

Fakultät II - Informatik, Wirtschafts- und Rechtswissenschaften

Oldenburger Schriften zur Wirtschaftsinformatik  
Hrsg.: Prof. Dr.-Ing. Jorge Marx Gómez

Jose Daniel Alvarez Coello

## **Vehicle Data Semantics for Driving Context Knowledge**





Carl von Ossietzky University of Oldenburg

School II - Computing Science, Business Administration, Economics, and Law  
Department of Computing Science

***Vehicle Data Semantics for Driving Context Knowledge***

The School of Computing Science, Business Administration, Economics and Law  
of the University of Oldenburg has accepted as partial fulfillment of the requirements  
for the degree and title of

**Doctor of Engineering (Dr.-Ing.)**

the above-mentioned thesis by

**Jose Daniel Alvarez Coello**

born on 22.04.1988 in Cuenca, Ecuador

Supervisors: Prof. Dr.-Ing. Jorge Marx Gómez  
Prof. Dr. Frank Köster

Date of thesis defence: 09.12.2022

Oldenburger Schriften zur Wirtschaftsinformatik

Band 34

**Jose Daniel Alvarez Coello**

**Vehicle Data Semantics  
for Driving Context Knowledge**

Shaker Verlag  
Düren 2023

**Bibliographic information published by the Deutsche Nationalbibliothek**

The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at <http://dnb.d-nb.de>.

Zugl.: Oldenburg, Univ., Diss., 2022

Copyright Shaker Verlag 2023

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior permission of the publishers.

Printed in Germany.

ISBN 978-3-8440-8967-7

ISSN 1863-8627

Shaker Verlag GmbH • Am Langen Graben 15a • 52353 Düren

Phone: 0049/2421/99011-0 • Telefax: 0049/2421/99011-9

Internet: [www.shaker.de](http://www.shaker.de) • e-mail: [info@shaker.de](mailto:info@shaker.de)

## Part I

### Front matter



# Abstract

Data is an asset, and its relevance is acknowledged worldwide. Modern vehicles represent a tremendous data source that provides insights into the driving context. Despite its value, data must be prepared before it can support decision-making. Data and software architectures in the automotive industry have been traditionally application-centric. This type of architecture causes repeated data engineering processes due to the missing adoption of standards and the lack of semantic data models. Consequently, several isolated data models are developed to serve specific use cases.

There is currently no standard semantic data model to cope with the variety of the vehicle's dynamic properties and the outcomes obtained by applications that process them. Additionally, such properties produce data streams that impose the so-called one-pass constraint on the architecture. It is then crucial to either process, store, or forward the observed data; otherwise, it is lost forever. Since vehicles have constrained computation, storage, and bandwidth resources, data streams must be handled appropriately to maximize their use.

A data model alone is useless. In practice, all the entities and relationships represented in a data model must be complemented with techniques capable of extracting them from the actual data. On the one hand, the existing semantic models for vehicle data have focused primarily on static information, missing out on the representation of the temporal domain. Moreover, other models applicable to sensors suggest the semantic annotation of raw sensor data, which implies increased volume without adding the semantics of the data sequences. On the other hand, plenty of approaches exist to process the data to extract its meaning. Some ignore the one-pass constraint by accepting arbitrary iterations, while others focus on specific tasks and are not generically applicable. Therefore, it is desirable to rapidly process, interpret, and semantically annotate data streams to facilitate further use.

This thesis studies how to design an artifact that interprets vehicle data streams and represents their meaning in a human-understandable way so that instant, actionable, and reusable facts are obtained in the vehicle architecture. Following the Design Science Research methodology, it proposes an approach for working with vehicle data streams at a higher abstraction layer than raw sensor data. The proposed design consists of a semantic model for vehicle data and its dynamic properties, a stream analysis process for continuous data aggregation, a semantic annotation step for adding explicit context to the streams, and a knowledge formalization step in the form of rules for representing the logic behind the desired information.

