Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 137 (2018) 235–241

Procedia Computer Science

www.elsevier.com/locate/procedia

SEMANTiCS 2018 – 14th International Conference on Semantic Systems

LSane: Collaborative Validation and Enrichment of Heterogeneous Observation Streams

Matthias T. Frank[a], Sebastian Bader[b], Viliam Simko[a], Stefan Zander[c]

[a]FZI Research Center for Information Technology, Haid-und-Neu-Str. 10-14, 76131 Karlsruhe, Germany
[b]Fraunhofer Institute for Intelligent Analysis and Information Systems IAIS, Schloss Birlinghoven, 53757 Sankt Augustin, Germany
[c]Hochschule Darmstadt, Haardtring 100, 64295 Darmstadt, Germany

Abstract

The increasing amount of publicly available data streams of environmental observation stations opens up new opportunities: domain experts are provided with an extensive amount of observations covering large areas with high density of environmental sensors, which could hardly ever be provided by a single organization. However, these opportunities come at the cost of new challenges regarding trustworthiness and comparability of such observations. In this paper, we address the challenges of semantic validation and enrichment of heterogeneous observation streams by exploiting collaboratively created and curated annotations. For this purpose, we introduce and discuss the Linked Stream Annotation Engine (LSane) to validate observation messages from heterogeneous sensors. We enrich these observation messages with provenance information derived from annotations. We present an implementation of LSane with messages from public and private environmental observation stations, which are mapped to explicit semantics, and validate and enrich the mapped messages based on annotations from the LSane collaboration platform.

Keywords: Semantic Sensors; Semantic Stream Processing; Collaborative Shape Constraints; Collaborative Reasoning

1. Introduction

Publicly available data streams of environmental observation stations are continually growing in popularity and pervasiveness. Current examples are public observation stations for traffic noise[1] or air pollution[2] but also private observation stations like senseBoxes[3] or other weather stations, which offer their observations in machine-readable formats on the Web. The increasing availability of such data streams leads to new opportunities but also challenges: domain experts can exploit extensive observations that cover large areas with high density. However, trustworthiness of observations in publicly available data streams usually varies due to lacking provenance information, missing

[1] http://www4.lubw.baden-wuerttemberg.de/servlet/is/224275/
[2] http://mnz.lubw.baden-wuerttemberg.de/messwerte/aktuell/statDEBW080.htm
[3] https://sensebox.de/en/

10.1016/j.procs.2018.09.022

values or unreliable providers. In addition, incorporating values observed by heterogeneous sensors is a difficult task due to heterogeneous formats, undocumented syntax and ambiguous semantics of observation messages. Moreover, in practice groups of people usually work together to develop applications based on those streams. We therefore aim at enabling domain experts to take the opportunity of the high amount of environmental sensor data included in publicly available data streams by addressing the associated challenges and assisting them to properly annotate sensor data streams in a collaborative manner. Consequently, our work elaborates around the following research questions: 1) *How can heterogeneous messages of environmental observations be collaboratively validated with uniform shapes?* and 2) *How can we exploit these annotations to derive and apply rules for semantic enrichment to heterogeneous messages of environmental observations?* We propose the Linked Stream Annotation Engine (LSane) for annotating streams in teams through a Media Wiki extension and combine it with several state-of-the-art technologies of the semantic Web stack. An overview of the relevant LSane modules discussed in this paper is shown in Figure 1. We outline how groups of experts can work together to lift existing streams to RDF streams by using an intuitive configuration mechanism. Our implementation, which is based on a Kafka server, prototypically shows the suitability of our approach and has already proven its applicability in practical use cases.

