Towards Message-Driven Ontology Population - Facing Challenges in Real-World IoT*

David Graf[1], Wieland Schwinger[1], Elisabeth Kapsammer[1], Werner Retschitzegger[1], Birgit Pröll[1], and Norbert Baumgartner[2]

[1] Johannes Kepler University, Linz, Austria
`firstname.name@cis.jku.at`
[2] team GmbH, `firstname.name@te-am.net`

Abstract. Large-scale Internet-of-Things (IoT) environments as being found in critical infrastructures such as *Intelligent Transportation Systems (ITS)* are characterized by (i) massive *heterogeneity* of data, (ii) prevalent *legacy systems*, and (iii) continuous *evolution* of operational technology. In such environments, the realization of crosscutting services demands a *conceptual IoT representation*, most promising, in terms of a domain ontology. Populating the ontology's *A-Box*, however, faces some challenges, which are not sufficiently addressed by now. In this respect, the contribution of this short paper is three-fold: Firstly, in order to point out the complexity of addressed real-world IoT environments, we identify prevalent *challenges for (semi-)automatic ontology population by means of a real world example.* Secondly, in order to address these challenges, we elaborate on related work by *identifying promising lines of research* relevant for ontology population. Thirdly, based thereupon, we *sketch out a solution approach* towards message-driven ontology population.

Keywords: Internet-of-Things, (Semi-)Automatic Ontology Population, Intelligent Transportation Systems

1 Introduction

Conceptual IoT Representation. Large-scale Internet-of-Things (IoT) environments as being found in critical infrastructures such as *Intelligent Transportation Systems (ITS)* are characterized by massive *heterogeneity* of data provided by the variety of individual systems and data sources involved. Prevalent *legacy systems* often lacking structured and semantic information. Moreover, systems themselves are subject to continuous *evolution*, meaning that the underlying Operational Technology (OT), employed for monitoring and controlling the ITS' operation, is continuously added, removed or replaced. In such environments, to realize crosscutting services like service quality monitoring of OT [8] or failure reasoning, a *conceptual IoT representation*, being independent from the actual technology used, is an indispensable prerequisite.

* This work is supported by the Austrian Research Promotion Agency (FFG) under grant FFG Forschungspartnerschaften 874490.

(Semi-)Automatic Ontology Population. A promising paradigm to address *heterogeneity*, *evolution* and *legacy systems* are *semantic technologies* in terms of *ontologies* [7]. While the *T-Box*, or the OT types respectively, are manually specified by domain experts in terms of an *object type catalog* providing simple taxonomic relationships, it is simply not feasible from a practical point of view to manually populate a domain ontology's *A-Box* with thousands of objects, particularly in the light of continuous evolution of OT, thus (semi-)automatic ontology population is a must. Moreover, ontology population is aggravated by the fact that the majority of historically grown systems are lacking homogeneous object information in a machine interpretable format. The only machine-readable data source available about underlying OT, which can be used as a basis for population, is often their communication data recorded within various message logs of systems consisting of human interpretable service messages as well as failure messages.

Paper Contribution. Aiming a *conceptual representation of the underlying OT* in terms of a domain ontology, i.e., OT objects and their relationships in between, the contribution of this short paper is three-fold: Firstly, in order to point out the complexity of addressed real-world IoT environments, we identify prevalent *challenges for (semi-)automatic ontology population based on message logs by means of a real world example* in the ITS domain in Section 2. Secondly, in order to address these challenges, we elaborate on related work by *identifying promising lines of research* relevant for ontology population through message logs in Section 3. Thirdly, based thereupon, we *sketch out a solution approach* towards message-driven ontology population in Section 4.

