatability study (LOCU)

To present the data from various experiments in a comprehensive and comparable format, the performance data, previously depicted as cumulative distribution functions, are now represented as whisker plots. Each whisker plot displays the median and percentiles at the levels 1 %, 25 %, 75 %, and 99 %. The results for the LOCU system are shown in Figure 4.15, and those for the LLS system in Figure 4.16. On the shared x-axis, the experiment numbers are assigned, with LOCU experiments numbered from 1 to 10 and LLS experiments from 11 to 20. Consequently, each subplot within a figure illustrates the data distribution for a specific performance characteristic, segmented by experiment.

For Figure 4.15, focusing on the LOCU system. The whisker plots for horizontal accuracy across the experiments show a consistent value range of 0 mm to 300 mm, with medians between 100 mm and 140 mm. For localization repeatability, the data generally indicate lower values with medians between 40 mm and 70 mm. The time delay measurements exhibit a wide range, spanning from almost −1000 ms to 1000 ms, yet the median values consistently lie between 25 ms and 100 ms. Lastly, the measurement of the time gap shows great uniformity, with median values between 112 ms and 122 ms.

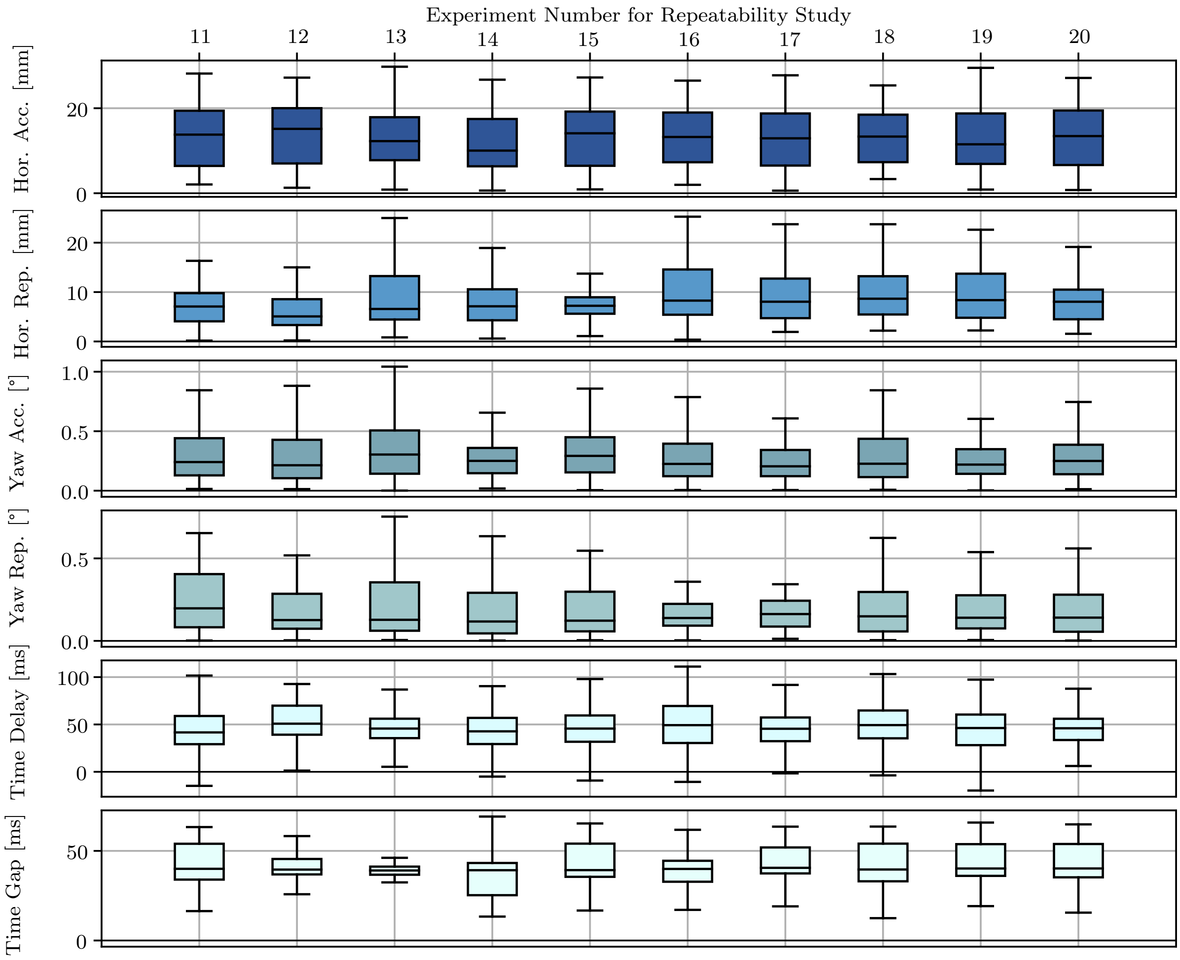


Figure 4.16: Whisker plots with performance characteristics for experiments of the repeatability study (LLS)

Subsequently, Figure 4.16 is considered, presenting the data for the LLS system. Its horizontal accuracy plots, ranging from 0 mm to 30 mm with medians between 10 mm and 15 mm, reveal notably lower errors than the LOCU system. The horizontal localization repeatability exhibits even lower median values, from 5 mm to 10 mm. Additionally, the LLS system's absolute heading accuracy and repeatability are analyzed. These plots show errors below 1° for accuracy and below 0.8° for localization repeatability. Both characteristics show narrow ranges with median values approximately between 0.2° and 0.35° for accuracy and between 0.1° and 0.2° for repeatability. The time delay measurements for the LLS system range from −10 ms to 110 ms, with median values between 40 ms and 50 ms. The time gap measurements show significant variations, ranging from 20 ms to 70 ms, with medians differing by sub-millisecond amounts.

This section's performance results supplement the previously established empirical basis for the upcoming discussion on the framework's validity. Concerning the repeatability of experiments, it is evident that the data distribution varies across experiments, depending on both the system and the performance characteristic in question. Nevertheless, the observed variations in the value ranges, particularly in the median values, are relatively low, suggesting a degree of repeatability for both systems across all evaluated performance characteristics. The following section delves into the analysis of the performance results presented using statistical methods.

4.2.3 Analysis of Results

This section aims to assess the hypothesis that no significant difference exists in the distribution of performance data across each characteristic and both systems, suggesting that the experiments are repeatable.

A widely used statistical method to test the repeatability of experiments by examining the significance of differences between group means is the so-called Analysis of Variance (ANOVA) [131]. However, for ANOVA to be applicable, the data must follow a normal distribution. Thus, a preliminary step involves verifying the normality of the data for each experiment and performance characteristic, previously visualized as whisker plots.

The *SciPy* implementation of the Kolmogorov-Smirnov test is used to assess the type of distribution of performance data [132, 133]. This test yields a p-value based on the chi-squared statistic, which reflects the degree to which the data deviates from a specified distribution. The p-value quantifies the likelihood of observing such a deviation if the null hypothesis that the data come from a particular distribution holds. Typically, this hypothesis is rejected for a p-value of less than 0.05. Table 4.7 summarizes the number of data sets for each performance characteristic that conform to a normal distribution, with a total of ten experiments for each system. For example, a count of five would mean that normal distribution characteristics were observed in five out of ten experiments. The table's resulting p-values suggest that none of the data sets for any experiment, system, and characteristic adheres to a normal distribution. Hence, this hypothesis must be rejected.

Given the rejection of the normal distribution hypothesis, alternative distributions are considered. Thus, the Kolmogorov-Smirnov test is now employed to assess whether the performance data follow a Weibull distribution. The results are additionally shown in Table 4.7. For the LOCU system, the test supports this null hypothesis for horizontal accuracy, horizontal repeatability, and time gap for each experiment. However, for time delay, the hypothesis is consistently

Table 4.7: Overview of amount of experiments that conform to the hypothesis that data of the respected performance characteristics comes from a normal or Weibull distribution ($p > 0.05$) for both systems

	Hor. acc.	Hor. rep.	Heading acc.	Heading rep.	Time delay	Time gap
LOCU						
Normal distribution	0	0	–	–	0	0
Weibull distribution	10	10	–	–	0	10
LLS						
Normal distribution	0	0	0	0	0	0
Weibull distribution	10	7	10	10	10	4

rejected in all experiments. For the LLS system, the p-values suggest the adherence to a Weibull distribution for all performance characteristics, including horizontal accuracy, heading accuracy, heading repeatability, and time delay. In particular, for horizontal repeatability, seven out of ten tests, and for time gap, four out of ten tests yield a p-value less than 0.05.

Next, the repeatability of the experiment is examined. As the classical ANOVA necessitates normally distributed data, instead the Kruskal-Wallis test, an alternative test for not normally distributed data, is used [134, 135]. This test is applied for both systems and each performance characteristic to calculate the p-values, summarized in Table 4.8. Given that all p-values significantly exceed 0.05, the original hypothesis is confirmed, proving the repeatability of the experiments. In the following chapter, the implications of these findings will be discussed and reflected.

Table 4.8: Overview of p-values for each performance characteristic for both systems

	Hor. acc.	Hor. rep.	Heading acc.	Heading rep.	Time delay	Time gap
LOCU	0.918	0.564	–	–	0.938	0.635
LLS	0.941	0.260	0.933	0.583	0.808	0.336

4.3 Study of Experiment Comparability

This study focuses on the demand of stakeholders for the comparability of the results of various experiments. With experiment repeatability confirmed, this section advances by identifying differences in the experiments, expressed by influencing factors, that significantly affect the T&E results. The focus of this study is on the LLS system and the evaluation of horizontal localization accuracy, as this is generally considered the main performance criterion. Similarly to the previous study on the repeatability of experiments, this section begins with an overview of the conducted experiments in Section 4.3.1, followed by presenting the results of the performance evaluation in Section 4.3.2. Next, statistical methods are used in Section 4.3.3. Finally, Section 4.3.4 offers a detailed interpretation of the results.

4.3.1 Experiments

To contribute to the analysis of the comparability of experiments, the experiments for this study must be designed in a way that they entail systematically varying experiment conditions. This is achieved by defining a set of potentially relevant influencing factors and two possible characterizations each. The experiments are then designed by varying one factor at a time, leading to distinct experiment specifications.

In the following, each influencing factor is introduced, using refined terms adapted from those listed in Table 4.1 to improve specificity and clarity. However, each influencing factor will be linked to the previously used categories. For this study, two characterizations are considered for each influencing factor introduced. These factors, along with their respective characterizations, are summarized in Table 4.9. The use of black and white bullets in this table serves to denote the association between experiments and their influences in the upcoming graphical presentation of the results. Influencing factors associated with white bullets align with those of previous studies, while black bullets indicate modifications.

Table 4.9: Parameter values for considered influencing factors (LLS)

Factor	**Values**	
Environment	○ *open*	● *aisle*
Map	○ *open*	● *aisle*
Reflector	○ *no*	● *yes*
Field of View (FoV)	○ 270°	● 180°
Dynamics	○ *no*	● *yes*

Environment (Static Objects) Describes the existence of static objects within the test area. Two distinct environments were established: a clear test area (*open*), consistent with previous studies, and an aisle environment (*aisle*), which introduces logistics objects to simulate real-world challenges. Figure 4.17 shows the ELT navigating the *aisle* environment.

Map (Amount and Location of Reference Nodes) The LLS system's localization relies on the creation of an environmental map, marking the map as a crucial influencing factor in this study. A map is generated for both *open* and *aisle* environments. For each environment's map, a specific global alignment is determined through an individual alignment experiment.

Reflector (Amount and Location of Reference Nodes) The accuracy of localization for the LLS system can be enhanced by placing specific reflectors within the environment. These reference points must be registered during the mapping process. In practice, a reflector is positioned centrally and no more than one meter from the test area's boundaries on each side. This characterization is denoted as *yes* when the system uses reflector positions for localization, and *no* when this information is excluded.

Figure 4.17: Robot navigating in “aisle” environment

FoV (Device Position) The FoV is recognized as a potentially significant influencing factor in this study due to its impact on the quantity of data points the LLS system can utilize for map matching. The FoV is configured to its maximum extent of 270° or reduced to 180°. Although the FoV settings are part of the system configuration for the experiments, they are in practice constrained by the device’s placement.

Dynamics (Dynamic Entities) Dynamics are introduced in some experiments by having a person walk ahead of the robot, maintaining a distance of 1.5 m to 2.5 m. This influencing factor simulates real-world movement scenarios. The scenarios with dynamic elements present are labeled “yes” whereas those without are labeled “no”.

In addition to the stated expression of the influencing factors, the experiments are configured similarly to the case study presented in Section 4.1. However, the path and evaluation poses are adjusted so that the robot is guided through the adapted environment. During experiment execution, localization and reference data were recorded as the robot followed its path, covering 64 predefined evaluation poses. Modifying one factor at a time results in 32 distinct experiments. For each, both the system and the environment were configured accordingly. The experiments were carried out within a single day, generating 32 sets of experiment data.

4.3.2 Performance Results

Similarly to the approach used in the earlier repeatability study (Section 4.2.2), this section leverages the *Performance Evaluation* methods of the *T&E 4Log Framework* to generate performance results from each set of experiment data. However, to streamline this study, the focus here is on examining only the horizontal position error for the LLS system.