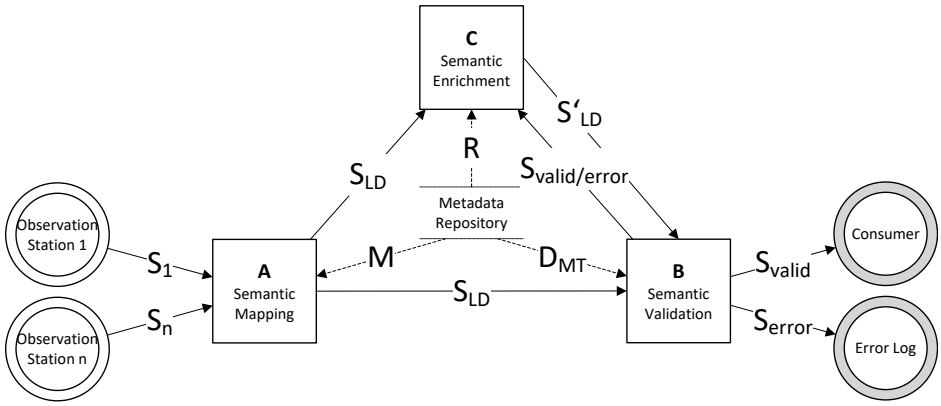

Fig. 1. Overview of the LSane modules: Module *A* [8] is used to map the key-value pairs of messages in streams S_{1-n} produced by observation stations $1-n$ to messages of triples with explicit semantics in stream S_{LD} based on metadata M provided by the central metadata repository. Module *B* provides a validation for the triples in messages of S_{LD} based on the message type definitions of D_{MT} of the metadata repository and produces messages with additional triples in streams S_{valid} or S_{error} depending on the result of the validation process. Module C is used to execute rules R defined in the metadata repository on streams S_{LD}, S_{valid} or S_{error} to infer new triples for S'_{LD}.

The remainder of this work is organized as follows: In Section 2, we provide and discuss the related literature. In Section 3 and 4, we detail the LSane approach wrt. validating and enriching observation messages based on collaboratively created annotations. A proof of concept for the LSane approach and a conclusion of the current state of our work are discussed in Section 5.

2. Related Work

In this section, we discuss related work in the fields of 1) heterogeneity and semantics of sensor streams, 2) semantic annotations for sensor streams and 3) semantic validation and enrichment of sensor streams as our work contributes to all three parts.

Heterogeneity and Semantics of Sensor Streams. The Semantic Sensor Network (SSN) ontology [3] defines basic concepts for observations and sensors, in particular applicable in the IoT domain. In the recent years, the SSN ontology became more and more the de facto standard vocabulary for describing sensors and sensing events. Together with the SAREF ontology [4] these two vocabularies are the most commonly used ones for the semantic description of sensing and actuation. Wiener, P. et al et al. [16] have shown that variety and veracity of spatio-temporal data of heterogeneous sensor observations are still an unsolved issue that has to be addressed in order to generate meaningful knowledge. They discuss an approach for continuous refinement of these data supported by semantic web services

and domain experts in their vision paper. However, additional research has to be carried out in order to proof this approach. Markovic et al. [14] have pointed out that streams of low-level observation messages from heterogeneous sensors are meaningless without higher-level context information that adds explicit semantics to the sensor data. With their work, the authors have shown that controlled semantic vocabularies like SSN or PROV-O can be exploited to explicitly model the provenance of sensor data. Further, the authors have shown how the vocabularies can be used for a semantic stream-based data processing framework [13]. Although they have deployed there approach to a relevant use case within the food safety domain, the authors have not stated how experts of this domain are enabled to exploit the expressiveness of the developed ontology or collaboratively annotate data streams of existing sensors. Mappings of relational or otherwise formatted data to Resource Description Framework (RDF) is possible with the RDB to RDF Mapping Language R2RML [5] or the broader applicable RDF Mapping Language RML [6] which also enables mappings from JSON, XML or CSV to RDF. The desired transformations are also formulated in RDF by defining the output graph structure by so-called Maps and URI templates. While R2RML strictly relies on tables and uses column names as resource and attribute identifiers of row-based data objects, RML also transforms JSON and XML data by identifying objects according to their keys. Even though some tools have been introduced in order to support the creation of mappings for both approaches, the possibility to collaboratively work on mappings was not part of the design requirements and is still missing. One example for a related vocabulary is the IoT-Lite ontology [15] that reuses the core SSN Device and Sensor definitions for a broader model of IoT resources with a strong focus on lightweight descriptions. Barnaghi et al. [2] provide a framework for stream annotations in combination with data from the Linked Open Data Cloud in order to improve the location attributes. They also provide a Web client to support the annotation but only on the level of stream elements and do not assist to create mappings for series of elements which typically is essential for any stream. The approach targets the manual annotation of stored streams but not real-time, automated transformation. The approach of Duy et al. [7] focuses—similar to the one presented in this paper—on sensor observations and describes those through the SSN and SWEET ontology. The streaming data is enriched with semantic concepts and provided by a SPARQL API on top of a graph store, linking observations and measurement stations by RDF predicates. As they have several modules in their framework that do not support a configurable mapping for the streams they implement the transformation in their program code. A cooperation of several experts is therefore not possible.