2 IoT Ontology Population Challenges by Example

In order to provide an illustrative picture of prevalent challenges when using messages to populate a domain ontology, we discuss these challenges based on a concrete example in the ITS domain. The IoT environment considered by our work comprises more than 100.000 OT devices of more than 200 different types[3] ranging from simple sensing and actuating devices (e.g., a lightning device) to complex systems consisting of several devices of various types (e.g., a radar systems), which are geographically distributed over 2.220 kilometers highway and 165 tunnels. During operation, these OT devices provide status information for the operators in terms of *a stream of messages* consisting of (i) human interpretable *message text*, (ii) a *unique identification* of the affected device, and (iii) *time* information, recorded within the logs of various systems.

Using these messages as a basis to populate a domain ontology, however, faces some challenges discussed in the following and exemplified by reflecting on the OT of an *emergency call system* being a crucial part for safety in tunnels on a highway. From an OT perspective, such an *emergency call system* consists of (i) two *door-contacts* (trigger a message when a door was opened), (ii) a

[3] Based on the domain specific *object type catalog* defined by the operating company.

SOS button to push in case of an emergency (triggers a message when pushed), and (iii) a *SOS phone* to make an emergency call (triggers a message when the phone is off-hook as well as when it is on-hook again), thus representing a composition in the T-Box of an ontology. These OT devices, as well as the *emergency call system* itself, are able to individually and independently send messages via various gateways to an operational monitoring and control system logging these service messages (e.g., the *door-contact* notifies ”door opened”) or failures messages (e.g., the *emergency call system* notifies a ”power supply error”). Regarding the A-Box of the ontology, naturally, there exist hundreds of *emergency call systems* (i.e., several of them in each tunnel) as naturally *door-contacts*, *SOS buttons* and *SOS phones* thereof. Whereby all of them finally need to be represented as the corresponding interrelated OT objects of their respective OT types belonging to the corresponding OT object of the composed OT type. This composition information is, however, not available and needs to be learned from messages. In the following, we discuss major challenges surrounding (1) *type-instantiation* and (2) *object-linking*, visualized in Figure 1.

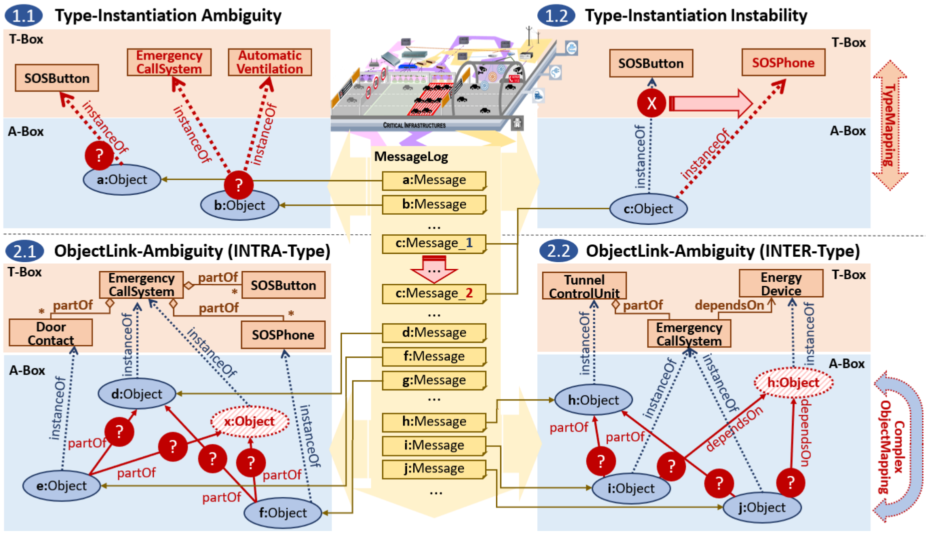

Fig. 1. Challenges of message-driven ontology population

(1.1) Type-Instantiation Ambiguity. Whereas the unique identification of the OT object is given through the affected device, their respective OT type information is not explicitly part of the message but rather just implicitly indicated through the message text, and to make matters worse, identical message text might be used by different OT types, which makes the type mapping of an OT object to the correct OT type ambiguous (cf. Fig. 1). For instance, the message ”SOS button pushed” (cf. ”a:Message”) indicates the OT type *SOSButton*, whereas, an example of a more generic message found like ”maintenance active”

(cf. "b:Message") might originate from an *EmergencyCallSystem* or from other OT like an *AutomaticVentilation* unit.