Figure 4.18 consolidates the results by presenting the distribution of data through whisker plots for each experiment, numbered from 1 to 32. For example, *Experiment 1* showcases the system's default setting (**Environment** (Env.): *open*; **Map**: *open*; **Reflector** (Ref.): *no*; **FoV**: 270°; **Dynamics** (Dyn.): *no*), while *Experiment 32* displays a fully modified configuration (**Environment**: *aisle*; **Map**: *aisle*; **Reflector**: *yes*; **FoV**: 180°; **Dynamic**: *yes*)). As such, the figure provides a comprehensive overview of the performance results obtained.

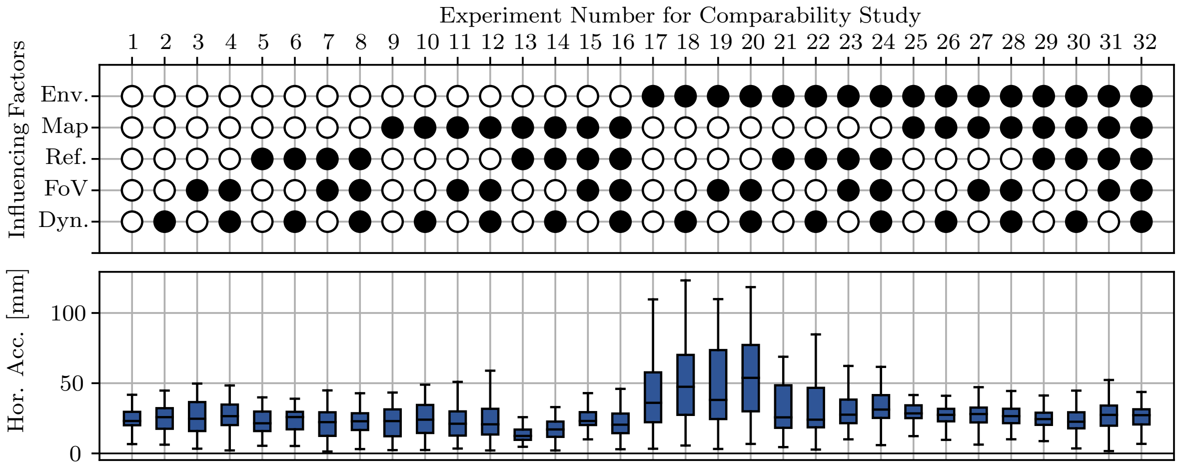


Figure 4.18: Whisker plots of horizontal position error under varying influences (LLS). White and black bullets indicate the association with the characterization of influencing factors as previously introduced in Table 4.9

Across the 32 experiments conducted, the horizontal error stayed consistently below 150 mm, predominantly ranging between 0 mm and 50 mm. Significant variations in error distributions are evident among the individual experiments. For example, a comparison between the results of the default setup (*Experiment 1*) and those obtained from *Experiment 17*, which shares the same configuration except the environmental setup, indicates its significant effect on the horizontal position error.

Although Figure 4.18 provides a comprehensive overview of the results, due to a large number of experiments, assessing the influence of factors solely through comparisons of whisker plots is challenging. Consequently, the following section complements the provided results by employing statistical methods to elucidate the effects of each factor more clearly.

Alternative approaches to visualizing performance data are possible. For example, the use of decision trees to intuitively visualize the impacts of influencing factors has previously been explored. While facilitating manual interpretation, this approach requires additional data processing steps and reduces the incorporated information. Hence, this work prioritizes the presentation of data through whisker plots. For a more detailed exploration of the decision tree approach, reference is made to an earlier publication [110].

4.3.3 Statistical Analysis

In this section, the performance results for the LLS system of the comparability study presented in Figure 4.18 are further examined. For this purpose, statistical methods are employed to determine the significance and rough scale of the effect of several influencing factors on the horizontal position accuracy.

Generally, a straightforward method to determine the effect of one variable on another is to fit a linear equation to the observed data by using simple linear regression. However, since this study seeks to explore the influence of multiple factors, multiple linear regression is used [136, pp. 17]. Multiple linear regression is particularly suitable for this research as it enables the exploration of complex interactions within experimental data, by treating each influencing factor as an independent variable.

To facilitate the utilization of multiple linear regression, all characterizations of the influencing factors are converted into binary dummy values. Characterizations marked with white bullets are assigned a value of “0”, while those marked with black bullets are represented by “1”. Converting also the continuous value for FoV to binary values results in greater consistency and reduces the risk of misinterpretation for values other than those used for the analysis. The dependent variable used for the regression analysis is the mean value of the horizontal position error.

The *Python* implementation of multiple linear regression from the *Statsmodels* library [137] is utilized to determine the linear coefficient k and the p-value for each influencing factor, which is then summarized in Table 4.10. Here, the p-value indicates the significance of each influencing factor. A result smaller than or equal to 0.05 indicates that a factor is significant. The coefficient k represents the change in the horizontal position error when the linear coefficient changes from one binary value to another, thus representing the effect of each influencing factor.

Table 4.10: Overview of linear coefficients k, p-values, and significance ($p \leq 0.05$) for each influencing factor (LLS)

Factor	k	p**-value**	**Significance**
Environment	-0.011	0.00	yes
Map	0.009	0.00	yes
Reflector	0.006	0.00	yes
FoV	-0.002	0.29	no
Dynamics	-0.002	0.25	no

The findings reveal that variations in the characterization of the influencing factors *Environment*, *Map*, and *Reflector* exert significant effects on the horizontal position error. Conversely, changes in the *Field of View (FoV)* ($k = -0.002$) and *Dynamics* ($k = -0.002$) demonstrate insignificant impacts. A negative coefficient of -0.011 for the *Environment* signifies a negative effect when transitioning from an *open* to an *aisle* environment, marking it as the factor with the greatest effect. The map alteration emerges as the second most impactful, with a coefficient of 0.009, indicating a positive change when moving from an *open* to an *aisle* map. Furthermore, the introduction of reflectors into the environment is shown to enhance localization, represented by a coefficient of 0.006.

The application of multiple linear regression has provided insight into the significance and impact of various influencing factors on the horizontal position error. However, this approach is not without its constraints, such as the assumed independence of variables or the utilization of binary dummy values in combination with assumed linearity, necessitating careful consideration. Hence, in the next section, the findings of the multiple linear regression analysis will be interpreted within the broader context of this study.

4.3.4 Interpretation of Results

In this scetion, the outcomes of the multiple linear regression are interpreted in conjunction with the previously presented visualization of the horizontal position accuracy. Firstly, the influencing factors ***Environment*** and ***Map*** are considered. The results of multiple linear regression indicate that the *empty* environment and the *aisle* map lead to greater accuracy. The method assumes that the influencing factors are independent variables. However, regarding the whisker plots presented in Figure 4.18, a cross-correlation between the two factors becomes apparent, since the difference between the results of *Experiment 1* to *8* and *Experiment 17* to *24* is substantially higher than the difference between the results *Experiment 9* to *16* and *Experiment 25* to *32*. Notably, *Experiment 17* to *Experiment 20* demonstrate considerably elevated error distributions, attributable to their combination of the environment (*aisle*) and map (*open*).

This identified connection between environment and map is logical, considering that the LLS system determines its position by matching the environment points measured with the pre-recorded map. To investigate this relationship and create starting points for further analysis, the following approach is taken. Initially, a new map was recorded for both environmental setups during an actual experiment. These new environment maps are then aligned with the original maps previously utilized for localization. This alignment employs a registration-based point cloud matching algorithm provided by the *Open3D Python* library [138]. The algorithm calculates a fitness value, indicating the ratio of points in the new map that matched well with the old map, with a fitness value of 100 % denoting an ideal match. The alignment of points in the four pairings of the new environment maps and the original maps is shown in Figure 4.19,

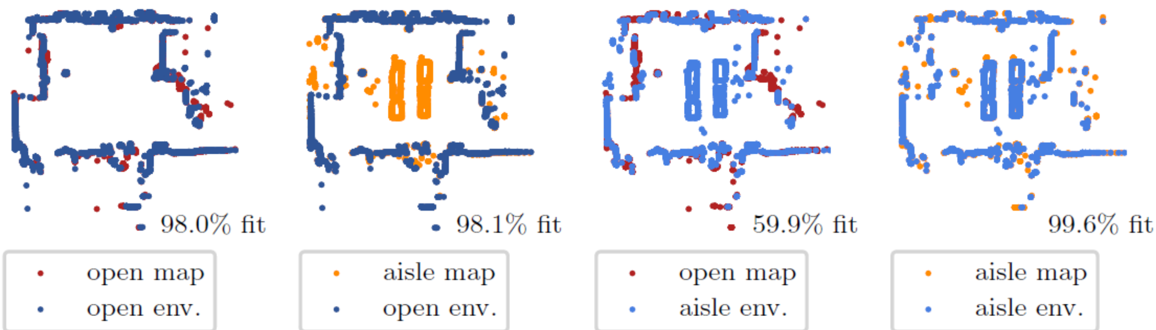


Figure 4.19: Map matching results for the four combinations of maps used for experiments and maps recorded as a representation of the environment. The dark blue and light blue points indicate measurements of the contour, reflecting the open and aisle environments, respectively. The orange and red points indicate measurements that have been used as a map for localization.

accompanied by the respective fitness values for each combination.

Despite the presence of outliers, the data points of each map allow the identification of the contours of the environment. Figure 4.19 shows that the *open* environment matches to a high degree of 98.0 % on its corresponding map. Similarly, the *aisle* environment matches to a high degree of 99.6 % on its respective map. This aligns with the expectation that each map accurately represents its specific environment. The discrepancies from a 100 % fit are attributed to outliers and variations in the robot's path during the mapping process. Notably, the *open* environment also demonstrates a high match rate of 98.1 % with the *aisle* map, indicating that the essential contours of the environment are captured on the map. However, the *aisle* environment matches to a significantly lower degree of 60.0 % with the *open* map. This suggests that when the sensor operates within the aisle environment, it does not find many matching points on the open map, leading to a considerable decline in horizontal position accuracy for the respected experiments.

Concluding the discussion on the effect of map and environment, it can be hypothesized that the presence of additional points in a map poses less of an issue for the system's horizontal position accuracy than the absence of points that represent significant objects in the environment. To validate this hypothesis, further empirical research is necessary. An intriguing approach for regression, contrary to the one of utilizing binary values for map and environment, involves treating the fitness value as a continuous variable to assess its impact. Nonetheless, further investigation falls beyond the scope of this dissertation.

The multiple linear regression analysis indicated a significant influence of the presence of ***Reflectors*** on the system's accuracy. This effect is further examined by analyzing the whisker plots displayed in Figure 4.18. To assess the impact of reflectors while maintaining consistency in other variables, comparisons are made between grouped experiments. *Experiment 1* to *4* is compared to *Experiment 5* to *8*, *Experiment 9* to *12* with the next four experiments, and so forth. A distinct improvement in performance attributed to the use of reflectors can be particularly observed among the lower-performing experiments when comparing *Experiment 17* to *20* with *Experiment 21* to *24*. This suggests that reflectors may not significantly increase accuracy when the measured contour points match closely with the map. However, in instances where there is a significant deviation between the environment and the map, the precise coordinates of known reflectors provide a notable advantage.

Finally, the results related to the influencing factors ***FoV*** and ***Dynamic*** are examined. The whisker plots provided reveal a reduced variability in the data distribution between *Experiment n* and the three subsequent experiments, with *n* being multiples of four. This pattern indicates a relatively minor influence of both factors *FoV* and *Dynamic*, a finding supported by multiple linear regression analysis. In the experimental setup, both factors ultimately limit the sensor's ability to measure points in the environment. The observed insignificance of this reduced view on the system's accuracy can be explained by the maintained data points that remain for location determination. This finding suggests that a slight diminishment in sensor data points does not critically affect the system's accuracy. To further validate this hypothesis, additional experiments with a specific focus are necessary.

Potentially, the findings of the comparability study could be used in conjunction with the results of the *Requirement Specification* of the case study as outlined in Section 4.1. This would allow for an investigation of which localization functions can be reliably enabled by an ILS under which environmental and system configurations. However, this exceeds the scope of this work.

In summary, three empirical studies were presented throughout this chapter. The first study demonstrated how the *T&E 4Log Framework* can be applied to assess the performance and suitability of both a *LIDAR*-based (LLS) and a *UWB*-based localization system (LOCU) for the AuC “Mobile Robots for Material Transport”. This exploration comprehensively covered the systematic application of the framework’s procedures.

The second study delved into the performance results of repetitive execution of experiments to examine the distribution type and experiment repeatability for both SuTs. The hypothesis of a normal distribution was disproved for both systems. Yet, the analysis revealed conformance to a Weibull distribution. Moreover, the data supported the repeatability of most experiments for both SuT, confirming the hypothesis of consistent performance under repeated conditions.

Lastly, the third study concentrated on the comparability of outcomes from systematically varying experiments. Visualization of performance results, along with the application of multiple linear regression, facilitated interpretation of the significance and influence of different influencing factors.

Conclusively, this chapter has established an empirical basis for the forthcoming discourse on the framework’s validity while offering an examination of the presented localization systems. Naturally, transferring the presented results to real-world application scenarios must be approached with caution. This matter, among others, will be thoroughly addressed in the forthcoming discussion chapter.

5 Discussion

This chapter completes the framework's validation by offering a comprehensive discussion based on the outcomes of the empirical examination presented in Chapter 4. Initially, the responses to the research questions posed are discussed in Section 5.1, followed by the discussion of the *T&E 4Log Framework* in its entirety in Section 5.2. The chapter concludes with a critical reflection in Section 5.3.

