Digital Shadows for Operational Technology Monitoring – A Comparison of Data-Driven Approaches*

David Graf
Johannes Kepler University Linz
david.graf@jku.at

Wieland Schwinger
Johannes Kepler University Linz
wieland.schwinger@jku.at

Werner Retschitzegger
Johannes Kepler University Linz
werner.retschitzegger@jku.at

Elisabeth Kapsammer
Johannes Kepler University Linz
elisabeth.kapsammer@jku.at

Norbert Baumgartner
team GmbH
norbert.baumgartner@te-am.net

Birgit Pröll
Johannes Kepler University Linz
birgit.proell@jku.at

Johannes Schönböck
University of Applied Sciences Upper Austria
johannes.schoenboeck@fh-hagenberg.at

Abstract

Critical infrastructures in domains like road traffic management heavily depend on Operational Technology (OT) to ensure safe operation. One faces, however, also tremendous challenges in OT monitoring (OTM), i.e., ensuring the proper functioning of the OT objects themselves, due to their inherent large-scale, heterogeneous, and evolutionary nature. Going beyond the current practice of monitoring single OT object states, the digital twin paradigm could enable a more holistic OTM - research being, however, still in its infancy. Thus, the contribution of this paper is threefold: Firstly, we discuss key challenges from a domain perspective and derive appropriate criteria for a systematic evaluation of data-driven approaches aiming at a digital representation of an OT infrastructure. Secondly, based on these criteria, we identify and discuss promising approaches, ranging from IT networks to Social Networks. Thirdly, based thereupon, we present lessons learned and open issues for further research.

Keywords: Operational Technology Monitoring, Road Traffic Management, Digital Shadow, Data-Driven Generation and Evolution

1. Introduction

Operational Technology Monitoring. Critical infrastructures like road traffic management use a wide range of Operational Technology (OT), such as traffic flow sensors, video surveillance systems, or digital road signs (Murray et al., 2017), for monitoring and controlling their smooth operation (Baumgartner et al., 2014) (cf. Figure 1). At least as important is the monitoring of the proper function of the OT infrastructure itself aka. *"Operational Technology Monitoring (OTM)"*, being the crucial prerequisite for ensuring an efficient and safe operation of the whole critical infrastructure (cf. Figure 1).

Digital Twin Paradigm for Holistic OTM. In practice, OTM mainly focuses, however, often on monitoring the states of individual OT objects, only (Pliatsios et al., 2020), leading to rather *isolated monitoring perspectives*, which aggravate diagnosis of real failure scenarios behind. Going beyond monitoring of single OT object states, i.e., *isolated OTM*, the *digital twin paradigm* seems to provide a promising solution striving towards a more *holistic OTM*. This would build the basis for reasoning about, e.g., failure root causes, impacts, effects, and criticality wrt. inter-dependent OT objects as well as for possible failure resolutions (cf. Figure 1). The benefits of digital twins have already been shown in a broad range of other domains like manufacturing (Brecher et al., 2021), cyber-physical systems (Josifovska et al., 2019; Steinmetz et al., 2018), software applications (Dalibor et al., 2022), and even natural systems (Bordeleau et al., 2020), thereby covering various use cases, ranging from monitoring and controlling a physical entity via simulating, testing and optimizing certain configurations to predicting future states and supporting decision making (Wen et al., 2022).

* Research jointly supported by ASEA-UNINET, BMBWF & OeAD [ASES 2022-2023/Uni Linz/2].

Digital Shadows for OTM Not Readily Available. Despite these various benefits, research on digital representations of OT infrastructures especially in the domain of road traffic management is still in its infancy and example use cases are lacking (Boyes and Watson, 2022; Rivera et al., 2022; Wen et al., 2022). This is not least because *digital twins often focus on rather small physical entities* (Boyes and Watson, 2022) and manually modeling complex systems or networks is not feasible, thus calling for *automatic, data-driven approaches* (Wen et al., 2022). However, not even for *digital shadows* (Kritzinger et al., 2018), the least ambitious form of a digital twin, it is clear how to automatically put forward a suitable digital model, capturing foremost the *structural aspects of OT objects and their inter-dependencies*, before dealing also with *behavioral aspects*, not least because of the inherent large-scale, heterogeneous, and evolutionary nature of our problem domain.

Contribution and Paper Structure. Therefore, the contribution of this paper is threefold: Firstly, we discuss key challenges from a domain perspective and derive appropriate criteria for a systematic evaluation of data-driven approaches, possibly suitable for the provision of a digital shadow of an OT infrastructure in road traffic management (cf. Section 2). Secondly, based on these criteria, in Section 3 we identify and discuss promising approaches, ranging from IT networks to social networks, for putting forward a digital shadow. Thirdly, based thereupon, we present lessons learned and open issues for further research in Section 4. Finally, Section 5 summarizes our main findings.