Besides the validation, the design is tested on two use cases related to stream reasoning tasks. The first is about accessing high-level information inferred from vehicle data streams, whereas the second is about supporting decision-making on the highway. Altogether, this study encounters some implications. The data stream abstraction enables individuals without domain knowledge to better understand the meaning behind vehicle data for further use. Consequently, a middle semantic layer in the vehicle architecture decouples raw data from data consumers, alleviating the burden of application logic for common reasoning tasks. The resulting graph data volume is, in most cases,



reduced compared to the semantic annotation of individual observations at the cost of information loss. Moreover, adopting and supporting the proposed semantic model for vehicle data will be a vital step towards a data-centric architecture in the automotive industry.

# Zusammenfassung

Daten sind ein Vermögenswert, und ihre Bedeutung ist weltweit anerkannt. Moderne Fahrzeuge stellen eine enorme Datenquelle dar, die Einblicke in den Fahrkontext bietet. Trotz ihres Wertes müssen die Daten vorbereitet werden, bevor sie die Entscheidungsfindung unterstützen können. Die Daten- und Software-Architekturen in der Automobilindustrie waren bisher anwendungsorientiert. Diese Art von Architektur verursacht wiederholte Data-Engineering-Prozesse aufgrund der fehlenden Übernahme von Standards und des Mangels an semantischen Datenmodellen. Infolgedessen werden mehrere isolierte Datenmodelle entwickelt, um spezifische Anwendungsfälle zu bedienen.

Derzeit gibt es kein standardisiertes semantisches Datenmodell, das die Vielfalt der dynamischen Eigenschaften des Fahrzeugs und die Ergebnisse der sie verarbeitenden Anwendungen bewältigen kann. Darüber hinaus erzeugen solche Eigenschaften Datenströme, die der Architektur die so genannte One-Pass-Beschränkung auferlegen. Es ist dann entscheidend, die beobachteten Daten entweder zu verarbeiten, zu speichern oder weiterzuleiten; andernfalls sind sie für immer verloren. Da Fahrzeuge nur über begrenzte Rechen-, Speicher- und Bandbreitenressourcen verfügen, müssen die Datenströme entsprechend gehandhabt werden, um ihre Nutzung zu maximieren.

Ein Datenmodell allein ist nutzlos. In der Praxis müssen alle Entitäten und Beziehungen, die in einem Datenmodell dargestellt sind, durch Techniken ergänzt werden, die in der Lage sind, sie aus den tatsächlichen Daten zu extrahieren. Einerseits haben sich die bestehenden semantischen Modelle für Fahrzeugdaten vor allem auf statische Informationen konzentriert und die Darstellung des zeitlichen Bereichs vernachlässigt. Außerdem schlagen andere Modelle, die auf Sensoren anwendbar sind, die semantische Annotation von Sensor-Rohdaten vor, was eine Erhöhung des Volumens bedeutet, ohne dass die Semantik der Datensequenzen hinzugefügt wird. Andererseits gibt es eine Vielzahl von Ansätzen zur Verarbeitung der Daten, um ihre Bedeutung zu extrahieren. Einige ignorieren die One-Pass-Beschränkung, indem sie beliebige Iterationen zulassen, während andere sich auf spezifische Aufgaben konzentrieren und nicht allgemein anwendbar sind. Daher ist es wünschenswert, Datenströme schnell zu verarbeiten, zu interpretieren und semantisch zu annotieren, um die weitere Nutzung zu erleichtern.

In dieser Doktorarbeit wird untersucht, wie ein Artefakt entworfen werden kann, das Fahrzeugdatenströme interpretiert und ihre Bedeutung in einer für den Menschen verständlichen Weise darstellt, so dass sofortige, umsetzbare und wiederverwendbare Fakten in der Fahrzeugarchitektur gewonnen werden. In Anlehnung an die Methodik der Design Science Research wird ein Ansatz für die Arbeit mit Fahrzeugdatenströmen auf einer höheren Abstraktionsebene als den rohen Sensordaten vorgeschlagen. Der vorgeschlagene Design besteht aus einem semantischen Modell für Fahrzeugdaten und ihre dynamischen Eigenschaften, einem Stream-Analyseprozess zur kontinuierlichen Datenaggregation, einem semantischen Annotationsschritt zum Hinzufügen von explizitem Kontext zu den Streams und einem Wissensformalisierungsschritt in Form von Regeln zur Darstellung der Logik hinter den gewünschten Informationen.