5.1 Responses to Research Questions

The response to each research question consists of various framework components that can be categorized into procedures, methods, concepts, tools, guidelines, and actions, which have been explored in a process-oriented manner throughout the presentation of the *T&E 4Log Framework* in chapter 3.3. In the following, the responses to each research question are summarized and ultimately discussed based on the findings of the empirical examination and the previously presented state of the art.

A comprehensive overview of the main components of the framework and their correspondences with the individual research questions is provided in Table 5.1. In addition to procedures and methods, the table highlights key concepts, main tools, and selected guidelines that underpin the *T&E 4Log Framework*. Although most of these components are novel developments, some established tools have been adopted. For such, the sources are stated in Table 5.1. Since the actions are numerous and intricately integrated into the methods, they are not explicitly summarized in this section. Examples of specific actions include "defining the dimensions of the test volume", "performing interpolation of GT for timestamp alignment", or "conducting calibration of the SuT and the reference ILS".

Research Question 1

RQ1 addresses the challenge of defining test scenarios that produce comparable and transferable T&E results within intralogistics contexts. It directly addresses two key stakeholder demands, aiming for results that are not only comparable but also transferable. The inherent conflict between these stakeholder demands has previously been highlighted in Section 2.2.2. In short, transferability requires that test scenarios closely mimic the specific application scenarios for which the T&E results are intended, demanding specificity, while comparability calls for systematic and standardized scenario definitions.

The *ISO/IEC 18305:2016* [17] favors comparability by proposing 14 predefined scenarios, which encompass a limited range of motion, ELT, and building types. However, this approach does not incorporate processes for systematically defining scenarios, which limits transferability.

Table 5.1: Overview of connections between research questions, procedures and methods, key concepts, main tools, and selected guidelines

	RQ1 How should test scenarios be defined that yield comparable and transferable T&E results in intralogistics contexts?	**RQ2** How should benchmarking of ILSs be conducted for repeatable, black-box, system-level testing in partially controlled test environments?	**RQ3** How should the suitability of ILSs for an intralogistics application be evaluated based on their localization performance?
Procedures and Methods	(*a*) *Application Description* • Process Description • Environment Description (*b*) *Scenario Definition* • Influencing Factor Identification • Influencing Factor Characterization	(*c*) *Experiment Specification* • Testbed Description • Process, Environment, and System Specification • Evaluation Pose Specification (*d*) *Experiment Execution* • Experiment Set-up • Experiment Data Recording • Alignment Data Recording (*e*) *Performance Evaluation* • Experiment Data Processing • Spatial Alignment • Performance Metric Determination	(*f*) *Requirement Specification* • Localization Function Identification • Requirement Margin Specification • Effective Localization Error Specification (*g*) *System Evaluation* • System Suitability Evaluation
Key Concepts	• Testbed-independent scenario as an aggregation of application-driven influencing factors • Linkage between indoor localization technologies and application-driven influencing factors • Characterization of influencing factors based on systematic application description	• Evaluation poses including orientation components and tolerances • Spatial alignments including global and local alignment based on data of dedicated experiment • Performance metrics including localization repeatability and system latency	• Specification based on the objective of ensuring reliable localization functions • Requirement margin including interest space and motion space • Effective localization error including relevant localization error, time gap error, and time delay error
Main Tools	• Correspondences between components of common indoor localization technologies and application-driven influencing factors (Table 3.2)	• PTP for time synchronization [114] • Umeyama-technique for global alignment determination [117] • Algorithm for associating data points with evaluation poses	• Equation 3.37 and 3.38 for conservative system suitability assessment
Selected Guidelines	• Resemblance between scenario and application description is significant for achieving transferable T&E results • Automated experiment execution recommended for achieving high repeatability and comparability of T&E results	• Minimum number of evaluation poses $k > 30$ recommended (derived from Poorter *et al.* [113]) • Minimum GT accuracy one order of magnitude higher than SuT accuracy [17] • Determination of percentiles according to sigma-levels	• Localization repeatability is significant if the interest space is defined within ILS coordinates • Time delay error is significant if the entity's current location is required • Additional localization errors must be considered as a result of the transformation between the entity's localization frame and the interest frame

Conversely, the *EVARILOS Benchmarking Handbook* [21] offers a more adaptable framework, though it primarily concentrates on RF-based ILS.

Within the *T&E 4Log Framework*, the definition of test scenarios is facilitated by the *Application Description* and *Scenario Definition* procedures. Initially, the AuC is abstracted within the *Application Description*, by describing the pertinent application processes and environment, thus establishing a base for formulating application-driven test scenarios. The case study presented in Section 4.1 proved the applicability of this approach. It revealed how the description of the application processes deepens the understanding of the AuC and pinpoints the activities for which absolute localization is crucial. Naturally, the level of detail should be appropriate for the scope of the study. The AuC can represent a generic, prototypical, or even a hypothetical scenario designed to showcase certain conditions. While the *Application Description* offers valuable guidance, it does not introduce novel key concepts.

The *Scenario Definition* procedure is used to formulate testbed-independent scenarios as an aggregation of application-driven influencing factors. These are directly linked with the components of indoor localization technology to allow systematic identification of potentially relevant influences. Using the taxonomy of localization technology presented in Section 2.1.2 and the correspondence table of application-driven influencing factors (Table 3.2), both concepts have been used in the exemplary case study to identify potentially relevant influencing factors. Assuming basic knowledge about the systems' technology, this approach has proven to be effective. However, for other influencing factors than those listed in Table 3.2, new correlations must first be established.

The tension between the comparability and transferability surfaces during the characterization of the influencing factors. Within the *T&E 4Log Framework*, the definition of a test scenario is achieved by the separate characterization of each potentially relevant factor. Although the resulting testbed-independent scenario outlined in Table 4.2 may not encompass all AuC characteristics, it pinpoints relevant influences, clearly differentiating it from scenarios applicable in alternate contexts.

In addition to the results of the case study discussed, the findings of the comparability study (Section 4.3), identified both significant and negligible influencing factors. Hence, this investigation demonstrated that varying scenarios, rooted in a practical application scenario, lead to distinct outcomes, which can be systematically examined thanks to their structured definitions. Regarding transferability, while it was shown that influencing factors derived from an application scenario can have a significant impact on the T&E results, the transfer of these results back to real-world application scenarios remains an area for further investigation.

Nevertheless, the presented approach offers a significant advantage over existing methodologies by allowing for flexible yet systematic scenario definition. The *T&E 4Log Framework* doesn't preclude the definition of standard cases. In contrast, the definition of standard cases is encouraged. However, the presented approach should be employed to enhance their relevance for real-world applications. In contrast to the *EVARILOS Benchmarking Handbook*, the *T&E 4Log Framework* leads to the definition of test scenarios that are testbed-independent. This higher level of abstraction allows the scenarios to serve as a basis for comparing T&E results obtained from experiments carried out in different test facilities. Furthermore, the *T&E 4Log Framework* sets itself apart from existing methodologies by offering guidance on achieving resemblance with real-world application scenarios.

Research Question 2

The *T&E 4Log Benchmarking Procedure* was presented as a response to RQ2 addressing repeatable, black-box, system-level testing in partially controlled environments. This central pillar of the *T&E 4Log Framework* offers a comprehensive approach for translating a test scenario into an experiment specification, conducting experiments, and evaluating system performance.

The case study presented in Section 4.1 showcased the framework's capability to derive performance metrics for two distinct ILSs based on an application-driven test scenario. This is achieved through the sequential utilization of methods, divided into the *Experiment Specification*, *Experiment Execution*, and *Performance Evaluation* procedure. The findings of the case study are complemented by the outcomes of the repeatability study presented in Section 4.2 providing the empirical foundation for the upcoming discussion on the *T&E 4Log Benchmarking Procedure*'s applicability and utility.

In the case study, the *Experiment Specification* procedure was applied to outline an experiment based on the provided test scenario. The methods facilitated a comprehensive and systematic specification of the testbed as well as a systematic derivation of process, environment, and system specifications. The evaluation poses were manually specified, supported by the *T&E 4Log App*, reflecting the intended path features "straights and curves within the horizontal plane". The system specification significantly influences system performance. To achieve increasingly comparable results across different test facilities, a standardized way to specify and deploy systems of certain types should ideally be agreed on.

Subsequently, the methods outlined in the *Experiment Execution* procedure were employed to produce sets of experiment and alignment data. Figure 4.7, which shows the horizontal positions of both SuTs, indicates successful data recording while demonstrating the relevance of spatial alignments. Facilitated by the *T&E 4Log App*, the robot successfully passed all evaluation poses within specified tolerances. Arising from a time discrepancy during synchronization over WLAN with PTP, a maximum measurement error of 0.55 mm was computed. Depending on the accuracy of the SuT, this demonstrates the importance of acknowledging this measurement error as an additional measurement uncertainty. The performance evaluation of the LLS system, has shown high horizontal position accuracy in the low centimeter range. This poses challenges for the accuracy of the GT to serve as a viable reference. In this range, it is difficult to prove the absolute accuracy of the GT, and a standardized method does not exist.

Finally, the *Performance Evaluation* procedure led to the determination of performance metrics for both SuTs, supported by the *T&E 4Log App*. The comparison of the alignment data presented in Figure 4.7 with the corresponding data points depicted in Figure 4.8 demonstrates the effective execution of the data processing steps for timestamp alignment and evaluation pose association with the *ClosestPoseAssociation* algorithm.

Furthermore, the effect of the global alignment process is demonstrated by comparing the horizontal positions before alignment (Figure 4.7) with those after alignment (Figure 4.9). Similarly, the relevance and effect of local alignment have been examined and illustrated in Figure 4.10. Accordingly, without applying these alignments, the computed performance metrics would be corrupted. Spatial alignment in this study is identified as a key concept of the *T&E 4Log Framework*. Although the Umeyama-technique is a well-established method for global alignment, the approach of determining both global and local alignment via a dedicated alignment experiment designed to minimize systematic errors represents a novel contribution. Nevertheless, it must be recognized that the accuracy of the alignment determination is contingent on the accuracy of the SuT itself. For systems characterized by low accuracy, conventional approaches present a favorable option.

The main outcomes of the *T&E 4Log Benchmarking Procedure* are the performance metrics. The establishment of relevant performance metrics based on the consideration of location data requirements revealed the critical role of localization repeatability, an attribute that is mostly overlooked in the context of T&E. This study introduced a novel concept for determining localization repeatability metrics based on the error derivation between two repeated approaches to the same evaluation pose. Unlike the *ISO 18646:2024-2*, this innovative concept significantly simplifies the process, requiring only two passes along the evaluation poses. The feasibility of this concept was successfully demonstrated in the case study presented for two distinct ILSs. It is important to note that localization repeatability, by measuring the deviation between two errors, as defined in this study, diverges from the one described by the *ISO 18646:2024-2*, describing localization repeatability through the deviation from the mean value. However, if the measurement error follows a standard normal distribution, the conversion from one value to the other can be described mathematically. Yet, the definition of localization repeatability used in this work aligns with the interpretation of the corresponding requirement parameter.

Unlike localization repeatability, system latency is considered in the *ISO/IEC 18305:2016* as well as in the *EVARILOS Benchmarking Handbook*. However, these methodologies fail to define an effective approach for determining latency for an ILS on a system level. Hence, this work proposes a novel approach to this challenge in Section 3.3.7. The outcomes from the empirical examination demonstrated the effectiveness of this approach for determining repeatable results for the LLS data while producing significantly less consistent results for the LOCU system. This indicates that the success of this approach is contingent upon the system's accuracy and scattering of provided location data relative to its latency. Nevertheless, this promising approach requires further exploration.

The metrics related to the remaining performance characteristics of the localization accuracy and the update rate are largely established. However, this work proposes an additional computation of localization error percentiles across various sigma levels in alignment with availability requirements. This contradicts the suggestion of Potortì *et al.* [15] to limit the number of performance metrics to increase comprehensibility for system users. Within the *T&E 4Log Framework*, this perspective is addressed by separating the provided information into performance results addressing system developers and testers, as well as suitability results addressing system users. The confidence of the percentiles determined by interpolation depends on the number of data points and should be further investigated, especially for the high percentiles. Regarding the update rate, it was initially assumed to be constant, hence the median was chosen for its characterization. However, both systems showed varying time gaps, suggesting further investigation. If the premise of a constant update rate is deemed inappropriate for most ILSs, employing percentiles to describe the update rate and potentially the system latency would be a more suitable approach. In addition to the proposed performance metrics, the experiment data could be used to derive various other metrics, such as velocity-dependent heading accuracy or position error depending on the distance to the origin. However, these derived metrics do not align with the system suitability assessment of the *T&E 4Log Framework* and are therefore not focused on in this work.

Emphasized in RQ2, the experiment repeatability is a crucial stakeholder demand, distinguishing the *T&E 4Log Benchmarking Procedure* from existing methodologies. The dedicated repeatability study confirmed the repeatability of experiments for both SuTs across most performance characteristics. By demonstrating the methodology's ability to generate consistent results, experiment repeatability builds the foundation for other stakeholder demands such as replicability and comparability. Nonetheless, it is vital to recognize that the repeatability of an experiment is not solely determined by the applied methodology but also by the characteristics inherent to the specific SuT. Hence, verifying the repeatability of experiments is strongly encouraged in the context of T&E studies.

In addition to experiment repeatability, RQ2 points out the necessity of employing black-box and system-level testing to allow T&E of diverse technologies across various applications. The *T&E 4Log Benchmarking Procedure* confirms its commitment to a system-level approach by testing the system in its entirety and based the evaluation on the system's location data output. The inner workings of the systems are merely recognized during the definition of test scenarios to identify potentially significant application-driven influencing factors, but not as part of the benchmarking procedure. As a result, the experiments carried out within the case study were uniform across both SuTs, ensuring a fair and unbiased comparison.