2. Challenges and Evaluation Criteria

Based on our long-standing experience regarding *situation awareness in road traffic management* (Baumgartner et al., 2010; Baumgartner et al., 2014; Graf et al., 2019) as well as *automatic conceptualizations of OT infrastructure models* in particular (Graf et al., 2022; Graf et al., 2022a, 2022b), we first identify the most relevant *domain challenges* in the following (cf. Figure 1), grounded in the specific characteristics of road traffic management. In a second step, we derive according *requirements* for the data-driven provision of a digital shadow focusing on structural aspects of an OT infrastructure in road traffic management in form of *evaluation criteria*, also backed up by existing surveys in the area of digital twins (Dalibor et al., 2022; Lugaresi and Matta, 2021a; Minerva et al., 2020; Wen et al., 2022) (cf. Figure 2). As underlying *rational behind categorizing* these challenges and the resulting criteria, we distinguish between the *static view of digital shadow provision* in terms of its *"generation"* on basis of an assumed non-changing OT infrastructure (cf. Subsection 2.1) and the *dynamic view of digital shadow provision* taking its *"evolution"* into account (cf. Subsection 2.2).

2.1. Generation of a Digital Shadow

First of all, the *generation of a digital shadow* in its simplest form in terms of a *static snapshot* of the structure of OT objects evidently faces two categories of challenges in our problem domain, as discussed in the following.

(1) Challenges of Scalability and Knowability. When considering a domain like road traffic management, the large scale and geographical distribution in terms of sheer size of such environments is naturally a major challenge. As an example, road traffic management operators in Austria are responsible for 2.249 km of motorways and expressways including 166 tunnels and more than 100.000 OT objects of more than 200 different object types, resulting in a stream of ten thousands of messages per hour recorded in log files of various systems.

When considering the state-of-practice (see Graf et al., 2022a), knowledge about all these OT objects and specifically about their inter-dependencies, is, if at all, frequently *limited to type-level*, only, and scattered across some (experienced) human operators. In case knowledge about inter-dependencies is explicitly documented, it is seldom available in a consistent *machine-processable* form, but rather in *proprietary formats* originating from *different sources*.

Consequently, from a practical point of view, it is not feasible to manually conceptualize the OT infrastructure in terms of a digital shadow with hundred thousands of objects and inter-dependencies in-between. Thus, inevitably, *automatic data-driven approaches* for the generation of a digital shadow are needed, facing the challenge of operating on data derived from operational artifacts of the OT infrastructure such as *log files* (cf. Graf et al., 2022a).

(2) Challenge of Heterogeneity. Furthermore, such large-scale and dynamic environments are naturally characterized by massive heterogeneity at different levels. For example, a single highway tunnel consists, on the one hand side, of thousands of diverse OT objects of different types, comprising *simple sensors* and *actuators* (e.g., fire detectors, ventilators, traffic jam detectors), *IT infrastructure* like gateways and servers, as well as, on the other hand side, of *complex software systems*, e.g, for video, noise, and air surveillance.

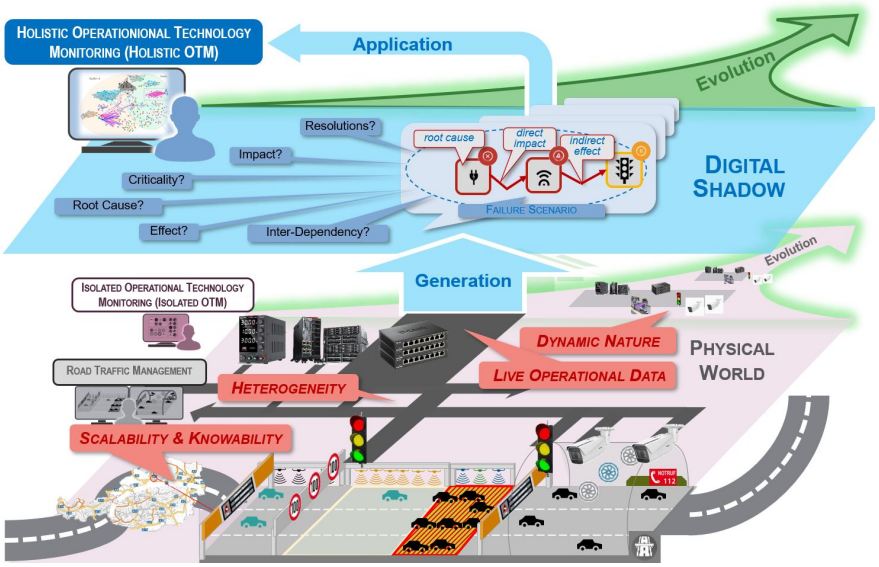

Figure 1. Employing Digital Shadows for Operational Technology Monitoring

Naturally, to realize complex OTM behaviour, these multiple OT objects are *inter-connected*, *directly or indirectly*, with a series of other OT objects in various ways, by *providing sensor data* (e.g., temperature measurement), *controlling input* (e.g., activating actuators), and *communicating* with each other (e.g., network hub), or are simply *physically dependent* on the proper working of others (e.g., power supply). On top of this, these OT objects *contextually interact differently*, i.e., having distinct inter-dependencies in divergent OTM situations.

Besides heterogeneity induced by different OT object types and inter-dependencies in between, the lifespan of OT, e.g., in a highway tunnel, is usually several decades, resulting in *legacy systems* which are naturally mixed up with newer types of OT of possibly different manufacturers.