(1.2) Type-Instantiation Instability. As new messages related to the same OT object come in eventually, thus more information about an OT object is collected over time, their respective mapping to OT types might need to be adjusted (cf. Fig. 1). For instance, a message "SOS - main error" (cf. "c:Message_1") originating from an OT object being an initial indication of an OT type *SOSButton*. However, a later message "call active" (cf. "c:Message_2") from the same OT object makes it necessary to revise the OT type mapping since this message indicates that the very same OT object is rather an instance of a *SOSPhone*. Hence, the type mapping might change incrementally after each message requiring to consider all previous messages not the most recent one, only.

(2.1) INTRA-Type Object Link-Ambiguity. As there is a vast number of OT objects and due to messages lacking information of how these OT objects interrelate to each other, complex object mapping is challenging (cf. Fig. 1). For instance, since multiple objects of the OT types *DoorContact*, *SOSButton*, *SOSPhone* and *EmergencyCallSystem* exist, it is unclear which of those belong to the same physical *EmergencyCallSystem* unit.

(2.2) INTER-Type Object Link-Ambiguity. While the different semantics (e.g., *partOf, dependsOn*) of relationships between OT objects may be captured in the T-Box of the ontology (i.e., at type level), they are missing at object level and thus are not readily available for the A-Box (cf. Fig. 1). For instance, the OT object of OT type *EmergencyCallSystem* is *energySupplied* by an OT object of OT type *EnergyDevice*, however, it is unclear on which particular *EnergyDevice* object the particular *EmergencyCallSystem* instance depends on.

In addition, both cases of object link-ambiguity are aggravated by the fact that due to operating on a stream of messages we do not have a complete picture of all existing OT devices at a certain point in time (depicted by the dashed objects in Fig. 1). For instance, if an *EmergencyCallSystem* is not used and still works accurate, thus does not send messages, we lack the information at this point in time that this *EmergencyCallSystem* even exist.

3 Identifying Promising Lines of Research

Addressing the discussed challenges, our systematic literature review follows a goal-oriented strategy, meaning that we first aim identifying promising lines of research regarding our goal (i.e., populating a domain ontology) such as being found in the areas of *ontology population* and *semantic annotation* from text or semi-structured data, surveyed most recently by [14], before considering research areas beyond. Related work is compared primarily based on (i) the *source* data structure, (ii) the *techniques* applied, (iii) the *target* data structure, and (iv) the ability to address our *challenges* (cf. Section 2). A summary of related work along these comparison dimensions is given in Table 1.

Ontology Population and Semantic Annotation. Promising data-driven techniques such as clustering and semi-supervised classification are used by [10] and

Table 1. Related approaches

			Ganino et al. [6]	Lin et al. [12]	Liu et al. [13]	Jayawardana et al. [10]	Reyes-Ortiz et al. [17]	Belkaroui et al. [3]	Matzner a. Scholta [15][a]	Ni et al. [16]	Jin et al. [11]	Endler et al. [5]	Jafari et al. [9]	Detro et al. [4][b]
Source	Structure	semi-structured		✓				✓	✓	✓	✓		✓	✓
		unstructured	✓		✓	✓	✓							
	Dynamicity	static	✓		✓	✓	✓	✓	✓	✓	✓		✓	✓
		dynamic		✓								✓		
Techniques	Information Retrieval	text-based			✓	✓	✓	✓						
	Machine Learning	supervised			✓		✓							
		semi-supervised				✓								
		unsupervised		✓					✓	✓	✓			
	Data Analysis	distance-based				✓			✓					
		similarity-based								✓	✓			
	Other Methods	tool-usage (e.g, GATE)	✓											✓
		complex-event-processing										✓		
		specific data-mining											✓	
Target	Content-Ontology	wine events						✓						
		health events												✓
		law events				✓								
		IoT data (e.g., sensor values)		✓	✓							✓		
	Resource-Ontology	OT												
		roles												✓
		web services					✓							
	Other Formalisms	organizational model							✓	✓	✓			
		document	✓											
		RBAC model											✓	
Challenges	Type-Instantiation	ambiguity and instability	✓	✓	✓	✓	✓	✓	✓					
	Obj. Link-Ambiguity	INTRA- and INTER-type		✓	✓			✓	✓	✓	✓	✓	✓	✓