Overall, the *T&E 4Log Benchmarking Procedure* integrates several procedures, methods, concepts, tools, and guidelines into a comprehensive benchmarking methodology for repeatable, black-box, system-level testing in partially controlled environments. Integrated key concepts, highlighted in Table 5.1 are the consideration of evaluation poses, the determination of global and local alignments through dedicated alignment experiments, as well as methods for assessing localization repeatability and system latency. The procedure incorporates various tools, including PTP for time synchronization [114], the Umeyama-technique for global alignment determination [117], and the *ClosestPoseAssociation* algorithm for associating data points with evaluation poses.

Research Question 3

Finally, RQ3 is dedicated to assessing how well an ILS suits to be used for an intralogistics application based on its localization performance. As such, the response to RQ3 serves to identify practically relevant performance metrics and ultimately to facilitate systematic system suitability assessment. Thus, it forms a vital element of an application-driven T&E methodology. The respective components of the *T&E 4Log Framework* are summarized in Table 5.1.

The response to RQ3 is embedded within the *Requirement Specification* and *System Evaluation* procedures of the *T&E 4Log Framework*. The suitability assessment is based on the concept of localization functions. Accordingly, the aim is to determine the presence or absence of an entity within or outside a multidimensional interest space, representing the primary purpose of location data in practical applications. Thus, location data requirements are specified to ensure reliable localization functions. By incorporating the performance metrics obtained from the *T&E 4Log Benchmarking Procedure* Equation 3.37 or 3.38 can be used to assess the suitability of an ILS for a specific localization function or an application that involves multiple localization functions. The equations are based on the requirement margin setting a limit for the allowable effective localization error comprising the relevant localization error, time gap error, and time delay error.

The applicability of the presented approach has been demonstrated by assessing the suitability of two distinct ILSs within the case study that deals with the application of mobile robots for material transport. The case study emphasized the importance of identifying localization functions to clarify the ultimate purpose of location data within an AuC. Defining the requirement margin in the subsequent step based on spatial considerations, such as interest and motion space, presents a systematic approach for specifying an upper limit for the effective localization error.

Determining the effective localization error necessitates informed assumptions about various process parameters, such as the ELT's velocity or the network transmission time. These are often unknown and require case-specific assumptions to be made. Additionally, the presented approach considers the distinction between the entity's interest frame and the entity's localization frame. As demonstrated for the "pallet placement" process through Equation 4.6, this distinction has emerged as a critical factor, particularly when the localization system's orientation estimate is either unavailable or significantly inaccurate.

The principle that the effective localization error should be lower than the requirement margin leads to a conservative estimate, facilitating data-driven system suitability assessment using Equation 3.37 or 3.38. The case study revealed that the LLS system is suitable for "pallet pick-up" and "pallet transport" but not for "pallet placement", whereas the LOCU system is considered unsuitable for all these processes. The outcome that the LOCU system is considered unsuitable for any of the examined application processes could arguably be made without benchmarking, simply due to its lack of providing heading information. Nevertheless, the outcomes are valuable. For example, based on the presented results the LOCU system may be considered suitable to provide position data, whereby data fusion techniques could be used to infer missing heading information.

The outcome of the assessment depends on the benchmarking results, the implementation of the ILS, and the various assumptions made, necessitating a careful evaluation. Yet, the presented approach offers a systematic exploration of various dependencies, facilitating the finding of an appropriate solution. As such, the outcomes of the comparability study presented in Section 4.3 can be used to determine the conditions under which a system is considered suitable.

Despite its potential, there are different limitations. Firstly, the presented approach focuses solely on evaluating the suitability of the system based on its localization performance. As discussed in Section 2.1.4, numerous additional user requirements such as weight, robustness, and cost are also crucial, presenting a multidimensional problem. To ultimately decide which is the ideal choice, it is imperative to consider all these user requirements. Moreover, the presented approach is grounded in the concept of localization functions. However, there are other requirements for location data, such as scattering of data to derive dynamic properties as presented in [4].

The complexity of the system assessment approach presented may appear too high for stakeholders involved in a simple system assessment. In part, this is due to the intricate nature of the AuC chosen to demonstrate the capabilities of the approach within the case study. However, for simpler applications, this approach becomes more manageable. Moreover, implementing the proposed methods within software tools for user guidance would offer a more accessible and streamlined process.

The T&E methodologies currently available largely overlook location data requirements. Only the *EVARILOS Benchmarking Handbook* allows the application-driven system assessment based on simple weighing factors for calculating the final score. In contrast, the presented methods provide a systematic approach for specifying requirements and assessing system suitability based on a system's localization performance within a holistic T&E methodology. Furthermore, the introduced approach for assessing system suitability progresses significantly beyond the elementary ideas of Hohenstein *et al.* [86], the domain-specific approach of Reid *et al.* [33], and the high-level strategies of Mautz [12] and Gladysz *et al.* [85] discussed in Section 2.1.4. The approach does not rely exclusively on benchmarking within partially controlled test environments. Hence, performance data obtained from real-world applications or building-wide testing could also be applicable.

5.2 Fulfillment of Research Objective

In addition to delivering responses to the research questions, integrating these responses into a holistic methodology is another key achievement of this work. The *T&E 4Log Framework* was developed as research artifact aimed at fulfilling the research objective. Building on the previous discussion of the responses to the individual research questions, this section provides an overarching perspective that focuses on the *T&E 4Log Framework* in its entirety. Hence, the extent to which the research objective has been met is discussed, drawing on findings from the empirical examination, thereby highlighting novel contributions, potentials, and limitations, while pointing out opportunities for further scientific exploration.

The research objective required the *T&E 4Log Framework* to be designed as an application-driven methodology. This was achieved through the adoption of a hybrid approach, inspired by Seltzer *et al.* [108], where both the testing and evaluation parts of the methodology are influenced by applications. The task of defining application-driven test scenarios was addressed by RQ1, as illustrated in the case study in Section 4.1. Moreover, this case study showcased the framework's capability to assess system suitability for a range of localization functions within an application, addressed by RQ3. This application-driven approach distinguishes the *T&E 4Log Framework* from existing methodologies, offering structured yet flexible guidance for conducting T&E driven by application needs. Since the inner workings of localization systems ultimately do not matter to an application as long as requirements are met, an application-driven T&E methodology requires black-box testing at the system level, as emphasized in RQ2. The presented case study demonstrated the implementation of this concept, resulting in similar experiments and evaluations for two distinct ILSs.

Moreover, the research objective emphasizes the T&E results to be meaningful for stakeholders within intralogistics contexts. Firstly, this focus on the intralogistics domain motivated experimentation in partially controlled test environments promising comparably high transferability of results. Furthermore, it guided the selection of significant indoor localization technologies to be considered and influenced the approaches for defining test scenarios and specifying location data requirements. The empirical examination showcased the usage of the *T&E 4Log Framework* within the context of intralogistics, highlighting its applicability and utility in this domain. Besides, the choice to center the dissertation on the intralogistics domain is grounded in the author's background and expertise in this area. Nonetheless, researchers are encouraged to explore how the findings of this research could be generalized and adapted to areas beyond intralogistics.

Another significant aspect of the research objective is the demand for meaningful results in partially controlled test environments. The term meaningfulness, which refers to the significance of the results for stakeholders, has been discussed in Section 3.2. This led to identifying feasibility, repeatability, comprehensibility, comparability, reproducibility, and transferability as key stakeholder demands for a T&E methodology. These partly conflicting qualities were considered for numerous design decisions throughout this research. In the following, the extent to which the *T&E 4Log Framework* satisfies these stakeholder demands is discussed individually.

Feasibility In the context of T&E, feasibility has been defined as the practicality of conducting T&E, considering the required resources, time, and technical limitations. For the *T&E 4Log Framework* this was effectively demonstrated through the presented case study. The inherent modularity of the framework increases practicality by supporting individual usage of specific procedures or methods. Moreover, the *T&E 4Log Framework* provides guidance that diminishes the need for specialized personnel to obtain meaningful T&E results. A testbed is the primary resource required for benchmarking. However, once established, such a testbed can facilitate benchmarking for a diverse range of test scenarios. Furthermore, testing in partially controlled environments enhances the potential for automation, significantly lowering manual labor, particularly for a large number of experiments. This provides significant benefits compared to the methodologies presented that rely on building-wide testing. Additionally, in such cases, providing accurate GT across large areas is costly, especially for evaluating highly accurate ILSs. Besides, T&E activities may interfere with ongoing operations and raise privacy concerns in cases involving camera-based localization.

Overall, T&E with the *T&E 4Log Framework* offers the potential for significantly reduced costs and effort, particularly when dealing with a large quantity or a diverse range of experiments. Nevertheless, the practicality of the methodology could be greatly enhanced by providing a software tool capable of partially automating benchmarking, documentation, and possibly even the dissemination and comparison of results across different testbeds. There have been initiatives to create such software tools, including the *EVARILOS Benchmarking Suite* [139] and the *evo* tool [106]. The *T&E 4Log App* presents a first implementation of the methods presented in this work. It is encouraged to further develop or refine the *T&E 4Log App* to expand its robustness, functionality, and applicability. However, ideally, a consensus on the methods to integrate should be previously agreed upon.

Comprehensibility In this work, comprehensibility is defined as the degree to which information is presented in a clear, organized, and understandable way. The *T&E 4Log Framework* addresses this demand by incorporating modularity and process orientation into its design principles. Breaking down the methodology into procedures with defined inputs and outputs and outlining the methods that are applied sequentially provides clear guidance. In addition, comprehensibility is improved by attributing information or tasks to the specific roles of stakeholders throughout the various stages of the T&E process. Furthermore, the methodology's development is meticulously documented within this dissertation, and its practical application is thoroughly presented through an illustrative case study, both of which contribute to its overall comprehensibility.

Repeatability Repeatability represents a crucial stakeholder demand that distinguishes this approach from those presented in the literature. It was explicitly highlighted in RQ2 and therefore already discussed in the previous section. As such, experiment repeatability has been confirmed within the repeatability study, establishing the foundation for the comparability, reproducibility, and transferability of the results.

Reproducibility The ability of independent researchers to achieve consistent results via T&E requires a high comprehensibility of the methods used and, ideally, high feasibility. In addition to these previously discussed qualities, reproducibility requires replicability. Similarly to repeatability, replicability is characterized by the ability to produce consistent results, however, based on testing in various facilities. Variations in the spatial environment and technical equipment across facilities introduce discrepancies in the experiments that are expected to show a notable effect on performance results. For this research, multiple testing facilities were not available to empirically examine this effect, and therefore the replicability of T&E with the *T&E 4Log Framework* has not yet been investigated. Hence, the research community is encouraged to explore this attribute. However, unlike other T&E methodologies, the *T&E 4Log Framework* facilitates these further investigations by addressing and confirming the repeatability of experiments.

Comparability In this research, comparability is understood as the ability to contrast results from distinct experiments or systems. This involves the comparability of experiments that may vary in influencing factors, systems, or even test facilities. As such, the *T&E 4Log Framework* is designed to offer a base to systematically analyze the results both within a single study and across various studies. It is underpinned by repeatability, which provides a reliable basis for identifying significant differences between experiment results.

Firstly, comparability of results from various systems is facilitated by adopting a black-box, system-level approach. The case study showcased how such experiments enabled the comparison of results from distinct ILSs. Comparability between results from experiments conducted under different influencing factors is achieved by offering a structured approach for defining test scenarios, followed by straightforward instructions for specifying, executing, and evaluating experiments. This was illustrated in the dedicated comparability study in Section 4.3. Exemplified through the use of multiple linear regression, this study further highlighted the benefits of comparable results.

Likewise, the ability to compare results from different test facilities is facilitated by defining new test scenarios or applying existing ones, followed by employing the methods outlined in the *T&E 4Log Benchmarking Procedure*. Nonetheless, similar to reproducibility, this aspect of comparability is constrained by replicability, an aspect that warrants further exploration.

Transferability Finally, transferability refers to the ability to produce results that apply to real-world application scenarios. The *T&E 4Log Framework* addresses this through the adoption of an application-driven T&E approach. Consequently, experiments reflect real-world application scenarios and the identification and evaluation of application-driven influencing factors is facilitated. Moreover, the *T&E 4Log Framework* enables the alignment of performance results with relevant location data requirements to assess the suitability of a system for an application. This represents a significant advance over existing methodologies. Yet, the exploration into how closely T&E results align with system performance in real-world application scenarios needs to be further explored. Currently, applying benchmarking results to real-world application scenarios requires careful consideration. Nevertheless, by adopting an application-driven approach with experiments that mirror real-world application scenarios, the *T&E 4Log Framework* moves on from idealized or arbitrarily defined experiments towards more meaningful T&E.

5.3 Critical Reflection

To complete the discussion of this work, this section provides a critical reflection on the research process itself. It starts with a reflection of the empirical examination, proceeds to assess the overarching validation concept, and concludes by discussing the overall scope of this dissertation.

Empirical Examination The empirical examination was structured into three parts. Initially, a case study was presented in Section 4.1, demonstrating the usage of all components of the framework. Thus, the case study offered an empirical basis for assessing the applicability and utility of the methodology. Nevertheless, certain aspects warrant a critical perspective. For example, although the specific case was developed in coordination with industrial partners, the process could have benefitted from a more practice-oriented approach. Incorporating concrete examples from the industry could have demonstrated the methodology's practicality and value more effectively. Furthermore, the provision of a single case study restricts the opportunity to thoroughly evaluate the adaptability of the proposed methodology across various applications and technologies, which should ideally require or provide location data with up to six DoFs. However, incorporating additional cases, especially doing this in close collaboration with industrial partners, would have significantly increased the effort for validation, which was not feasible within the constraints of available resources for this work. Instead, a complex AuC featuring diverse localization functions was chosen to showcase the potential of the framework. Furthermore, because of the foundational nature of this work, a detailed investigation of the repeatability and comparability of the results was given precedence.