Consequently, when generating a digital shadow derived from operational artifacts such as log files, being ultimately the least-common-denominator input, approaches are required to provide for an *appropriate abstraction* and reasonable *differentiation of OT objects* and their *inter-dependencies*, in the course of the digital shadow generation process as well as in form of a *semantically rich representation* in the generated digital shadow itself (cf. Graf et al., 2022).

Evaluation Criteria for Generation. The challenges described above entail that approaches for digital shadow generation need to be able to deal with certain characteristics of log data in terms of which kind of *"Input"* data (cf. Figure 2) is provided, the *"Quantity"* of data handled to give an indication of scalability, spanning from the number of log entries to the number of concrete OT objects therein (cf. below), and the data *"Quality"* assumptions, made with respect to unreliability, sparsity, and incompleteness as found in our problem domain. Specifically, regarding the *"Source"* of input data, the applicability of an approach is constrained, on the one hand side, by what kind of information is expected as input, being low-level OT *"Messages"*, distinct *"Events"* (i.e., aggregations of different messages reporting about the same real-world state), or mere observations of OT *"Objects"* and their *"States"*, as well as its ability to incorporate *"Additional Data"*, foremost in form of type-level information, and on the other hand side, in which way it is provided *("Data Ingestion")*, i.e., rather batch- or stream-oriented.

Turning from input data to *"Processing"* issues, an approach might address these in terms of *"Pre-Processing"*, aiming at *"Data Reduction"*, transcribing the representation often called *"Data Binarization"*, and partitioning the data in terms of *"Data Windowing"*. The lack of known inter-dependencies between OT objects can be tackled by applying different *"Computational Methods"*, ranging from probabilistic via statistical and similarity-based to algorithmic methods. Regardless of the employed method, a key determining factor of an approach's processing is its capability to finally differentiate between various types of inter-dependencies in some *"Post-Processing"* phase.

While the applicability of an approach is determined by the previous criteria, the usefulness of the digital

shadow is ultimately determined by the *"Output"* of the approach in terms of a structural conceptual model of the OT infrastructure. Using terminology from the *UML meta model for class diagrams*, we distinguish on the one hand side *"Core Elements"* of the digital shadow which are generated, i.e, *"Objects"*, their inter-dependencies in terms of *"Links"*, and current *"States"* of these objects. On the other hand side and most importantly, the approach should provide *"Additional Semantics"* about these core elements, along with digital shadow's *"Fidelity"* (Rivera et al., 2020), meaning its conformance to reality.

Finally, Figure 2 shows an example for the generation of a digital shadow. As time proceeds from t_1 to t_3, more and more OT objects are encountered in the log file and reflected within the digital shadow, together with inter-dependencies, which are established at t_3 as links having "controls" and "supplies" semantics.

2.2. Evolution of a Digital Shadow

Operating an OT infrastructure is naturally a dynamic affair that manifests itself in two challenges which need to be addressed conjunctively on top of the previously stated ones.

(1) Challenge of Dynamic Nature. Overall, OT objects are subject to *unintended changes* (i.e., failures of parts of OT) as well as affected by *intended changes* (i.e., OT objects are added, removed, or changed) on a daily basis (cf. Graf et al., 2019). The former includes failures of OT objects itself, e.g, a camera break down, as well as OT impacted by the failure of other OT, e.g., the camera is not able to connect to the image server due to a network switch break down. The latter, intended changes, are primarily related to maintenance actions, for instance, a traffic sensor employed at the roadside is removed because of its expected "end of life", as well as applying new technology because of some highway sections are renewed or refurbished.

Naturally enough, it is clearly crucial and most important for OTM that the structural information in the *digital shadow needs to co-evolve with the physical world*. This includes specifically both, OT objects as well as their inter-dependencies that need to be constantly adjusted to reflect the current OT infrastructure accordingly.

(2) Challenge of Live Operational Data. On top of that, from a practical point of view, operational data can be collected during live operation of the OT infrastructure, only. This is aggravated by the fact that operational data is often recorded in operational log files when "something interesting" happened, only. Thus, to a large extent, log files contain data resulting from unintended changes aka. "state messages" that report on failures of OT objects. Regarding intended changes, these are only recorded in case these OT objects report or are reported on actively and thus made available in the operational data (cf. Graf et al., 2022a).

In either case, exploitable data for digital shadow evolution becomes available in the course of system operation over time, resulting in the fact that information about the OT infrastructure is mostly lacking behind and thus delayed and incomplete observable.

Evaluation Criteria for Evolution. The challenge of dynamic nature of the OT infrastructure itself inter-twined with the challenge of life operational data channel into two main requirements wrt. evolution of a digital shadow. Firstly, *if and how it adjusts to changes* of the OT infrastructure *("Adjustment to Evolution")* as being reported in terms of *"Add" and "Remove"* of *"Instance Level"* information about objects and their links, but also about *"Type Level"* information, and about the current *"State"* of the OT objects, respectively.