Annotations of Sensor Streams. Kolozali [11] developed a framework specifically suitable for IoT stream processing. Their Stream Annotation Ontology (SOA) – which is also implemented by IoT-Lite – defines basic stream concepts and allows an RDF modeling of streaming events. Additional concepts are also defined for quality of service, quality of information and provenance features. The proposed stream annotation framework supports the transformation of non-RDF to RDF streams and therefore enables the creation of unambiguously defined stream events through the mentioned ontologies. Nevertheless, the framework misses a straightforward configuration module for those mappings; neither does it support collaborative work on the regarded streams. The most widely known and most successful project for collaboratively combining knowledge and information currently is certainly Wikipedia. While the MediaWiki engine supports both access but also contributions from nearly any human user – with some restrictions regarding content quality – many automated bots are currently actively maintaining the Wikipedia sites. Nevertheless, its content is not natively machine-processable as e.g. no meaning is attached to links between Wiki sites. Semantic Wikipedia [12] closes this gap by extending MediaWiki and allows more specific, semantically defined annotations and relations by introducing RDF into MediaWiki. Amiguet-Vercher et al. [1] have discussed the challenges of creating, propagating and consuming semantic annotations of data streams of observation messages within sensor networks, in particular in cases where the intended semantics of data streams changes over time. They have deployed there approach on a network of environmental observation stations in the Alps, where e.g. snow on a sensor could cause a total change of the meaning of observed values. However, the authors do not state how these annotations can be maintained in a collaboratively way. Although the authors have described the annotation propagation on a local level for single processing elements that also covers significance of semantic annotations, further research has to provide additional insight for annotation propagation on a workflow level.

Validation and Transformation of Sensor Streams. The Shapes Constraint Language (SHACL) [10] introduces a W3C recommendation for validation mechanisms on RDF graphs. The definition of required attributes, cardinality of relations or datatype restrictions in the form of shapes is an important aspect to ensure data quality for any productive

system. Some tools are already created to assist the creation of SHACL shapes, e.g. a Protégé plugin and as a part of TopBraid Composer. As SHACL shapes are also defined in RDF, they share the same format as the validated data in contrast to e.g. plain SPARQL Rules. This eases the required technology stack and reduces the amount of used libraries.

3. Approach

Based on the related work introduced and discussed in Section 2, we propose the Linked Stream Annotation Engine (LSane) and demonstrate how collaboratively created annotations can be both effectively and efficiently employed to validate and enrich semantics of data streams of observation stations on-the-fly. The underlying rationale resembles a generic framework of loosely coupled and platform-independent components that communicate over message brokers or Web APIs. Our approach is built on the assumption that streams of heterogeneous observation messages which consist of plain key-value pairs are mapped to messages that consist of triples with explicit semantics, for example based on a metadata management platform as described in [8]. This section describes the basic concept of LSane and outlines how it complements the previous system presented in [8] by introducing collaborative validation and enrichment mechanisms of heterogeneous observation streams.

Validation. For the validation of observation streams, we extend the collaborative annotation platform of [8] with shape definitions of message types D_{MT} to distinguish different categories of observation messages. An overview of the semantic validation process is shown in Figure 2.

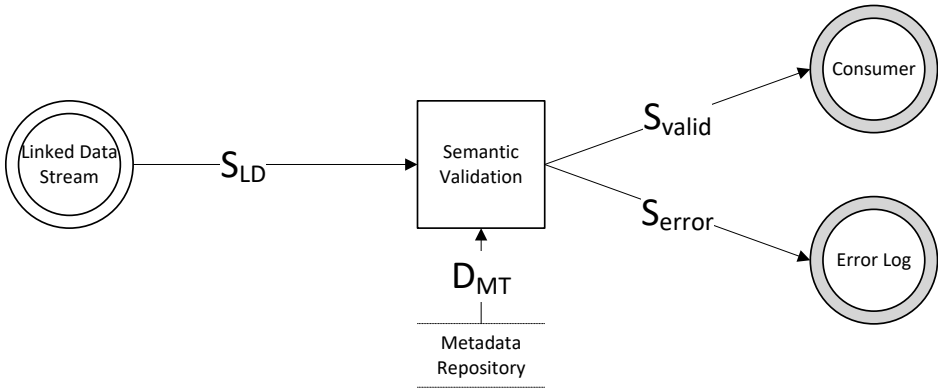

Fig. 2. Validation process: Messages with explicit semantics in S_{LD} are validated and enriched based on message type definitions D_{MT} of the collaborative semantic repository.