Repeatability and comparability were examined within the context of a particular testbed and specific localization systems. Investigating different systems, or assessing the same systems within varied facilities or under different experimental setups, might yield divergent outcomes. Therefore, the results related to repeatability and comparability cannot be attributed exclusively to the *T&E 4Log Framework*. Moreover, controlling, identifying, and generalizing experiment conditions poses significant challenges. Although they vary to some extent, they are assumed to remain constant. Yet, the presented examination demonstrated the feasibility of producing repeatable and comparable results and outlined approaches to carry out further investigations.

Validation Validating algorithms or technical systems is typically done by comparing them against each other or to a baseline, utilizing established metrics. The methodology introduced in this work lays the groundwork for such comparative studies of ILSs. However, validating the methodology itself poses significant challenges, since there is no universal truth for a system's localization accuracy or an application's location data requirements. Thus, validating a T&E methodology involves thorough discussions of their potential and benefits. However, which methodology is preferable remains contingent on the specific use case.

In this work, the discussion of the potential and limitations of the research artifact was focused on various stakeholder demands. The examinations of feasibility and comprehensibility leveraged the outcomes of the exemplary case study, while the discussions of repeatability and comparability were underpinned by the findings of the specific studies. However, as noted in the preceding section, it was not possible to conduct a comprehensive empirical examination of replicability and transferability within the scope of this work. In addition, a discussion of how well the research meets the demands of stakeholders could be improved through systematic engagement with various stakeholders. However, conducting such an in-depth investigation goes beyond what was achievable within the scope of this research.

Finally, this research introduced several novel concepts, such as the determination of spatial alignments, localization repeatability, and system latency, integrating them into an overarching methodology. These developments were crucial in assembling the components of this holistic approach, linking relevant location data requirements with application-driven benchmarking. Their use has been successfully demonstrated in the presented case study and subsequently discussed in Section 5.1. Nevertheless, several of them require further and more extensive validation.

Research Scope Initially, the goal of crafting an application-driven T&E methodology might appear straightforward. However, diving deeper into the topic revealed that the state of the art hinges on various aspects. Benchmarking ILSs is standard practice, yet the critical discussion on what aspects of accuracy matter in practice, how to specify those requirements, and how to determine them meaningfully is often overlooked. This research pursued the ambitious goal of integrating all of these components, leading to the development and exploration of several innovative concepts. The comprehensive nature of this approach, which incorporates both overarching and granular concepts, resulted in a methodology of considerable complexity. Some ideas presented may be too complex or impractical to gain widespread acceptance. However, in alignment with DSR the main goal of this work was not to create an easily adoptable standard but to advance research on T&E by expanding the current knowledge base.

6 Conclusions

To conclude this work, this chapter begins by pinpointing the theoretical contributions made by this research in Section 6.1. Section 6.2 proceeds by highlighting the practical implications of the results and underlining the relevance of this dissertation. Finally, Section 6.3 provides directives for stakeholders to further push the boundaries of current understanding and derive value from these advancements.

6.1 Theoretical Contributions

The theoretical contributions of this dissertation are given as significant expansions to the current knowledge base. These contributions are the validated responses to the three posed research questions and the overarching research artifact. In the following, a brief summary of the responses to the research questions and of the *T&E 4Log Framework* is provided.

RQ1: How should test scenarios be defined that yield comparable and transferable T&E results in intralogistics contexts?

The first research question was addressed by developing and validating a systematic approach for creating application-driven and testbed-independent test scenarios. This strategy encompasses methods, concepts, tools, and guidelines for outlining the processes and environments of real-world or hypothetical applications, identifying potentially significant influencing factors depending on the indoor localization technology used, and characterizing these influencing factors based on the application. The applicability and utility of this approach were demonstrated through an empirical examination. By offering a systematic yet adaptable design, aimed at achieving comparable and transferable T&E results, this approach marks a substantial advancement over existing methodologies, which typically rely on providing standard test cases rather than guiding the specification of relevant cases.

RQ2: How should benchmarking of ILSs be conducted for repeatable, black-box, system-level testing in partially controlled test environments?

The *T&E 4Log Benchmarking Procedure* addresses RQ2 by offering a detailed strategy for converting test scenarios into experiment specifications, executing experiments, and evaluating system performance. This yields a set of performance metrics tailored to match location data requirements. The presented approach incorporates novel concepts such as methods for determining spatial alignments through specific alignment experiments, and techniques for evaluating localization repeatability and system latency. Unlike current methodologies that focus on building-wide testing, the *T&E 4Log Benchmarking Procedure* provides an approach for testing within partially controlled test environments, aiming for results that are repeatable and transferable. Its adherence to repeatable, black-box, system-level testing within partially controlled environments was demonstrated through empirical examination that underscored its practicality and potential.

RQ3: How should the suitability of ILSs for an intralogistics application be evaluated based on their localization performance?

The *T&E 4Log Framework* introduces an approach to evaluate the suitability of systems for intralogistics applications, grounded in the concept of localization functions. These localization functions define the primary purpose of absolute location data in a real-world application to ascertain an entity's presence or absence within or outside a multidimensional interest space. Following this fundamental idea, the methodology guides the determination of a requirement margin that defines the maximum localization error to allow for a reliable localization function. Furthermore, the approach involves the estimation of the effective localization error based on various process parameters like the ELT's velocities and coordinate transformations, as well as performance metrics concerning the system's accuracy, localization repeatability, update rate, and latency. A system is ultimately deemed suitable when the requirement margin exceeds the effective localization error. This novel approach was demonstrated in a case study. Limitations remain as the presented approach is grounded in the underlying concept of localization functions and offers a conservative estimation of system suitability. Nevertheless, it marks a significant advancement over existing methodologies, which fall short of providing any methods for adequately specifying location data requirements, making this work's systematic quantification approach a first of its kind.

Research Objective: Design of an application-driven T&E methodology, that provides meaningful results to stakeholders in the context of intralogistics, based on repeatable, black-box, system-level testing within partially controlled test environments.

Finally, the research objective is addressed through the *T&E 4Log Framework*, which integrates the responses to the posed research questions within an application-driven methodology. Its modular, process-oriented design is tailored to meet the demands of stakeholders. The key features that set this methodology apart from existing ones are the focus on repeatable testing within partially controlled test environments and the application-driven approach. The *T&E 4Log Framework* is focused on localization functions and location data requirements for systems that deliver absolute location data. The framework was discussed based on an empirical examination, including a case study and two studies addressing specific stakeholder demands. This underscored the value of repeatable experiments for systematically analyzing performance results. The main limitation for T&E based on tests outside of real-world application scenarios remains the transferability of the results. This issue has been addressed by aiming to reflect real-world conditions through abstraction as application-driven influencing factors. Yet, further examinations of the gap between T&E results and real-world performance are crucial.

6.2 Practical Implications

While the previous section highlighted the novelty of this work's contributions, the current section emphasizes its relevance. As such, the practical implications of this research for different stakeholders are outlined below.

For system users and system integrators, this work's contributions offer guidance on evaluating system suitability, supporting the identification of reliable solutions. Given that an ILS determines the key capabilities of an indoor localization application, choosing a system that is both reliable and cost-effective is crucial, particularly when deploying systems on a large scale. Therefore, this work's contributions offer considerable benefits to system users who either apply the proposed methods themselves or request their usage by system providers. Additionally, system users ultimately benefit from increased market competition, thanks to the transparency that arises from the increased comparability of T&E results.

System developers rely on T&E to identify the strengths and weaknesses of systems, thereby ensuring their reliability in operation and providing evidence to direct their development efforts. For ILSs, this is especially critical considering the complexity and heterogeneity of indoor localization technologies and the varied and often safety-critical applications. Following the methods for application-driven T&E presented in this dissertation empowers system developers to systematically examine system behavior in a meaningful way at a reasonable cost. Moreover, the approach for the systematic specification of location data requirements presented in this work allows system developers to understand the impact of performance characteristics on system suitability, guiding their development efforts effectively. Additionally, this approach can be used to identify applications for which an ILS is suitable. Finally, customer trust and satisfaction can be improved by providing performance metrics that are relevant and reliable. Yet, transforming these potentials into tangible outcomes, such as lower development costs, enhanced products, and greater customer satisfaction depends on the stakeholders' initiative to employ these methods.

Beyond system users and system developers, researchers gain substantially from this work's contributions. They can utilize the proposed methods to validate their ideas and concepts through T&E, which is of particular importance as comparability, transferability, and reproducibility represent key pillars of scientific research. The necessity of a methodology for partially controlled test environments is even more apparent for public research institutions that often lack the resources for extensive building-wide testing. Furthermore, consistent use of the methods presented in this dissertation would strengthen the evidence of review papers summarizing and comparing the results of various scientific studies. Hence, the implications of this work benefit the scientific advancement in the field of indoor localization.

For all stakeholders, the guidance offered in this work reduces the necessity for expertise in conducting meaningful T&E. Especially at an early stage of system or application design, testing in partially controlled environments presents a cost-effective alternative to extensive building-wide testing. Furthermore, adopting an established methodology enhances the feasibility of outsourcing T&E to specialized test facilities. Moreover, based on common approaches, researchers, developers, and system users are empowered to create shared databases encompassing test scenarios and location data requirements for diverse indoor localization applications in conjunction with the performance results of various systems and experiments. This collaborative effort could furthermore facilitate data-driven studies concerning the performance and suitability of ILSs on a large scale. Ultimately, the presented methodology or individual elements have the potential to evolve into or contribute to a standard, which could then serve as a basis for jurisdiction.

The potential practical implications resulting from this research work are manifold. However, some of the benefits rely on the method's widespread adoption. It remains to be seen whether the *T&E 4Log Framework* will find acceptance in its entirety. Nevertheless, by providing a thorough investigation of T&E in the context of intralogistics, addressing various challenges, presenting a comprehensive methodology, and highlighting areas for further research, this work paves the way for future efforts moving the field toward more unified practices.

6.3 Stakeholder Directives

The previous sections outlined the theoretical contributions and practical implications of this dissertation. This final section provides directives for further expanding the knowledge base through relevant contributions and unlocking the various practical potentials.

Development and research in the area of T&E do not conclude with the *T&E 4Log Framework*. The overall methodology and its components must evolve through scholarly discourse within the research community drawing on empirical studies and the experiences of diverse stakeholders. The main areas that warrant deeper investigation include the following.

- Examining transferability of results obtained from experiments in partially controlled test environments to real-world application scenarios
- Assessing the replicability of results in different test facilities
- Exploring the potential and limitations for the adoption of the proposed methods to domains beyond intralogistics
- Enhancing and further investigating methods for determining test scenarios, spatial alignments, localization repeatability, system latency, and location data requirements

The approach for system assessment outlined in this dissertation addresses solely location data requirements, as these can be effectively determined through benchmarking in partially controlled test environments. However, identifying the ideal system requires considering other user requirement parameters pointed out in Section 2.1.4, such as robustness, size, and costs. Therefore, researchers are encouraged to complement the presented methodology through a broader system selection approach.

Furthermore, the adoption of methodological approaches can be greatly facilitated by offering reliable software tools that substantially reduce the risk of errors, the level of required expertise, and the effort for conducting meaningful T&E. Although initial efforts to develop such software tools have been made, primarily focusing on performance evaluation, the investigation of the underlying methods is frequently overlooked. The *T&E 4Log App* was developed as a software prototype to validate the benchmarking methods introduced in this work, with the source code made publicly available [122]. The research community is encouraged to further develop this initial version collaboratively into a user-friendly software tool grounded in validated methods.

Ultimately, the realization of the practical benefits relies on stakeholders employing the introduced concepts. Leveraging the presented methods can support their refinement and broader acceptance, whereby system users emerge as key drivers by requesting performance evaluations aligned with common methodologies.

Continued refinement, validation, and expansion of the proposed methodology and its components, together with the provision of a user-friendly software tool, could significantly increase the adoption of methodological approaches for application-driven T&E in intralogistics and beyond. Fueled by the various practical implications, this will promote greater adoption of ILSs and ultimately contribute to the ongoing digital transformation across industries.

Bibliography

[1] G. Nagy, B. Illés, and Á. Bányai, "Impact of industry 4.0 on production logistics," *IOP Conference Series: Materials Science and Engineering*, vol. 448, 2018.

[2] H. Huang and G. Gartner, "Current trends and challenges in location-based services," *International Journal of Geo-Information*, vol. 7, no. 6, 2018.

[3] A. Mütze, L. Hingst, N. Rochow, T. Miebach, and P. Nyhuis, "Use cases of real-time locating systems for factory planning and production monitoring," in *Conference on Learning Factories (CLF)*, Graz, Austria, 2021.

[4] J. Schyga, H. Rose, J. Hinckeldeyn, and J. Kreutzfeldt, "Analysis of the operation of industrial trucks based on position data," *Logistics Journal*, vol. 17, 2021.

[5] N. Hesslein, M. Wesselhöft, J. Hinckeldeyn, and J. Kreutzfeldt, "Industrial indoor localization: Improvement of logistics processes using location based services," in *Advances in Automotive Production Technology–Theory and Application: Stuttgart Conference on Automotive Production (SCAP)*, Stuttgart, Germany, 2021.