A scenario of adjustment to evolution at instance level is also exemplary depicted in Figure 2, showing the evolution of the digital shadow at t_4, where an OT object in the physical world ("ups_7") is replaced by another one ("ups_8") (i.e., "Remove" and "Add"), requiring within the digital shadow not only a replacement of the former OT object but also the replacement of two existing links.

While adjusting to evolution is fundamental to reflect the current state of the physical OT infrastructure within the digital shadow, the dynamic nature of the OT infrastructure is only conveyed completely, if, secondly, thereupon the *consequence of evolution* of the OT infrastructure, again in terms *"Add" and "Remove"*, is explicated in the digital shadow *("Explication of Evolution")*, for each of the core concepts, respectively.

All criteria motivated above are used in the following to systematically identify and evaluate approaches possibly suitable for the provision of digital shadows for an OT infrastructure in road traffic management.

3. Evaluation of Approaches

Approaches for the provision of digital twins can be, according to (Bordeleau et al., 2020), categorized into *(i) component solutions* provided, among others, by Microsoft, Amazon and Eclipse, *(ii) off-the-shelf solutions* often provided directly by OEMs for common industry use cases, and finally *(iii) custom-based solutions*. While it would be preferable to resort to component- or off-the-self solutions as they provide functionality out of the box, they are, due to challenges prevalent in our domain, not reasonably applicable.

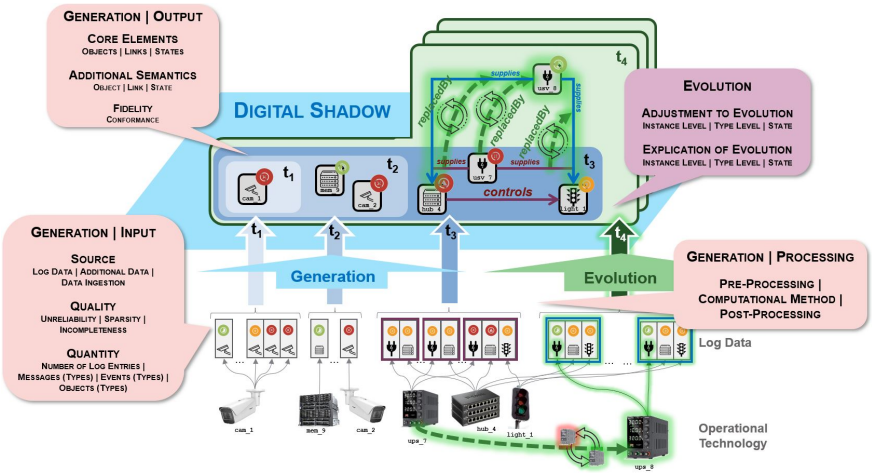

Figure 2. Criteria for Digital Shadow Generation and Evolution

Hence, since appropriate custom-based solutions for representing OT infrastructures especially in the domain of road traffic management are still in their infancy and example use cases are lacking (Boyes and Watson, 2022; Rivera et al., 2022; Wen et al., 2022), we broaden our investigation, adhering to (Wen et al., 2022), by identifying relevant approaches in domains part of the wide area of *"complex networked systems"*, thereby deliberately going beyond our target domain of OTM in road traffic management.

In any case, however, possibly eligible approaches need to have in common to tackle a conceptualization of structural aspects of objects and their inter-dependencies from log files in a data-driven way. Therefore, we carefully selected nine approaches to be evaluated from the following domains stretching from *(i) Transportation networks* and *(ii) IT networks* to *(iii) IoT networks*, *(iv) Manufacturing networks*, and *(v) Social networks*. Complementing this selection, we also included our own approach (Graf et al., 2022), targeting OT infrastructures in road traffic management.

It has to be noted that, although the requirements imposed on a digital representation in each of these domains are naturally partly different and also the considered approaches tackle them from diverse angles, there is some common ground which we try to exploit on basis of our criteria catalogue. Thus, in the following, based on this criteria catalogue, we give a brief coherent summary of each of these approaches, thereby especially emphasizing the most distinguishing aspects as well as highlighting applicability issues with regard to our problem domain. The complete, systematic evaluation results can be found in Figure 3.

Rivera et al., 2020. A comprehensive reference model to create a digital shadow for a *transportation network* is proposed by Rivera et al. In contrast to many other approaches solely focusing on modeling the behavioral aspects of a digital shadow, they additionally *focus on modeling its structural aspects*, being primarily relevant for our work in the light of modeling OT objects and their inter-dependencies. They allow for *domain knowledge inclusion*, being applicable for *batch- as well as stream-based data*. Since their work presents a reference model, implementation details to mine structural information from log data are lacking as well as concrete methods to tackle evolution of a digital shadow. Nevertheless, they provide a first proposal for *evolution adjustment mechanisms*.

Lugaresi and Matta, 2021b. An approach for generating a digital representation of a *manufacturing line* is provided by Lugaresi & Matta. Their *algorithmic-based approach* uses event information, in a *batch- or stream-based manner*, in order to automatically discover log data of manufacturing systems comprising the *system topology*, which is similar to our use case of mining OT objects and their inter-dependencies. In addition, the approach includes so-called *removal and aggregation rules* in order to address evolution of the underlying systems (e.g., objects are added or removed). Finally, the approach is applied on two test cases and a real manufacturing line.