Having S_{LD} as the input stream of observation messages with triples that are mapped to explicit semantics, we validate each observation message O with the shapes for this stream using D_{MT} derived from the semantic repository. The semantic validation is defined as the split function $SemVal(O, D_{MT})$ that takes an observation message O and the message type definition D_{MT} as input:

$$SemVal(O, D_{MT}) = \begin{cases} O_{valid} = O \cup R_{valid} & \text{if } D_{MT} \subseteq O \\ O_{error} = O \cup R_{error} & \text{if } D_{MT} \supset O \end{cases}$$

If all triples of O are valid according to D_{MT}, the validation function returns the message O_{valid}, defined as the union of O and the set of triples of the validation result R_{valid}. If the validation of D_{MT} fails for at least one triple of O, the validation function returns the error message O_{error}, defined as the union of O and the set of triples of the validation result R_{error}. The output of the $SemVal$ function applied to data stream S_{LD} are data streams S_{valid} for consumers of validated observation messages and S_{error} which can be used to log validation errors which are forwarded to the stream broker.

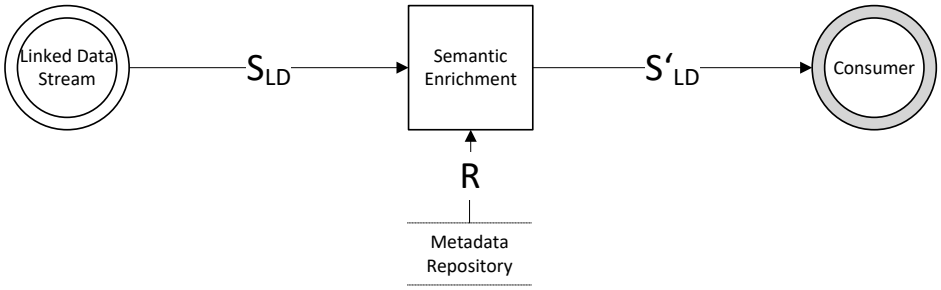

Fig. 3. Enrichment process: Messages with explicit semantics in S_{LD} are enriched based on rules R of the collaborative semantic repository and published to S'_{LD}.

Enrichment. Depending on the use case, it could be necessary to transform observed values to other representations in order to fulfill the requirements of D_{MT}. For example, if D_{MT} requires a thermodynamic temperature given as degree Fahrenheit and a sensor only delivers observations of thermodynamic temperature given as degree Celsius, $SemVal$ returns O_{error} although a simple transformation would fulfill the requirement. Therefore, we enrich observation messages O to O' by adding derived statements that make this implicit information explicit based on context information from the metadata repository. An overview of the semantic validation process is shown in Figure 3. The semantic enrichment function $SemEnr$ of LSane can use this information to derive a new statement in O' based on rule R and the original observed value, in our example the value in degree Celsius, contained in O as $SemEnr(O, R) = O'$. Using O' for the validation in $SemVal$ leads to the valid result O'_{valid}, regardless whether D_{MT} requires a thermodynamic temperature given as degree Fahrenheit or degree Celsius. Applying the $SemEnr$ function to data stream S_{LD} results in the new data stream S'_{LD} which can again be used as input for $SemVal$. As a consuming application is typically subscribed to S_{valid}, the previously created result O_{error} is ignored.

4. Implementation

In this section, we describe the implementation of the LSane approach as introduced in Section 3. The implementation comprises several steps: 1) Extending the annotation platform to define constraints for observation messages, 2) implementing a validation engine for observation messages, and 3) implementing semantic stream enrichment. An overview of relevant components used for the implementation is shown in Figure 4.