[6] U. Behrje, M. Himstedt, and E. Maehle, "An autonomous forklift with 3d time-of-flight camera-based localization and navigation," in *International Conference on Control, Automation, Robotics and Vision (ICARCV)*, Singapore, 2018.

[7] B. Molter and J. Fottner, "Semi-automatic pallet pick-up as an advanced driver assistance system for forklifts," in *Intelligent Transportation Systems Conference (ITSC)*, Auckland, New Zealand, 2019.

[8] L. Barreto, A. Amaral, and T. Pereira, "Industry 4.0 implications in logistics: An overview," *Procedia Manufacturing*, vol. 13, 2017.

[9] R. F. Brena, J. P. García-Vázquez, C. E. Galván-Tejada, D. Muñoz-Rodriguez, C. Vargas-Rosales, and J. Fangmeyer, "Evolution of indoor positioning technologies: A survey," *Journal of Sensors*, 2017.

[10] S. Pasricha, "Overview of indoor navigation techniques," in *Position, Navigation, and Timing Technologies in the 21st Century: Integrated Satellite Navigation, Sensor Systems, and Civil Applications*, Y. J. Morton, F. van Diggelen, J. J. Spilker, B. W. Parkinson, S. Lo, and G. Gao, Eds. Newark, US: John Wiley & Sons Incorporated, 2021, vol. 2.

[11] F. Potortì, A. Crivello, and F. Palumbo, "The evaal evaluation framework and the ipin competitions," in *Geographical and Fingerprinting Data to Create Systems for Indoor Positioning and Indoor/Outdoor Navigation*, ser. Intelligent Data-Centric Systems, J. Conesa, A. Pérez-Navarro, J. Torres-Sospedra, and R. Montoliu, Eds., Amsterdam, Netherlands: Academic Press, 2019.

[12] R. Mautz, “Indoor positioning technologies,” Habilitation, ETH Zurich, Zurich, 2012.

[13] D. Lymberopoulos, J. Liu, X. Yang, R. R. Choudhury, V. Handziski, and S. Sen, “A realistic evaluation and comparison of indoor location technologies,” in *International Conference on Information Processing in Sensor Networks (IPSN)*, S. Nath, B. Krishnamachari, and A. Rowe, Eds., Seattle, USA, 2015.

[14] D. Bousdar Ahmed, L. E. Diez, E. M. Diaz, and J. J. Garcia Dominguez, “A survey on test and evaluation methodologies of pedestrian localization systems,” *IEEE Sensors Journal*, vol. 20, 2020.

[15] F. Potortì, A. Crivello, P. Barsocchi, and F. Palumbo, “Evaluation of indoor localisation systems: Comments on the iso/iec 18305 standard,” in *International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, Nantes, France, 2018.

[16] V. Andiappan and Y. K. Wan, “Distinguishing approach, methodology, method, procedure and technique in process systems engineering,” *Clean Technologies and Environmental Policy*, vol. 22, no. 3, 2020.

[17] *Information technology – real-time locating systems – test and evaluation of localization and tracking systems*, ISO/IEC 18305:2016 standard, 2016.

[18] T. van Haute, E. de Poorter, F. Lemic, V. Handziski, N. Wirström, T. Voigt, A. Wolisz, and I. Moerman, “Platform for benchmarking of rf-based indoor localization solutions,” *IEEE Communications Magazine*, vol. 53, no. 9, 2015.

[19] F. Lemic, V. Handziski, N. Wirstrom, T. van Haute, E. de Poorter, T. Voigt, and A. Wolisz, “Web-based platform for evaluation of rf-based indoor localization algorithms,” in *IEEE International Conference on Communication Workshop (ICCW)*, London, UK, 2015.

[20] F. Potortì, S. Park, A. R. Jiménez Ruiz, P. Barsocchi, M. Girolami, A. Crivello, S. Y. Lee, J. H. Lim, J. Torres-Sospedra, F. Seco, R. Montoliu, G. M. Mendoza-Silva, M. C. Del Pérez Rubio, C. Losada-Gutiérrez, F. Espinosa, and J. Macias-Guarasa, “Comparing the performance of indoor localization systems through the evaal framework,” *Sensors*, vol. 17, no. 10, 2017.

[21] T. Van Haute, E. De Poorter, I. Moerman, F. Lemic, V. Handziski, A. Behboodi, A. Wolisz, N. Wirström, T. Voigt, P. Crombez, and G. Glorioso, *Evarilos deliverables – d2.4 final version of the evarilos benchmarking handbook*, [Report]. Available: https://evarilos.eu/deliverables/D2.4_final_version_of_the_handbook.pdf. (Accessed: 14 May 2024), 2015.

[22] F. Potorti, P. Barsocchi, M. Girolami, J. Torres-Sospedra, and R. Montoliu, “Evaluating indoor localization solutions in large environments through competitive benchmarking: The evaal-etri competition,” in *International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, Banff, Canada, 2015.

[23] N. Moayeri, M. O. Ergin, F. Lemic, V. Handziski, and A. Wolisz, “Perfloc (part 1): An extensive data repository for development of smartphone indoor localization apps,” in *International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, Valencia, Spain, 2016.

[24] N. Moayeri, C. Li, and L. Shi, “Perfloc (part 2): Performance evaluation of the smartphone indoor localization apps,” in *International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, Nantes, France, 2018.

[25] B. Van Herbruggen, J. Fontaine, and E. D. Poorter, “Anchor pair selection for error correction in time difference of arrival (tdoa) ultra wideband (uwb) positioning systems,” in *International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, Lloret de Mar, Spain, 2021.

[26] G. Yammine, M. Alawieh, G. Ilin, M. Momani, M. Elkhouly, P. Karbownik, N. Franke, and E. Eberlein, “Experimental investigation of 5g positioning performance using a mmwave measurement setup,” in *International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, Lloret de Mar, Spain, 2021.

[27] N. Zhou, L. Lau, R. Bai, and T. Moore, “Evolutionary-inspired strategy for particle distribution optimization in auxiliary particle filtering algorithm based indoor positioning,” in *International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, Lloret de Mar, Spain, 2021.

[28] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, “A benchmark for the evaluation of rgb-d slam systems,” in *International Conference on Intelligent Robots and Systems (IROS)*, Vilamoura-Algarve, Portugal, 2012.

[29] A. R. Hevner, “A three cycle view of design science research,” *Scandinavian Journal of Information Systems*, vol. 19, no. 2, 2007.

[30] C. R. Kothari, *Research methodology: Methods & techniques*, 2. edition, reprint. New Delhi et al.: New Age International, 2004.

[31] A. R. Hevner, S. T. March, J. Park, and S. Ram, “Design science in information systems research,” *MIS quarterly*, vol. 8, 2004.

[32] L. M. Vaughn and F. Jacquez, “Participatory research methods – choice points in the research process,” *Journal of Participatory Research Methods*, vol. 1, 2020.

[33] T. G. R. Reid, S. E. Houts, R. Cammarata, G. Mills, S. Agarwal, A. Vora, and G. Pandey, “Localization requirements for autonomous vehicles,” *SAE International Journal of Connected and Automated Vehicles*, vol. 2, no. 3, 2019.

[34] K. Al Nuaimi and H. Kamel, “A survey of indoor positioning systems and algorithms,” in *International Conference on Innovations in Information Technology (IIT)*, Abu Dhabi, United Arab Emirates, 2011.

[35] *The international vocabulary of metrology – basic and general concepts and associated terms*, JCGM 200:2012 standard, 2012.

[36] G. M. Mendoza-Silva, J. Torres-Sospedra, and J. Huerta, “A meta-review of indoor positioning systems,” *Sensors*, vol. 19, no. 20, 2019.

[37] J. Xiao, Z. Zhou, Y. Yi, and L. M. Ni, “A survey on wireless indoor localization from the device perspective,” *ACM Computing Surveys (CSUR)*, vol. 49, no. 2, 2016.

[38] A. Pérez-Navarro, J. Torres-Sospedra, R. Montoliu, J. Conesa, R. Berkvens, G. Caso, C. Costa, N. Dorigatti, N. Hernández, S. Knauth, E. S. Lohan, J. Machaj, A. Moreira, and P. Wilk, "Challenges of fingerprinting in indoor positioning and navigation," in *Geographical and Fingerprinting Data to Create Systems for Indoor Positioning and Indoor/Outdoor Navigation*, J. Conesa, A. Pérez-Navarro, J. Torres-Sospedra, and R. Montoliu, Eds., Academic Press, 2019.

[39] X. Guo, N. Ansari, F. Hu, Y. Shao, N. R. Elikplim, and L. Li, "A survey on fusion-based indoor positioning," *IEEE Communications Surveys and Tutorials*, vol. 22, 2020.

[40] N. Samama, *Indoor positioning: Technologies and performance*. John Wiley & Sons Inc, 2019.

[41] M. Stahlke, S. Kram, C. Mutschler, and T. Mahr, "Nlos detection using uwb channel impulse responses and convolutional neural networks," in *International Conference on Localization and GNSS (ICL-GNSS)*, Tampere, Finnland, 2020.

[42] T. Kolakowski, *Uwb channel impulse responses registered in a furnished apartment (version 1.0)*, [Dataset]. Available: http://doi.org/10.5281/zenodo.4742391, 2021.

[43] F. Zafari, A. Gkelias, and K. K. Leung, "A survey of indoor localization systems and technologies," *IEEE Communications Surveys and Tutorials*, vol. 21, no. 3, 2019.

[44] S. A. Mohamed, M.-H. Haghbayan, T. Westerlund, J. Heikkonen, H. Tenhunen, and J. Plosila, "A survey on odometry for autonomous navigation systems," *IEEE Access*, vol. 7, 2019.

[45] N. F. Ribeiro and C. P. Santos, "Inertial measurement units: A brief state of the art on gait analysis," in *Portuguese Meeting on Bioengineering (ENBENG)*, Coimbra, Portugal, 2017.

[46] W. Sakpere, M. Adeyeye-Oshin, and N. B. Mlitwa, "A state-of-the-art survey of indoor positioning and navigation systems and technologies," *South African Computer Journal*, vol. 29, no. 3, 2017.

[47] Z. Farid, R. Nordin, M. Ismail, *et al.*, "Recent advances in wireless indoor localization techniques and system," *Journal of Computer Networks and Communications*, 2013.

[48] T. Ballal and C. J. Bleakley, "3d location and orientation estimation using angle of arrival," in *International Symposium on Intelligent Signal Processing (ISPACS)*, Budapest, Hungary, 2009.

[49] P. K. Panigrahi and S. K. Bisoy, "Localization strategies for autonomous mobile robots: A review," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 8, 2022.

[50] S. Thrun, W. Burgard, and D. Fox, *Probabilistic Robotics*. Cambridge, MA, USA: MIT Press, 2005.

[51] J. Aulinas, Y. Petillot, J. Salvi, and X. Lladó, "The slam problem: A survey," in *International Conference of the Catalan Association for Artificial Intelligence (CCIA)*, Sant Martí d'Empúries, Spain, 2008.

[52] L. Batistić and M. Tomic, “Overview of indoor positioning system technologies,” in *International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, Opatija, Croatia, 2018.

[53] *Wireless lan medium access control (mac) and physical layer (phy) specifications*, IEEE 802.11 standard, 2022.

[54] *Bluetooth core specification version 5.3*, standard, 2021.

[55] J. Torres-Solis, T. H. Falk, and T. Chau, *A review of indoor localization technologies: towards navigational assistance for topographical disorientation*. Rijeka, Croatia: IN-TECH Open Access Publisher, 2010.

[56] Y. Gu, A. Lo, and I. Niemegeers, “A survey of indoor positioning systems for wireless personal networks,” *IEEE Communications surveys & tutorials*, vol. 11, 2009.

[57] L. Mainetti, L. Patrono, and I. Sergi, “A survey on indoor positioning systems,” in *International Conference on Software, Telecommunications and Computer Networks (SoftCOM), Split, Croatia*, Split, Croatia, 2014.

[58] J. Kunhoth, A. Karkar, S. Al-Maadeed, and A. Al-Ali, “Indoor positioning and wayfinding systems: A survey,” *Human-centric Computing and Information Sciences*, vol. 10, 2020.

[59] J. A. del Peral-Rosado, R. Raulefs, J. A. López-Salcedo, and G. Seco-Granados, “Survey of cellular mobile radio localization methods: From 1g to 5g,” *IEEE Communications Surveys and Tutorials*, vol. 20, no. 2, 2018.

[60] C. Reinke and P. Beinschob, “Strategies for contour-based self-localization in large-scale modern warehouses,” in *International Conference on Intelligent Computer Communication and Processing (ICCP)*, Cluj-Napoca, Romania, 2013.

[61] L. Huang, “Review on lidar-based slam techniques,” in *Proceedings of the 2021 International Conference on Signal Processing and Machine Learning (CONF-SPML)*, Stanford, US, 14 November, 2021.

[62] Q. Zou, Q. Sun, L. Chen, B. Nie, and Q. Li, “A comparative analysis of lidar slam-based indoor navigation for autonomous vehicles,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 7, 2021.

[63] N. Li, C. P. Ho, J. Xue, L. W. Lim, G. Chen, Y. H. Fu, and L. Y. T. Lee, “A progress review on solid-state lidar and nanophotonics-based lidar sensors,” *Laser & Photonics Reviews*, vol. 16, no. 11, 2022.

[64] J. Kallwies, B. Forkel, and H.-J. Wuensche, “Determining and improving the localization accuracy of apriltag detection,” in *International Conference on Robotics and Automation (ICRA)*, Paris, France, 2020.