[a] survey, [b] preliminary work and not yet validated

[17] to populate a domain ontology, the latter being closely related regarding the target data structure by populating an ontology with resources in terms of web services. Both, however, use primarily text documents as data source and do not use semi-structured stream data originating from logs. Closely related is the approach of [3] populating an event ontology for monitoring vineyards grounded on an heterogeneous IoT sensor network aiming to mine causality relationships between events occurred during the life cycle of a wine production. Although the data originates also from IoT sensors, the target ontology primarily focus on events and not on a representation of the underlying IoT environment. Related with respect to techniques used are approaches in the area of semantic annotation, i.e., the process to annotate entities in a given text with semantics [6] (e.g., using ontology classes), such as the work of [12][13], which, however, address primarily the input data source themselves (often in terms of text documents) as target data structure.

Compared to our work, approaches discussed so far mainly differ regarding the source data structure, i.e., they do not consider streaming log data. Hence, in order to identify promising lines of research also with respect to the source data structure, in the following, we focus on related work having similar requirements to ours in that respect namely (i) using semi-structured log data as data source, (ii) extracting knowledge about hidden resources (e.g., people, systems, roles) and their relationships in between, and (iii) being confronted with stream data.

All these requirements are also tackled by work in the area of process mining [18], more precisely in one of its sub-field *organizational process mining* [1].

Organizational Process Mining. While the work of [2] provides a systematic review of automated process discovery methods, most promising mining approaches are reviewed by [15] aiming to "derive the underlying organizational structure of a CPS" from event logs. Discussed techniques such as "metrics based on (possible) causality" focusing on temporal succession of activities, or "metrics based on joint cases" focusing on frequency and correlation of resources, seems to be eminently suitable to derive relationships between objects also in the IoT domain. Furthermore, variations of distance measures, e.g, those used by [16], as well as traditional clustering techniques applied to event logs, e.g., those used by [11], are promising approaches to transfer to the IoT domain. In addition, since "time is a key relation between pieces of information" [5], time-based approaches are highly relevant for our work such as the organizational mining approach of [9]. With respect to the source data structure and the techniques used, closely related is the approach of [4] in terms of semantically annotating event log information in the health-care domain.

As Table 1 shows, although approaches discussed so far are related to our work in some of the comparison dimensions, none of them are directly applicable to our requirements since none aim a conceptual representation of OT as target data structure, especially not based on message streams as data source structure.

4 Envisioned Approach

Based on previously discussed challenges and identified promising lines of research, we finally sketch out a solution approach towards message-driven ontology population aiming a conceptual representation of OT, in the following.

Addressing the *type-instantiation challenges*, we envision to employ text-similarity-based techniques like used by [10][17] in terms of mapping message texts originating from a particular OT object to the most similar OT type, textually described by domain specific documents and technical specifications. Thereby a crucial aspect is to incrementally verify already existing type mappings since we have to be aware of new information provided by the message stream.