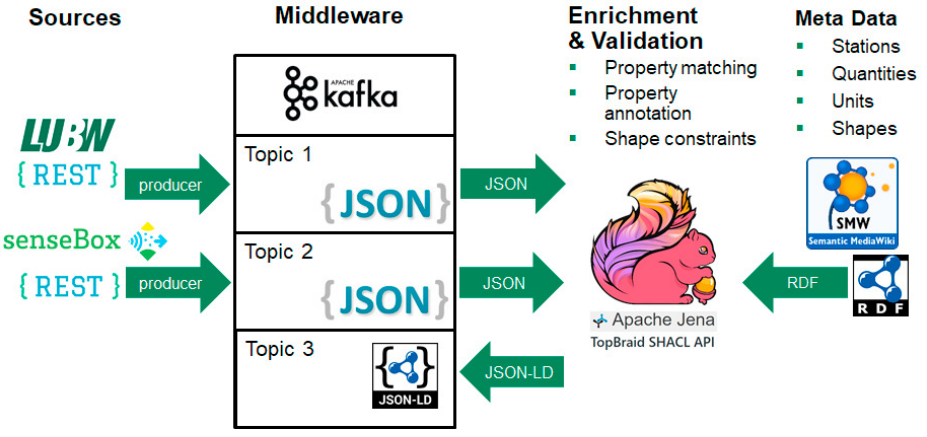

Fig. 4. Implementation: LSane relies on Apache Flink for stream processing and Apache Jena for RDF modeling. It receives JSON messages from the Apache Kafka message broker, maps them to RDF based on the metadata in Semantic MediaWiki and uses TopBraid's SHACL API for validation and enrichment.

Extension of the Annotation Platform. The underlying engine that provides the functionality for the LSane annotation platform is used to import shared vocabularies, add own terms, model entities and their relations and provide advanced functionality for querying and exporting semantic statements in a Wiki-based manner as well as reasoning and enrichment of semantic data. Depending on the requirements of the use case, Semantic MediaWiki (SMW) [12] or Linked Data Wiki [9] can be used interchangeable as the underlying engine for transforming input from MediaWiki templates to RDF statements in LSane. As an example, "temp" could be used as the key $k \in \mathbb{STR}$ to describe a member of an observation message serialized in JavaScript Object Notation (JSON). This key is explicitly mapped by a domain expert to the Uniform Resource Identifier (URI) that identifies the concept of thermodynamic temperature and degree Celsius is assigned as measurement unit for the values $v \in \mathbb{VAL}$. However, in order to define constraints for observation messages, a system of rules has to be included. For this reason, we employ the Shapes Constraint Language (SHACL) vocabulary within the annotation platform and enable domain experts to intuitively define general requirements for observation messages, which are independent from their provenance on a semantic abstraction layer. Domain experts can select required properties and cardinality constraints using forms of the annotation platform and apply these constraints to a set of observation streams. As an example, a domain expert could define a shape constraint for temperature observations. According to this shape, a temperature observation has to include exactly one member which is an instance of thermodynamic temperature and has a value $v \in \mathbb{VAL}$ that can be processed as floating-point number. Due to the employed SHACL vocabulary, the annotation platform delivers an RDF representation of this shape constraint in SHACL that can be interpreted by the validation engine.

Validation Engine. The validation engine of LSane employs the SHACL API provided by TopBraid. It holds an Apache Jena RDF model D_{MT} of the shape constraints from the annotation platform and validates each observation message O of the observation stream S_{LD} using that shape. As an example, the message type definition D_{MT} defined as *property quantity:ThermodynamicTemperature: minCount=1, maxCount=1* applied to an observation message O which does not include this property would cause the error message *Property needs to have at least 1 values, but found 0* for the *sh:minCount* component. $SemVal(O, D_{MT})$ uses that error message to create a new message in S_{error} as the message O used for this example does not contain a thermodynamic temperature and does therefore violate the minimum cardinality constraint of D_{MT}.

Semantic Stream Enrichment. The semantic stream enrichment of LSane works in the same way as the validation engine: rules that are collaboratively defined in the annotation platform are applied to observation messages using the SHACL API. However, SHACL distinguishes different kind of rules: whereas the validation engine considers only statements of D_{MT} which contains instances of sh:property, the semantic stream enrichment applies the set of rules R which contains instances of sh:rule. Instances of sh:rule could be further distinguished in rule types like sh:TripleRule or sh:SPARQLRule. For LSane, we employ rule type sh:SPARQLRule as it allows to encode SPARQL Protocol and RDF Query Language (SPARQL) construct queries without the need of further specification.

5. Demonstration and Conclusion

Proof of Concept. To demonstrate the LSane approach, we test the implementation introduced in Section 4 with message patterns of public and private environmental observation stations and a shape constraint definition for temperature observations. As data input for the demonstration, we have generated a number of 10.000 example observations that use the same message patterns as they are used by two concrete environmental observation stations. The reason for using generated observation messages rather than a live feed is that we can adjust the frequency of emitted messages for performance testing and create reproducible results while still ensuring the compatibility with concrete environmental observation stations. The shape of the observation message is characterized by a set of ten key-value pairs of observed values and metadata. Without the knowledge of a domain expert, the implicit semantics of observed values cannot be evaluated. Besides observed values, the observation message does also contain spatial and temporal information. The spatial information is given as latitude and longitude for the World Geodetic System 1984 and the temporal information as milliseconds since 1/1/1970 for both message patterns. In contrast to the message pattern of the public observation station, the message pattern of the private station contains only one observed value together with an ISO 8601 timestamp and some addition metadata. The shape constraint definition has therefore to define two

requirements: an observation message needs to include exactly one floating-point number that represents the value of an observed thermodynamic temperature and exactly one timestamp that states the time when the value was observed.