[65] Y. Desmarais, D. Mottet, P. Slangen, and P. Montesinos, “A review of 3d human pose estimation algorithms for markerless motion capture,” *Computer Vision and Image Understanding*, vol. 212, 2021.

[66] R. Azzam, T. Taha, S. Huang, and Y. Zweiri, “Feature-based visual simultaneous localization and mapping: A survey,” *SN Applied Sciences*, vol. 2, 2020.

[67] Y. Wu, F. Tang, and H. Li, “Image-based camera localization: An overview,” *Visual Computing for Industry, Biomedicine, and Art*, vol. 1, 2018.

[68] P. Nogueira, “Motion capture fundamentals,” in *Doctoral Symposium in Informatics Engineering (DSIE)*, Porto, Portugal, 2012.

[69] M. Knitt, J. Schyga, A. Adamanov, J. Hinckeldeyn, and J. Kreutzfeldt, “Estimating the pose of a euro pallet with an rgb camera based on synthetic training data,” *Logistics Journal*, vol. 18, 2022.

[70] R. Bostelman, J. Falco, and T. Hong, “Performance measurements of motion capture systems used for agv and robot arm evaluation,” *Autonomous Industrial Vehicles: From the Laboratory to the Factory Floor*, 2016.

[71] S. Thiede, B. Sullivan, R. Damgrave, and E. Lutters, “Real-time locating systems (rtls) in future factories: Technology review, morphology and application potentials,” *Procedia CIRP*, vol. 104, 2021.

[72] W. Fang and Z. An, “A scalable wearable ar system for manual order picking based on warehouse floor-related navigation,” *International Journal of Advanced Manufacturing Technology*, vol. 109, 2020.

[73] M. Nardello, L. Santoro, F. Pilati, and D. Brunelli, “Preventing covid-19 contagion in industrial environments through anonymous contact tracing,” in *International Workshop on Metrology for Industry 4.0 & IoT (MetroInd4.0&IoT)*, 2021.

[74] B. P. Sullivan, P. G. Yazdi, A. Suresh, and S. Thiede, “Digital value stream mapping: Application of uwb real time location systems,” *Procedia CIRP*, vol. 107, 2022.

[75] S. Thiede, P. Ghafoorpoor, B. P. Sullivan, S. Bienia, M. Demes, and K. Dröder, “Potentials and technical implications of tag based and ai enabled optical real-time location systems (rtls) for manufacturing use cases,” *CIRP Annals*, vol. 71, 2022.

[76] H. Unger, T. Markert, and E. Müller, “Evaluation of use cases of autonomous mobile robots in factory environments,” *Procedia Manufacturing*, vol. 17, 2018.

[77] E. Ackerman, “A robot for the worst job in the warehouse: Boston dynamics’ stretch can move 800 heavy boxes per hour,” *IEEE Spectrum*, vol. 59, 2022.

[78] M. Beul, D. Droeschel, M. Nieuwenhuisen, J. Quenzel, S. Houben, and S. Behnke, “Fast autonomous flight in warehouses for inventory applications,” *IEEE Robotics and Automation Letters*, vol. 3, no. 4, 2018.

[79] DHL, *Dhl logistics trend radar: Sixth edition*, [Online]. Available: https://www.dhl.com/global-en/home/insights-and-innovation/insights/logistics-trend-radar.html. (Accessed: 14 May 2024), 2022.

[80] M. Köseoğlu, O. M. Çelik, and Ö. Pektaş, “Design of an autonomous mobile robot based on ros,” in *International Artificial Intelligence and Data Processing Symposium (IDAP)*, Malatya, Turkey, 2017.

[81] *Systems and software engineering – life cycle processes – requirements engineering*, ISO/IEC/IEEEE 29148:2018 standard, 2018.

[82] *Systems and software engineering – systems and software quality requirements and evaluation – system and software quality models*, ISO/IEC 25010:2011 standard, 2011.

[83] J. Schyga, M. Knitt, J. Hinckeldeyn, and J. Kreutzfeldt, “Method for specifying location data requirements for intralogistics applications,” in *Conference on Production Systems and Logistics (CPSL)*, Santiago de Querétaro, Mexico, 2023.

[84] F. Jordan, A. Bernardy, M. Stroh, J. Horeis, and V. Stich, “Requirements-based matching approach to configurate cyber-physical systems for smes,” in *IEEE Portland International Conference on Management of Engineering and Technology (PICMET)*, Portland, USA, 2017.

[85] B. Gladysz and K. Santarek, “An approach to rtls selection,” in *International Conference on Production Research (ICPR)*, Posnan, Poland, 2017.

[86] F. Hohenstein and W. Günthner, “Anforderungen und fähigkeiten gegenwärtiger stapler-lokalisierung,” in *Hamburger Staplertagung*, Hamburg, Germany, 2013.

[87] J. Schyga, J. Hinckeldeyn, and J. Kreutzfeldt, “Meaningful test and evaluation of indoor localization systems in semi-controlled environments,” *Sensors*, vol. 22, no. 7, 2022.

[88] S. Adler, S. Schmitt, K. Wolter, and M. Kyas, “A survey of experimental evaluation in indoor localization research,” in *International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, Banff, Canada, 2015.

[89] S. Pasricha, “Overview of indoor navigation techniques,” in *Position, Navigation, and Timing Technologies in the 21st Century: Integrated Satellite Navigation, Sensor Systems, and Civil Applications*, Y. J. Morton, F. van Diggelen, J. J. Spilker, B. W. Parkinson, S. Lo, and G. Gao, Eds. Newark, US: John Wiley & Sons Incorporated, 2021, vol. 2.

[90] Y. Zhao, J. S. Smith, S. H. Karumanchi, and P. A. Vela, “Closed-loop benchmarking of stereo visual-inertial slam systems: Understanding the impact of drift and latency on tracking accuracy,” in *International Conference on Robotics and Automation (ICRA)*, Paris, France, 2020.

[91] R. Kümmerle, B. Steder, C. Dornhege, M. Ruhnke, G. Grisetti, C. Stachniss, and A. Kleiner, “On measuring the accuracy of slam algorithms,” *Autonomous Robotics*, vol. 27, no. 4, 2009.

[92] J. Torres-Sospedra, I. Silva, L. Klus, D. Quezada-Gaibor, A. Crivello, P. Barsocchi, C. Pendão, E. S. Lohan, J. Nurmi, and A. Moreira, “Towards ubiquitous indoor positioning: Comparing systems across heterogeneous datasets,” in *International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, Lloret de Mar, Spain, 2021.

[93] G. M. Mendoza-Silva, P. Richter, J. Torres-Sospedra, E. S. Lohan, and J. Huerta, “Long-term wifi fingerprinting dataset for research on robust indoor positioning,” *Data*, vol. 3, 2018.

[94] R. Montoliu, E. Sansano, J. Torres-Sospedra, and O. Belmonte, “Indoorloc platform: A public repository for comparing and evaluating indoor positioning systems,” in *International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, Sapporo, Japan, 2017.

[95] Y. Liu, W. Gao, and Z. Hu, "A large-scale dataset for indoor visual localization with high-precision ground truth," *International Journal of Robotics Research*, vol. 41, no. 2, 2022.

[96] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, Providence, USA, 2012.

[97] M. Helmberger, K. Morin, B. Berner, N. Kumar, G. Cioffi, and D. Scaramuzza, "The hilti slam challenge dataset," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, 2022.

[98] F. Potorti *et al.*, "Off-line evaluation of indoor positioning systems in different scenarios: The experiences from ipin 2020 competition," *IEEE Sensors Journal*, vol. 22, no. 6, 2021.

[99] M. C. Pérez-Rubio, C. Losada-Gutiérrez, F. Espinosa, J. Macias-Guarasa, J. Tiemann, F. Eckermann, C. Wietfeld, M. Katkov, S. Huba, J. Ureña, J. M. Villadangos, D. Gualda, E. Díaz, R. Nieto, E. Santiso, P. del Portillo, and M. Martínez, "A realistic evaluation of indoor robot position tracking systems: The ipin 2016 competition experience," *Measurement*, vol. 135, 2019.

[100] R. Ichikari, K. Kaji, R. Shimomura, M. Kourogi, T. Okuma, and T. Kurata, "Off-site indoor localization competitions based on measured data in a warehouse," *Sensors*, vol. 19, no. 4, 2019.

[101] N. Moayeri, C. Li, and L. Shi, "Perfloc (part 2): Performance evaluation of the smartphone indoor localization apps," in *International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, Nantes, France, 2018.

[102] Filip Lemic, Vlado Handziski, and Adam Wolisz, *Evarilos deliverables – d4.3b report on the cooperation with evaal and microsoft/ipsn initiatives*, [Report]. Available: https://www2.tkn.tu-berlin.de/tkn-projects/evarilos/deliverables/D4.3b.pdf. (Accessed: 14 May 2024), 2015.

[103] N. Moayeri, "Response to "evaluation of indoor localisation systems: Comments on the iso/iec 18305 standard"," in *WiP Proceedings of the International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, Lloret de Mar, Spain, 2021.

[104] *Information technology – real time locating systems device performance test methods – part 62: High rate pulse repetition frequency ultra wide band air interface*, ISO/IEC 24770-62:201 standard, 2015.

[105] *Information technology – robotics – performance criteria and related test methods for service robots*, ISO 18545:2021 standard, 2021.

[106] M. Grupp, *Evo: Python package for the evaluation of odometry and slam*, [Documentation]. Available: https://github.com/MichaelGrupp/evo. (Accessed: 14 May 2024), 2017.

[107] J. Schyga, J. Hinckeldeyn, B. Bruss, C. Bamberger, and J. Kreutzfeldt, "Application-driven test and evaluation framework for indoor localization systems in warehouses," in *WiP Proceedings of the International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, Lloret de Mar, Spain, 2021.

[108] M. Seltzer, D. Krinsky, K. Smith, and X. Zhang, “The case for application-specific benchmarking,” in *Workshop on Hot Topics in Operating Systems (HotOS)*, Washington DC, US, 1999.

[109] S. Mathur and S. Malik, “Advancements in the v-model,” *International Journal of Computer Applications*, vol. 1, no. 12, 2010.

[110] J. Schyga, S. Plambeck, J. Hinckeldeyn, G. Fey, and J. Kreutzfeldt, “Decision trees for analyzing influences on the accuracy of indoor localization systems,” in *International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, Beijing, China, 2022.

[111] R. Haberfellner, O. de Weck, E. Fricke, and S. Vössner, *Systems engineering – Fundamentals and Applications*. Springer, 2019.

[112] Z. Chen, Y. Liu, S. Li, and G. Wang, “Study on the multipath propagation characteristics of uwb signal for indoor lab environments,” in *International Conference on Ubiquitous Wireless Broadband (ICUWB)*, 2016.

[113] E. de Poorter, T. van Haute, E. Laermans, and I. Moerman, “Benchmarking of localization solutions: Guidelines for the selection of evaluation points,” *Ad Hoc Networks*, vol. 59, 2017.

[114] *Precision clock synchronization protocol for networked measurement and control system*, IEC/IEEE 61588:2021 standard, 2021.

[115] G. Fontana, M. Matteucci, and D. G. Sorrenti, “Rawseeds: Building a benchmarking toolkit for autonomous robotics,” *Methods and Experimental Techniques in Computer Engineering*, 2014.

[116] J. Craig, *Introduction to Robotics: Mechanics and Control*, 3. edition. Pearson, 2004.

[117] S. Umeyama, “Least-squares estimation of transformation parameters between two point patterns,” *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 13, no. 04, 1991.

[118] B. El-Haik, *Axiomatic quality: integrating axiomatic design with six-sigma, reliability, and quality engineering*. John Wiley & Sons, 2005.

[119] P. Stephan, I. Heck, P. Krau, and G. Frey, “Evaluation of indoor positioning technologies under industrial application conditions in the smartfactorykl based on en iso 9283,” *IFAC Proceedings Volumes*, vol. 42, no. 4, 2009.

[120] D. Bouvet and G. Garcia, “Gps latency identification by kalman filtering,” *Robotica*, vol. 18, no. 5, 2000.

[121] L. Shi, N. Moayeri, and C. Li, “On the use of lookahead to improve wi-fi fingerprinting indoor localization accuracy,” in *International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, LLoret de Mar, Spain, 2021.

[122] J. Schyga, *Repository for evaluating indoor localization systems with the t&e 4log framework*, [Repository]. Available: https://collaborating.tuhh.de/JakobSchyga/dissertation_js. (Accessed: 14 May 2024), 2024.

[123] SICK AG, *Agv system localization: More efficiency in intralogistics*, [Whitepaper]. Available: https://cdn.sick.com/media/docs/4/14/814/whitepaper_mobile_platforms_en_im0104814.pdf. (Accessed: 14 May 2024), 2022.

[124] SICK AG, *Locu – tag-based localization*, [Product Info]. Available: https://fcc.report/FCC-ID/WRMLOCU2A/4677719.pdf. (Accessed: 14 May 2024), 2023.

[125] SICK AG, *Lidar-loc – modular lidar localization based on natural contours*, [Product Info]. Available: https://cdn.sick.com/media/familyoverview/2/72/272/familyOverview_LiDAR-LOC_g541272_en.pdf. (Accessed: 14 May 2024), 2023.

[126] SICK AG, *Microscan3 – safety laser scanner*, [Product Info]. Available: https://www.sick.com/media/docs/7/57/757/Operating_instructions_microScan3_Safety_laser_scanner_en_IM0063757.PDF. (Accessed: 14 May 2024), 2016.