Messager et al., 2019. A *statistic-based approach* to mine functional dependencies from large event logs of *IT networks* is pursued by Messager et al., especially emphasizing on *generic applicability*, thereby, however, neglecting any form of semantic differentiation. The approach is able to deal with data quality issues like *temporal unreliability* and *sparsity of events*, whereas an application to data streams is not supported. Regarding evolution support, since encountered objects

						Rivera		Lugaresi		Messager		Kobayashi		Harper		Antonello		Song		Psorakis		Hallac		Graf	
Application Domain						Transportation Networks		Manufacturing Networks		IT Networks		IT Networks		IT Networks		IT Networks		IoT Networks		Social Networks		Social Networks		OT Networks	
GENERATION	Input	Source	Log Data	Messages		✓		×		✓		✓		✓		✓		×		×		✓		✓	
				Events		✓		✓		×		×		✓		×		✓		✓		×		×	
				Objects		✓		✓		✓		✓		✓		✓		✓		✓		✓		✓	
				Ratio Messages/Events/Objects		?		?		×		n:m		n:m		×		?		×		×		n:m	
				State information		✓		✓		×		×		×		×		×		×		×		✓	
			Additional Data	Type Information		✓		×		×		✓		✓		×		×		×		×		✓	
				Historical Data		✓		×		×		×		×		×		×		✓		×		✓	
			Data Ingestion	Batch		✓		✓		×		×		×		✓		×		✓		✓		×	
				Semi-Stream		×		×		✓		✓		✓		×		×		×		×		✓	
				Stream		✓		✓		×		×		×		×		✓		×		✓		✓	
		Quantity		Number of Log Entries		2.5M/day		?		13M		35M		2k		18M		?		?		2.2k		2.2k	
				Messages	Message Types	?	?	?	?	?	n.a.	1.8k	?	?	?	13k	n.a.	?	n.a.	?	n.a.	?	n.a.	900k	15k
				Events	Event Types	?	?	?	?	n.a.	n.a.	n.a.	n.a.	?	?	n.a.	n.a.	?	n.a.	?	n.a.	?	n.a.	n.a.	n.a.
				Objects	Object Types	926	?	?	?	13k	n.a.	158	3	?	n.a.	n.a.	n.a.	?	n.a.	?	n.a.	6	n.a.	17k	100
		Quality		Unreliability		?		?		✓		×		✓		?		×		?		?		✓	
				Sparcity		?		?		✓		?		?		✓		×		?		?		✓	
				Incompleteness		?		?		✓		✓		?		?		?		?		?		✓	
	Processing		Pre-Processing	Data Reduction		?		?		×		irrelevant/duplicate		?		×		?		✓		×		×	
				Data Binarization		n.a.		n.a.		✓		✓		✓		n.a.		?		n.a.		✓		✓	
				Data Windowing		n.a.		n.a.		✓		✓		×		×		✓		✓		✓		✓	
			Computational Method	Probabilistic		?		×		✓		✓		✓		×		✓		✓		✓		×	
				Statistical		?		×		✓		✓		✓		✓		✓		✓		✓		✓	
				Similarity-Based		?		×		×		×		✓		×		×		×		×		×	
				Algorithmic		?		✓		×		×		×		✓		×		×		×		×	
			Post-Processing	Type-Differentiation		?		?		×		messageType		eventType		×		×		×		×		✓	
	Output	Core Elements	Objects			?		✓		✓		✓		✓		✓		✓		✓		✓		✓	
			Links			?		✓		✓		✓		✓		✓		✓		✓		✓		✓	
			States			✓		?		×		×		×		?		?		×		?		✓	
		Additional Semantics	Object	Type		?		×		×		×		×		×		×		×		×		✓	
				Clustering		×		×		×		×		×		✓		×		×		×		×	
				Confidence		?		?		?		?		?		?		?		?		?		✓	
			Link	Type		?		?		×		×		×		×		×		×		×		✓	
				Direction		?		?		×		✓		×		×		✓		×		×		✓	
				Confidence		?		?		✓		✓		×		✓		?		?		✓		✓	
			State	Confidence		?		?		?		?		?		?		?		?		?		✓	
		Fidelity	Conformance			✓		✓		?		?		?		?		?		?		?		×	
EVOLUTION	Adjustment to Evolution	Instance Level	Object	Add		✓		✓		✓		✓		×		×		×		×		×		✓	
				Remove		✓		✓		×		✓		×		×		×		×		×		×	
			Link	Add		✓		✓		✓		✓		✓		×		✓		✓		✓		✓	
				Remove		✓		✓		✓		✓		×		×		✓		✓		✓		✓	
		Type Level	Object Type	Add		✓		n.a.		n.a.		×		×		n.a.		n.a.		n.a.		n.a.		×	
				Remove		✓		n.a.		n.a.		×		×		n.a.		n.a.		n.a.		n.a.		×	
			Link Type	Add		✓		?		n.a.		n.a.		n.a.		n.a.		n.a.		n.a.		n.a.		×	
				Remove		✓		?		n.a.		n.a.		n.a.		n.a.		n.a.		n.a.		?		×	
		State	State information			✓		✓		×		×		×		×		×		×		×		✓	
	Explication of Evolution	Instance Level	Object	Add		×		×		×		×		×		×		×		×		×		×	
				Remove		×		×		×		×		×		×		×		×		×		×	
			Link	Add		×		×		×		×		×		×		×		✓		✓		×	
				Remove		×		×		×		×		×		×		×		✓		✓		×	
		Type Level	Object Type	Add		×		n.a.		n.a.		×		×		n.a.		n.a.		n.a.		n.a.		×	
				Remove		×		n.a.		n.a.		×		×		n.a.		n.a.		n.a.		n.a.		×	
			Link Type	Add		×		?		n.a.		n.a.		n.a.		n.a.		n.a.		n.a.		n.a.		×	
				Remove		×		?		n.a.		n.a.		n.a.		n.a.		n.a.		n.a.		n.a.		×	
		State	State information			×		?		×		×		×		×		×		×		×		×	