Neben der Validierung wird das Design an zwei Anwendungsfällen getestet, die

mit Stream-Reasoning-Aufgaben zusammenhängen. Im ersten Fall geht es um den Zugriff auf High-Level-Informationen, die aus Fahrzeugdatenströmen abgeleitet werden, während es im zweiten Fall um die Unterstützung der Entscheidungsfindung auf der Autobahn geht. Insgesamt stößt diese Studie auf einige Implikationen. Die Abstraktion von Datenströmen ermöglicht es Personen ohne Fachwissen, die Bedeutung von Fahrzeugdaten für die weitere Verwendung besser zu verstehen. Folglich entkoppelt eine mittlere semantische Ebene in der Fahrzeugarchitektur die Rohdaten von den Datenkonsumenten und entlastet die Applikationslogik für allgemeine Schlussfolgerungsaufgaben. Das resultierende Graphdatenvolumen wird in den meisten Fällen im Vergleich zur semantischen Annotation einzelner Beobachtungen reduziert, allerdings auf Kosten von Informationsverlusten. Darüber hinaus ist die Übernahme und Unterstützung des vorgeschlagenen semantischen Modells für Fahrzeugdaten ein wichtiger Schritt auf dem Weg zu einer datenzentrierten Architektur in der Automobilindustrie.

# Declaration

I declare that I have written this thesis independently, that I have not used any sources or aids other than those indicated, and that I have followed the general principles of scientific work and publications as stipulated in the guidelines of good scientific practice of the Carl von Ossietzky University of Oldenburg.

A handwritten signature in blue ink, appearing to be 'JD Coello', written in a cursive style.

*Jose Daniel Alvarez Coello*



# Contents

## I Front matter

Abstract	i
Zusammenfassung	iii
Declaration	v
List of Figures	xiii
List of Tables	xv
Code Listings	xvii
Acronyms	xix

## II Main matter

<b>1 Introduction</b>	<b>1</b>
1.1 Structure of this thesis	2
1.2 Nomenclature and conventions	3
1.2.1 Figures' informative boxes	3
1.2.2 Prefixes and namespaces	3
1.2.3 Rule language	3
1.2.3.1 Rules in a nutshell	4
1.2.3.2 Example rule	4
1.2.3.3 About rules in this document	5
1.2.4 Text abbreviations	6
1.3 Motivation	6
1.4 Problem context	9
1.5 Goal	11
1.6 Solution requirements	12
1.7 Background	13
1.7.1 Differences between Data, Information and Knowledge	13
1.7.1.1 An illustrative example of DIKW	13
1.7.2 Knowledge Representation	15
1.7.2.1 Syntax	15
1.7.2.2 Semantics	16
1.7.2.3 Pragmatics	16
1.7.2.4 Ontology	16
1.7.2.5 Graph data	16
1.7.3 Data streams	17
1.7.3.1 Stream processing	17