Addressing the *object link-ambiguity challenges*, we envision the usage of temporal patterns like [16] by applying distance measures or clustering techniques [11] with respect to certain timestamps of messages to identify relationships between OT objects. The rational behind is based on two hypotheses, namely (1) logical relationships between OT objects result in nearly simultaneous messages in case of a cross-device function failure, e.g., a shared energy-supply, and (2) physical relationships between OT objects result in typical functional temporal patterns of messages, e.g., in case of an *emergency call system* within 10 minutes typically the messages "door opened", "SOS button pushed", and "emergency call active" before again "door opened" are triggered.

First experiments towards this envisioned approach using samples of data have already shown promising results. For this in a first step we are now investigating in more detail on available real-world data expecting to elaborating on a prototype as the following step.

References

1. Appice, A.: Towards Mining the Organizational Structure of a Dynamic Event Scenario. J. of Intelligent Information Systems 50(1), 165–193 (2018)
2. Augusto, A. et al.: Automated Discovery of Process Models from Event Logs: Review and Benchmark. IEEE Transactions on Knowledge and Data Engineering 31(4), 686–705 (2018)
3. Belkaroui, R. et al.: Towards Events Ontology Based on Data Sensors Network for Viticulture Domain. In: Proc. Int. Conf. on the IoT, pp. 1–7. ACM (2018)
4. Detro, S. et al.: Enhancing Semantic Interoperability in Healthcare Using Semantic Process Mining. In: Proc. Int. Conf. on Information Society and Technology, pp. 80–85 (2016)
5. Endler, M. et al.: Towards Stream-based Reasoning and Machine Learning for IoT Applications. In: Intelligent System Conf., pp. 202–209. IEEE (2017)
6. Ganino, G. et al.,: Ontology Population for Open-Source Intelligence: A GATE-based Solution. Software Practice and Experience 48(12), 2302–2330 (2018)
7. Graf, D., Kapsammer E., Schwinger W., Retschitzegger W., Baumgartner N.: Cutting a Path Through the IoT Ontology Jungle - a Meta Survey. In: Int. Conf. on Internet of Things and Intelligence Systems. IEEE (2019)
8. Graf, D., Retschitzegger W., Schwinger W., Kapsammer E., Baumgartner N., Pröll B.: Towards Operational Technology Monitoring in Intelligent Transportation Systems. In: Int. Conf. on Management of Digital Eco-Systems. ACM (2019)
9. Jafari, M. et al.: Role mining in access history logs. J. of Computer Information Systems and Industrial Management Applications 1 (2009)
10. Jayawardana, V. et al.: Semi-Supervised Instance Population of an Ontology using Word Vector Embeddings. In: Proc. Int. Conf. on Advances in ICT for Emerging Regions, pp. 217–223. IEEE (2017)
11. Jin, T. et al.: Organizational Modeling from Event Logs. In: Proc. Int. Conf. on Grid and Cooperative Computing, pp. 670–675. IEEE (2007)
12. Lin, S. et al.: Dynamic Data Driven-based Automatic Clustering and Semantic Annotation for Internet of Things Sensor Data. Sensors and Materials 31(6), 1789–1801 (2019)
13. Liu, F. et al.: Device-Oriented Automatic Semantic Annotation in IoT. J. of Sensors 2017, 9589,064:1–9589,064:14 (2017)
14. Lubani, M. et al.: Ontology Population: Approaches and Design Aspects. J. of Information Science 45(4), 502–515 (2019)
15. Matzner, M. and Scholta, H.: Process Mining Approaches to Detect Organizational Properties in CPS. In: European Conf. on Information Systems (2014)
16. Ni, Z. et al.: Mining Organizational Structure from Workflow Logs. In: Proc. Int. Conf. on e-Education, Entertainment a. e-Management, pp. 222–225. IEEE (2011)
17. Reyes-Ortiz, J. et al.: Web Services Ontology Population through Text Classification. In: Proc. Fed. Conf. on Computer Science and Information Systems, pp. 491–495. IEEE (2016)
18. Van Der Aalst, W. et al.: Process Mining Manifesto. In: Proc. Int. Conf. on Business Process Management, pp. 169–194. Springer (2011)