Legend: ✓ supported × not supported ? unknown n.a. not applicable

Figure 3. Evaluation of Approaches

and inter-dependencies are based on probabilities of occurring events, *emerging objects* and their resulting inter-dependencies can be dealt with. The approach falls short, however, when objects disappear.

Kobayashi et al., 2017. The approach proposed by Kobayashi et al. specifically addresses the identification of *cause-impact relationships* from *IT network* syslog data. They use *supervised machine learning* to apply pre-processing like removing irrelevant and duplicate data together with probabilistic and statistical methods to mine unknown correlations of the internal network structure. It is important to note, that this approach is one out of two, where the output in terms of a graph-based model provides *directed edges* representing inter-dependencies. In contrast to other approaches, types of events are distinguished as a first step towards *semantic differentiation*. Regarding evolution, the approach is *adaptive to changes at instance level* (adding and removing objects) due to the independent (e.g., day-wise) processing of recently received log data.

Harper and Tee, 2019. A similar goal in the area of *IT networks* as Kobayashi et al. is pursued by Harper and Tee in terms of identifying system errors by focusing on the *impact of cascading errors*. The main assumption is, that a failure leads to almost immediate failures at dependent nodes, thus allowing to mine the underlying structure of a system. By applying additional steps during data binarization, their approach

counteracts timestamp fuzziness and delayed log entries, which is a crucial aspect in our domain. Like others, their *statistical and probabilistic approach* was designed for batch execution, but might be able to easily extend to blocking stream processing, i.e., creating independent results for each block. Evolution support is limited in that links can be added in batch mode, only.

Antonello et al., 2021. The *algorithmic-based approach* of Antonello et al., applying a novel association rule mining method, targets again the area of *IT networks* focusing on the identification of *sparse inter-dependencies* between components of various systems. It is important to note, that this approach is one out of two, which identifies groups of functionally dependent components, i.e., *functional clusters*. Like others, their approach relies on structured log files which are processed in a *batch-wise* manner. Although this approach is promising regarding the generation of a digital shadow, evolution support is entirely missing, also due to its operation on batch data, only.

Song et al., 2022. The statistical approach proposed by Song et al. addresses an *IoT network* aiming to mine *frequently changing inter-dependencies* between IoT-based objects. It is important to note, that this approach is able to operate on *streaming data*, where sequences are segmented into time windows, having as a result *directed inter-dependencies*. Regarding evolution, the approach is focused on handling changes with respect to inter-dependencies but there is neither support for emerging and disappearing objects, since the number of objects is fixed over the whole generation process, nor for explication of evolution.

Psorakis et al., 2012. An approach from the area of *social networks* based on timestamped logs and location information is proposed by Psorakis et al. By applying a community detection algorithm, groups of *"objects" having a statistical correlation* are identified. Regarding evolution, similar to Song et al., they *focus on emerging and disappearing inter-dependencies*, but do not deal with emerging and disappearing objects. This approach is worth mentioning due to being one out of two approaches (cf. Hallac et al. below), which *explicate change*, meaning that through comparison of daily community memberships, one is able to *identify change*, *when it took place* and *what changed*.

Hallac et al., 2017. Also targeting social networks, the speciality of Hallac et al. is the *multi-dimensional input data*. It was designed to infer "time-varying" social networks from timestamped sensor readings, focusing on *streaming aspects*. Beyond that, *explication of evolution* is supported, in that, based on the temporal deviation values between consecutive timestamps, *change points can be derived*. To be more concrete, it can be explicated that a *change happened*, *when it happened*, *what changed*, *at what scope* (e.g., globally or locally), and *how fast it changed*. Despite these evolution explication features, *adjustment to evolution is limited* to changing links between *fixed objects*.