1.7.3.2	Stream reasoning . . . . .	17
<b>2</b>	<b>Methodology</b>	<b>19</b>
2.1	Design Science Research methodology . . . . .	19
2.1.1	Types of artifacts . . . . .	20
2.1.2	Research framework . . . . .	20
2.1.2.1	Design cycle . . . . .	21
2.2	Research Questions . . . . .	22
2.2.1	Main Research Question . . . . .	22
2.2.2	Knowledge Questions . . . . .	22
2.2.2.1	Problem-related KQ . . . . .	22
2.2.2.2	Test-related KQs . . . . .	23
2.3	Research methods . . . . .	23
2.4	Assumptions and other aspects . . . . .	23
<b>3</b>	<b>Problem investigation</b>	<b>25</b>
3.1	Related areas . . . . .	25
3.2	Semantization of data streams . . . . .	26
3.2.1	Semantization of individual observations . . . . .	27
3.2.1.1	Data models at low-level of abstraction . . . . .	28
3.2.2	Semantization of analytical outcomes . . . . .	29
3.2.2.1	Data models at high-level of abstraction . . . . .	30
3.2.3	Other modeling approaches . . . . .	30
3.2.3.1	AUTOSAR . . . . .	30
3.2.3.2	NGSI-LD information model . . . . .	30
3.2.3.3	ASAM standards . . . . .	30
3.3	Analysis of data streams . . . . .	30
3.3.1	Statistical summaries . . . . .	31
3.3.1.1	Data compression using statistics . . . . .	32
3.3.2	Machine Learning methods on streams . . . . .	32
3.3.2.1	Deep learning . . . . .	32
3.3.2.2	A note on datasets for ML . . . . .	33
3.3.3	Stream mining . . . . .	33
3.3.4	Stream processing with windowing functions . . . . .	33
3.3.5	Stream reasoning . . . . .	34
3.4	Requirements satisfaction of existing approaches . . . . .	35
3.5	A note on information systems . . . . .	36
3.5.1	Steaming platforms . . . . .	36
3.5.2	Cloud services for connected vehicles . . . . .	38
3.5.3	Graph databases . . . . .	38
<b>4</b>	<b>Solution approach</b>	<b>39</b>
4.0	Approach overview . . . . .	39
4.0.1	A zoomed-out view of the solution approach . . . . .	39
4.0.2	A zoomed-in view of the solution approach . . . . .	42
4.1	Semantic model . . . . .	42
4.1.1	VSSo 2.0: A core ontology in sync with VSS . . . . .	43
4.1.2	XDVP: An ontology eXtension for Dynamic Vehicle Properties . . . . .	45
4.1.2.1	Scope . . . . .	45
Domain & Intended use. . . . .	45	
Competency questions. . . . .	46	
4.1.2.2	Reuse . . . . .	46

	<i>Reusing VSSo as the basis.</i> . . . . .	46
	Reusing SSN and SOSA for annotation of individual observations. . . . .	47
	Reusing IoT-Stream for time intervals and streams. . . . .	47
	Reusing QUDT for units and quantity kinds. . . . .	47
4.1.2.3	Terms . . . . .	48
4.1.2.4	Classes . . . . .	48
4.1.2.5	Properties . . . . .	48
4.1.2.6	Constraints . . . . .	49
4.1.2.7	Instances . . . . .	49
4.1.2.8	Joining all modeling pieces together . . . . .	50
	<i>Overall view of entities for static and dynamic data.</i> . . . .	50
	<i>Data streams and their features.</i> . . . . .	50
	<i>Introducing variable kinds.</i> . . . . .	52
	<i>Data records hold the values of a data stream.</i> . . . . .	52
	<i>Application produces or consumes a data stream.</i> . . . . .	52
4.2	Stream analysis . . . . .	53
4.2.1	Data stream management . . . . .	54
4.2.2	Analysis of numerical data streams . . . . .	55
4.2.2.1	Exact mean and standard deviation . . . . .	55
4.2.2.2	Trend detection with dynamic threshold . . . . .	56
4.2.2.3	Encoding trends and temporal information . . . . .	56
	Trend letter and words . . . . .	57
	Trend codes and their sum . . . . .	57
	Decision difference and their sum . . . . .	57
	Ratio of slopes and their sum . . . . .	57
	Temporal mask . . . . .	57
4.2.2.4	Symbolic representations . . . . .	57
4.2.2.5	Feature vectors for complex models . . . . .	57
4.2.3	Analysis of categorical data streams . . . . .	58
4.3	Semantic annotation . . . . .	58
4.3.1	Annotation of numerical data streams . . . . .	58
4.3.2	Annotation of categorical data streams . . . . .	59
4.4	Knowledge formalization . . . . .	59
4.4.1	Extracted concepts from a numerical data stream . . . . .	59
4.4.2	Extracted concepts from a categorical data stream . . . . .	59
4.4.3	Examples . . . . .	59
<b>5</b>	<b>Validation</b> . . . . .	<b>61</b>
5.1	Validating the VSSo core model . . . . .	61
5.1.1	Expert opinion on VSSo core . . . . .	61
5.1.2	Query comparison between VSSo 1 vs. VSSo 2 . . . . .	61
5.1.2.1	ExQ 1: How many seats does this <i>Vehicle</i> have? . . . . .	62
5.1.2.2	ExQ 2: What is the local current temperature on the driver side? . . . . .	63
5.1.2.3	ExQ 3: What properties in this vehicle refer to “ <i>speed</i> ”? . . . . .	64
5.2	Validating the XDVP model . . . . .	65
5.2.1	Expert opinion on XDVP . . . . .	65
5.2.2	Competency questions’ satisfaction with SPARQL queries . . . . .	65
5.2.2.1	Common selection queries . . . . .	65
	<i>Selecting a specific Vehicle.</i> . . . . .	65
	<i>Selecting the latest VehicleJourney.</i> . . . . .	65