[127] Qualisys AG, *Qualisys stories – leveraging mocap data for localization of logistical systems*, [Blogpost]. Available: https://www.qualisys.com/stories/itl-tuhh/. (Accessed: 14 May 2024), 2023.

[128] C. Hansen, D. Gibas, J.-L. Honeine, N. Rezzoug, P. Gorce, and B. Isableu, "An inexpensive solution for motion analysis," *Journal of Sports Engineering and Technology*, vol. 228, no. 3, 2014.

[129] Qualisys AG, *Real-time latency tests of a qualisys system in the sensory-motor systems lab at eth, zurich, switzerland*, [Report]. Available: https://www.qualisys.com/news/real-time-latency-tests-of-a-qualisys-system-in-the-sensory-motor-systems-lab-at-eth-zurich-switzerland/. (Accessed: 14 May 2024), 2009.

[130] Open Source Robotics Foundation, Inc., *Turtlebot2 – open-source robot development kit for apps on wheels*, [Online]. Available: https://www.turtlebot.com/turtlebot2. (Accessed: 14 May 2024).

[131] L. Stahle and S. Wold, "Analysis of variance (anova)," *Chemometrics and intelligent laboratory systems*, vol. 6, no. 4, 1989.

[132] *Scipy 1.13.0 – kolmogorov-smirnov test*, [Documentation]. Available: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.kstest.html. (Accessed: 14 May 2024), 2024.

[133] V. W. Berger and Y. Zhou, "Kolmogorov-smirnov test: Overview," *Wiley StatsRef: Statistics reference online*, 2014.

[134] W. H. Kruskal and W. A. Wallis, *Use of ranks in one-criterion variance analysis*, 1952.

[135] *Scipy 1.13.0 – kruskal-wallis test*, [Documentation]. Available: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.kruskal.html. (Accessed: 14 May 2024), 2024.

[136] D. J. Olive, *Linear regression analysis*. Springer Cham, 2017.

[137] *Statsmodels 0.14.1 – linear regression*, [Documentation]. Available: https://www.statsmodels.org/stable/regression.html. (Accessed: 14 May 2024), 2023.

[138] Q.-Y. Zhou, J. Park, and V. Koltun, "Open3d: A modern library for 3d data processing," 2018.

[139] T. Van Haute, E. De Poorter, I. Moerman, V. Handziski, A. Behboodi, F. Lemic, A. Wolisz, N. Wirström, T. Voigt, P. Crombez, and G. Glorioso, *Evarilos deliverables - d2.5 evarilos benchmarking suite*, [Report]. Available: https://www2.tkn.tu-berlin.de/tkn-projects/evarilos/deliverables/D2.5_EVARILOS_benchmarking_suite.pdf. (Accessed: 14 May 2024), 2014.

Appendix

Overview of the *T&E 4Log App*

The *T&E 4Log App* facilitates the specification, execution, and evaluation of experiments following the *T&E 4Log Framework*. The app allows users to specify, execute, and evaluate experiments, and view results. It is built using the *FLASK Python* web development framework. Detailed usage and installation instructions are provided on *GitLab* alongside the commented program code and exemplary experiment data [122]. In the following, a brief overview of the functionality is provided by describing the various tabs, each dedicated to a specific function, such as project management, specification, execution, evaluation, or database operations.

Index

The *Index* tab (Figure A.1) offers an introduction to the app, an overview of its functionalities, and links to further documentation.

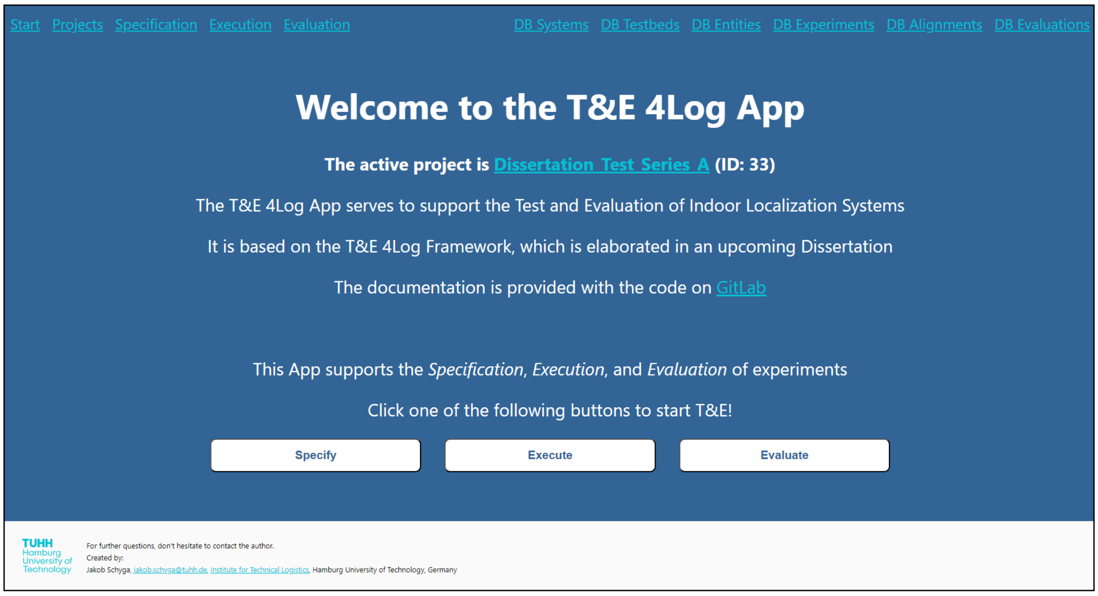


Figure A.1: Exemplary screenshot of *Index* tab to introduce user

Project Management

The *Project* tab (Figure A.2) lets you view, create, edit, and delete projects. Each project is associated with an automatically generated directory and a *SQLite3* database.

Figure A.2: Exemplary screenshot of *Project* tab for project management

Specification

Experiments can be created and configured in the *Specification* tab (Figure A.3). This process involves setting up evaluation poses.

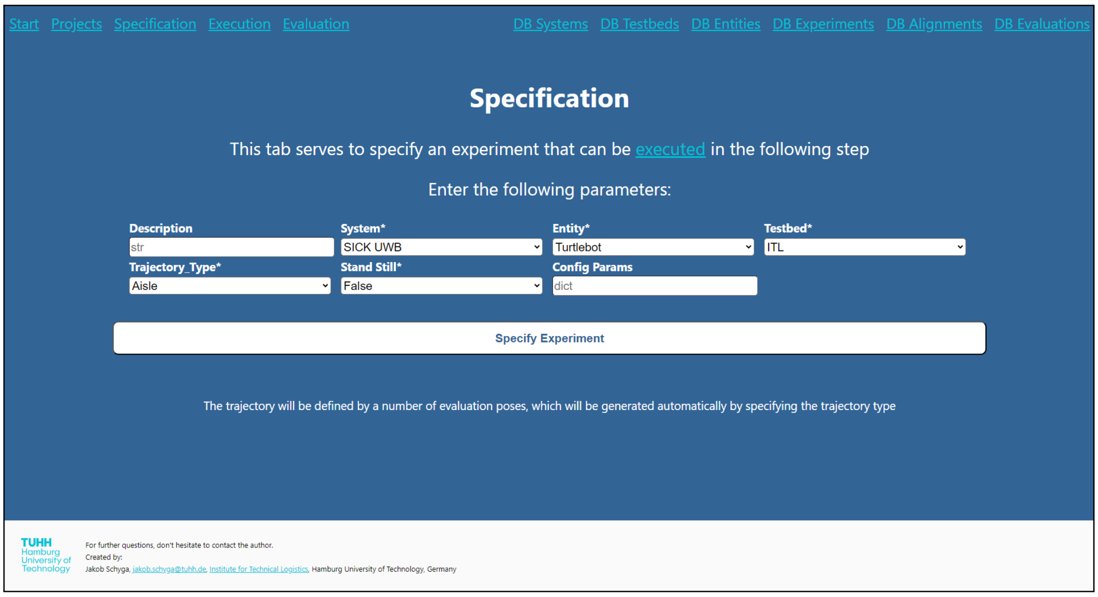


Figure A.3: Exemplary screenshot of *Specification* tab to assist systematic experiment specification

Execution

The *Execution* tab facilitates automated or manual guidance of the entity throughout the test area within the ITL testbed while recording reference and location data as *rosbag* files through ROS.

Evaluation

The *Evaluation* tab enables the evaluation of previously recorded experiment data. It encompasses the determination and alignment of global and local alignments and ultimately computes various performance metrics. The evaluation data are displayed in an interactive *Bokeh* dashboard. The dashboard enables visualization and provides a summary of the evaluation data. In addition to various data tables, it includes interactive plots to offer a comprehensive view of the evaluation data (Figure A.4).

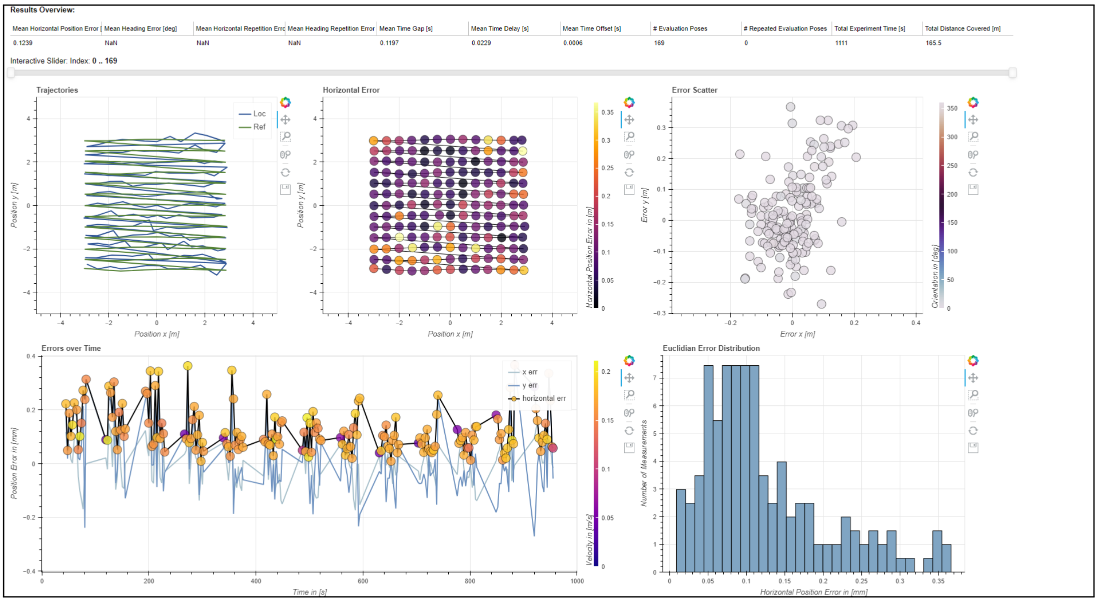


Figure A.4: Exemplary screenshot of interactive dashboard plot view

Create, Read, Update, Delete (CRUD) Database Operations

The tabs *DB Systems*, *DB Entities*, *DB Testbeds*, *DB Experiments*, and *DB Evaluations* allow viewing, adding, editing, and deleting database entries. Figure A.5 shows exemplarily a screenshot of the *DB Experiments* tab.

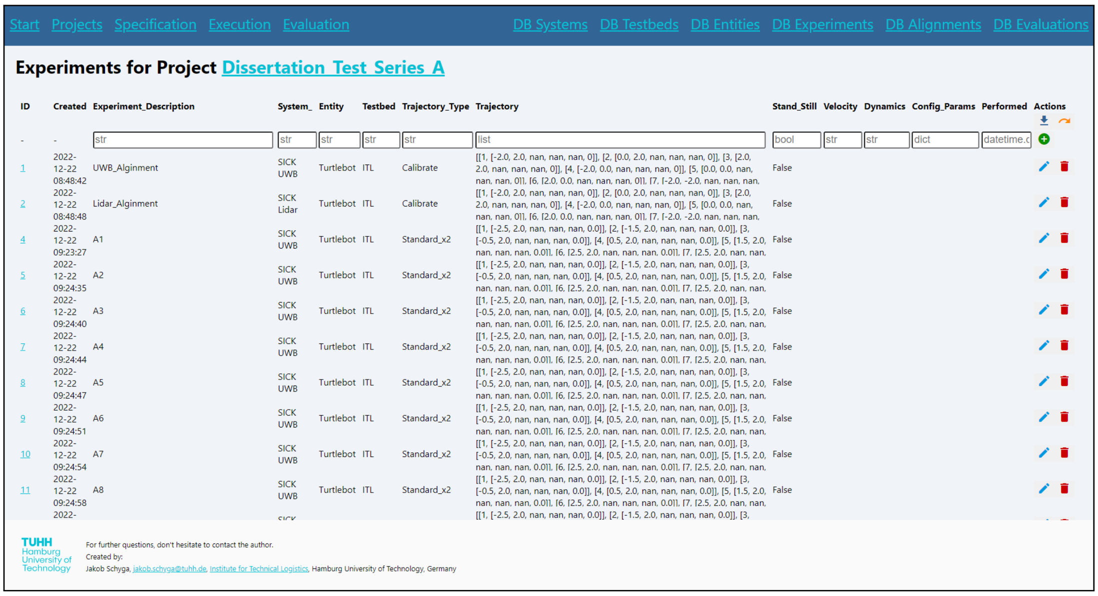


Figure A.5: Exemplary screenshot of *DB Experiment* tab for experiment management