Graf et al., 2022. Finally, rounding up this evaluation, we include our own approach, aiming at automatic conceptualizations of *OT infrastructure models in road traffic management*. To allow for a semantically rich OT infrastructure model, being crucial for the purpose of OTM, we follow a *hybrid approach* combining a *data-driven method* grounded on temporal coincidences in log data, with an exploitation of a *structural meta model of the OT infrastructure* along the entire generation process. This allows for a semantic differentiation of inter-dependencies, forming the basis for a proper application of the digital shadow in terms of, e.g., root cause analysis or predictive maintenance. As a first step towards evolution support, domain-specific knowledge can be dynamically adjusted for coping with emerging and disappearing OT objects and their inter-dependencies. Finally, the applicability of the approach has been demonstrated on bases of real-world log data of an OT infrastructure.

4. Lessons Learned and Open Issues

In the following, we change our perspective from an approach-centered discussion towards a bird-eyes view on existing research. Thereby, we present lessons learned from previously discussed approaches as well as open issues along our criteria catalogue, again categorized into generation and evolution of digital shadows (cf. Figure 4, for an overview).

4.1. Generation of a Digital Shadow

Data quality partly considered only. It can be observed that most approaches consider input data to be *reliable*, *frequently available* as well as *complete*. For example, (Kobayashi et al., 2017) assume a correct order of occurrences in the log. This, however, stands in conflict with the data characteristics of our domain. Nevertheless, some approaches include parametric mechanisms to compensate data incompleteness, unreliability, and sparseness to some extent during a pre-processing step as supported by (Graf et al., 2022a; Messager et al., 2019).

Scalability mostly unclear. Only about half of the approaches report coherently about the data volume they operate on (Antonello et al., 2021; Graf et al., 2022; Hallac et al., 2017; Harper and Tee, 2019; Kobayashi et al., 2017; Messager et al., 2019). However, most of them do not consider the volume

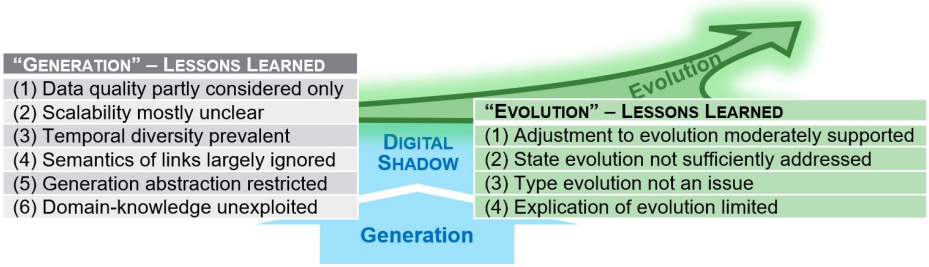

Figure 4. Lessons Learned and Open Issues

of input data in relation to occurrences of objects and inter-dependencies therein as well as possible sparsity and uneven distribution. The reason could be that those, operating in a batch-wise fashion (Antonello et al., 2021; Harper and Tee, 2019; Kobayashi et al., 2017; Messager et al., 2019; Psorakis et al., 2012), are not intended specifically for (near) real-time employment which would be, however, desirable for maintaining a "live" digital shadow. Overall, the work of (Kobayashi et al., 2017; Messager et al., 2019) seems to be *most promising* since they explicitly focus on large-scale logs.

Temporal diversity prevalent. It occurs that the investigated approaches broadly fall into one of two categories: those, who consider *time of occurrence* (Kobayashi et al., 2017; Song et al., 2022) and those who draw from *partial order of occurrences* as (Messager et al., 2019). In either case, time passed is of minor importance in those approaches. Whereas time of occurrence is a determining factor also in our domain at hand, some objects may have inter-dependencies only observable with delay or in the light of long-lasting events. This is aggravated by heterogeneity that may lead to divergence between the occurrence of the event and when it is reported, leading to a mix of time instance information alongside time duration information. Consequently, approaches are needed to be able to identify links regardless of temporal diversities, whereby (Graf et al., 2022; Messager et al., 2019) take a first step in this direction by not relying on exact timestamps.

Semantics of links largely ignored. While all approaches identify inter-dependencies in terms of links between objects, most are restricted to the *mere existence of a link* as well as *between two objects only* - with the exception of (Harper and Tee, 2019) since focusing on cascading failures. While some approaches complement links with meta-information, deriving, e.g., the link's confidence (Messager et al., 2019), they generally *lack any further link qualification*. Apart of two approaches (Kobayashi et al., 2017; Song et al., 2022), *link direction is not an issue*, leaving the important aspect of cause-impact aside. Most crucial, however, is, that *only one kind of link between objects* is identified. This limits the generated digital shadow to express the semantics of contextual links and thus to distinguish between interactions in normal operational scenarios from those in failure cases. Consequently, approaches are needed to express the existence of links between objects together with their specific semantics in a differentiated way like (Graf et al., 2022).

Generation abstraction restricted. Alongside the previous limitation is that only a *minority* of approaches provide *abstraction capabilities* in the course of generating the digital shadow (Graf et al., 2022; Harper and Tee, 2019; Kobayashi et al., 2017). These are, however, urgently needed for our purposes, not least since commonly log files stemming from OT contain data not being tailored to serve as input for generating the structural aspects of a digital shadow. Thus, abstraction mechanisms should mediate between input data and the digital shadow, including *duplicate removal* and *mappings of low-level messages* to more *abstract events*.