	<i>Selecting a specific DataStream of the latest VehicleJourney.</i>	66
5.2.2.2	Answering about background knowledge (i.e., static) . . .	66
	<i>How many driving journeys does a vehicle have in a particular period?</i> . . . . .	66
	<i>What are the topics of the available streams?.</i> . . . . .	67
	<i>What analytics are applied to a particular stream?.</i> . . . . .	67
	<i>Who are the producers and consumers of a particular stream?</i> 68	
	<i>What features are present in a given stream, and what are their quantity type and unit associated?</i> . . . . .	68
5.2.2.3	Answering about data stream values . . . . .	68
	<i>What data records of a data stream are available in the latest journey?</i> . . . . .	68
	<i>How long has the latest seen value of a given categorical feature remained unchanged?</i> . . . . .	69
	<i>What is the value of a feature (e.g., average) in a given time window?.</i> . . . . .	69
5.3	Analyzing the use of resources . . . . .	69
5.3.1	Annotations per stream . . . . .	69
5.3.1.1	Annotations of a categorical stream . . . . .	70
5.3.1.2	Annotations of a numerical stream . . . . .	70
5.3.2	Graph data volume comparison . . . . .	71
5.3.2.1	Graph data sizes from a categorical data stream . . . . .	72
5.3.2.2	Graph data sizes from a numerical data stream . . . . .	74
5.4	Building the data graph . . . . .	75
5.5	Complex prediction tasks . . . . .	77
<b>6</b>	<b>Use case one: Accessing high-level information inferred from vehicle data streams</b>	<b>79</b>
6.1	Use case overview . . . . .	79
6.2	Reasoning tasks . . . . .	80
6.2.1	Vehicle motion . . . . .	80
6.2.2	Semantic speed . . . . .	81
6.2.3	Idle mode . . . . .	82
6.2.4	Road slope . . . . .	83
<b>7</b>	<b>Use case two: Supporting decision making on the highway</b>	<b>85</b>
7.1	Use case overview . . . . .	85
7.2	Reasoning tasks . . . . .	86
7.2.1	Distance-based states . . . . .	87
7.2.2	Safety conditions' satisfaction . . . . .	89
7.2.3	Decision recommendation . . . . .	90
<b>8</b>	<b>Summary</b>	<b>91</b>
8.1	Answers to Research Questions . . . . .	91
8.1.1	Answer to main Research Question . . . . .	91
8.1.2	Answer to Knowledge Questions . . . . .	93
8.1.2.1	Problem-related KQ . . . . .	93
8.1.2.2	Test-related KQs . . . . .	94
8.2	Discussion . . . . .	96
8.2.1	Advantages of using the proposed design . . . . .	96
8.2.2	Limitations in the proposed design . . . . .	96
8.3	Future work . . . . .	97

III	Back matter	
A	RML mapping rules	103
B	Graph data experiment test results	107
	Bibliography	113