Domain-knowledge unexploited. Most of the approaches are *agnostic* wrt. the types of objects they are employed on, to identify links in-between. While this *facilitates general applicability*, one fails in taking advantage of existing domain knowledge, in terms of considering, e.g., messages recorded and the events they report or knowledge about inter-dependencies which are potentially possible. That could leverage, on the one hand side, abstraction in the course of digital shadow generation and, on the other hand side, help in expressing the digital shadow as semantically rich as desired. Only (Graf et al., 2022) uses *type-level domain knowledge* in order to enrich the underlying data-driven mechanisms during all abstraction levels of the generation process, an idea which can be easily mapped onto the generic reference model of (Rivera et al., 2020), also foreseeing the usage of domain knowledge.

4.2. Evolution of a Digital Shadow

Adjustment to evolution moderately supported. Several approaches (Graf et al., 2022; Harper and Tee, 2019; Kobayashi et al., 2017; Lugaresi and Matta, 2021b; Messager et al., 2019; Rivera et al., 2020; Song et al., 2022) are robust to *adjust to some limited form of evolution* meaning that they are inherently able to regard new objects and in so learn new links in-between. However, since biased towards identifying links based on the presence of objects, in case of objects not being encountered, they disregard proper adjustment by removing the corresponding object in the digital shadow and consequently also lack proper removal of corresponding links (Hallac et al., 2017; Harper and Tee, 2019; Psorakis et al., 2012; Song et al., 2022). Inherently, there is also *neither identification of replacement* of objects nor of links.

State evolution not sufficiently addressed. Although all approaches focus on generation of a structural model of the digital shadow, most of them *do not provide means to reflect an objects' current state* and *its evolution* within the digital shadow. The only exceptions are those explicitly providing digital twin functionality (Lugaresi and Matta, 2021b; Rivera et al., 2020).

Type evolution not an issue. As stated, not all approaches utilize additional knowledge about types of objects. Of those approaches who do (Harper and Tee, 2019; Rivera et al., 2020), *type information is assumed to be available and stable from the beginning*. In our problem domain, however, evolution of domain knowledge is subject to change over time since new types of OT are permanently incorporated as well as new usage requirements might lead to adjusting existing ones. In either case, this rises alongside the question of how to deal with different evolving versions of types in course of the generation of a digital shadow.

Explication of evolution limited. As already mentioned, while some approaches are robust to adjust to some limited form of evolution, only two approaches go beyond by being capable to explicate evolution (Hallac et al., 2017; Psorakis et al., 2012). These two approaches are able to identify that a change takes place, *when it happened*, *what changed* and in case of (Psorakis et al., 2012) even at *what scope*. A feasible and for our purposes quite useful further extension of this work would be to incorporate an explication of change points derived from a comparison of previous versions, thereby allowing an appropriate co-evolution of the digital shadow even in a predictive way based on trends of previous changes.

5. Summary and Conclusion

In this paper, we investigated on digital shadows for OTM in road traffic management by providing an *evaluation of data-driven approaches* based on criteria derived from existing surveys and challenges of our problem domain. Summarizing, the following conclusions can be drawn: Firstly, although the investigated approaches are capable to operate on log data, the impacts of *data quality issues*, *scalability*, and *temporal diversity* are *not comprehensively understood*. Secondly, while basic inter-dependencies are identified from log data, *differentiation to express adequate semantics is not an issue*. Thirdly, though data-driven approaches benefit from *general applicability*, they are *lacking to exploit available domain-knowledge* for digital shadow generation. Fourthly, while some approaches are *robust wrt. evolution* allowing for an adjustment of the digital shadow over time, *explication of evolution is lacking* - needed, however, to put forward a digital shadow in the light of the dynamic nature of OT in road traffic management.

References

Antonello, F., Baraldi, P., Shokry, A., Zio, E., Gentile, U., & Serio, L. (2021). A novel association rule mining method for the identification of rare functional dependencies in complex technical infrastructures from alarm data. *Expert Systems with Applications*, *170*.

Baumgartner, N., Gottesheim, W., Mitsch, S., Retschitzegger, W., & Schwinger, W. (2010). Beaware!—situation awareness, the ontology-driven way. *Data & Knowledge Engineering*, *69*(11), 1181–1193.

Baumgartner, N., Mitsch, S., Mueller, A., Retschitzegger, W., Salfinger, A., & Schwinger, W. (2014). A tour of beaware–a situation awareness framework for control centers. *Information Fusion*, *20*, 155–173.

Bordeleau, F., Combemale, B., Eramo, R., Brand, M. V. D., & Wimmer, M. (2020). Towards model-driven digital twin engineering: Current opportunities and future challenges. *Proc. of the 1st Int. Conf. on Systems Modelling and Management*, 43–54.

Boyes, H., & Watson, T. (2022). Digital twins: An analysis framework and open issues. *Computers in Industry*, *143*.

Brecher, C., Dalibor, M., Rumpe, B., Schilling, K., & Wortmann, A. (2021). An ecosystem for digital