Conclusion. In this paper, we have introduced the LSane approach for collaborative definitions of semantic shapes and enrichment rules for heterogeneous message streams. We have discussed related work in the fields of heterogeneity and semantics of sensor streams, semantic annotations for sensor streams and semantic validation and enrichment of sensor streams and identified the research gap for collaborative definitions of shapes and rules for observation messages. Based on these findings, we have provided a formal description of the semantic validation function $SemVal(O, D_{MT})$ and the semantic enrichment function $SemEnr(O, R)$ which both exploits collaboratively created semantic annotations of domain experts. For the implementation of LSane, we have extended an existing annotation platform to define constraints for observation messages using SHACL, implemented a validation engine for observation messages and semantic stream enrichment based on the SHACL API. For a proof of concept of LSane, we have used the shapes of concrete public and private environmental observation stations to generate streams of observation messages and validate these streams with shape definitions from the annotation platform. With our work, we have shown that heterogeneous messages of environmental observations can be collaboratively validated using semantic annotations of SHACL shapes and also that collaboratively created annotations of rules can be exploited for semantic enrichment of heterogeneous messages of environmental observations. Further research has to be carried out to investigate the applicability and scalability of LSane on a large amount of observation streams and the usability for domain experts in different domains.

References

[1] Amiguet-Vercher, J., Wombacher, A., Klifman, T.E., 2010. Annotations: dynamic semantics in stream processing, in: Nica, A., Varde, A.S. (Eds.), PIKM 2010, Toronto, Ontario, Canada, ACM. pp. 1–8. doi:10.1145/1871902.1871904.

[2] Barnaghi, P.M., Wang, W., Dong, L., Wang, C., 2013. A Linked-Data Model for Semantic Sensor Streams, in: IEEE iThings/GreenCom/CPSCom 2013, Beijing, China, IEEE. pp. 468–475. doi:10.1109/GreenCom-iThings-CPSCom.2013.95.

[3] Compton, M. et al, 2012. The SSN ontology of the W3C semantic sensor network incubator group. J. Web Sem. 17, 25–32. doi:10.1016/j.websem.2012.05.003.

[4] Daniele, L., den Hartog, F., Roes, J., 2015. Created in close interaction with the industry: the smart appliances reference (saref) ontology, in: International Workshop Formal Ontologies Meet Industries, Springer. pp. 100–112.

[5] Das, S., Sundara, S., Cyganiak, R., 2012. R2RML: RDB to RDF Mapping Language. URL: http://www.w3.org/TR/r2rml/. W3C Recommendation.

[6] Dimou, A., Vander Sande, M., Colpaert, P., Verborgh, R., Mannens, E., Van de Walle, R., 2014. RML: A Generic Language for Integrated RDF Mappings of Heterogeneous Data, in: 7th Workshop on Linked Data on the Web.

[7] Duy, T.K., Quirchmayr, G., Tjoa, A., Hanh, H.H., 2017. A semantic data model for the interpretion of environmental streaming data, in: ICIST, pp. 376–380. doi:10.1109/ICIST.2017.7926788.

[8] Frank, M.T., Simko, V., . Semantic Data Stream Mapping and Shape Constraint Validation Based on Collaboratively Created Annotations (in press), in: ICWE 2018.

[9] Frank, M.T., Zander, S., 2017. Exploiting Linked Open Data for Enhancing MediaWiki-based Semantic Organizational Knowledge Bases, in: SciTePress (Ed.), KEOD. doi:10.5220/0006587900980106.

[10] Knublauch, H., Kontokostas, D., 2017. Shapes Constraint Language (SHACL): W3C Recommendation 20 July 2017. URL: https://www.w3.org/TR/shacl/.

[11] Kolozali, S. et al, 2014. A Knowledge-Based Approach for Real-Time IoT Data Stream Annotation and Processing, in: iThings/GreenCom/CPSCom 2014, Taipei, Taiwan, IEEE Computer Society. doi:10.1109/iThings.2014.39.

[12] Krötzsch, M., Vrandecic, D., Völkel, M., Haller, H., Studer, R., 2007. Semantic Wikipedia. J. Web Sem. 5, 251–261. doi:10.1016/j.websem.2007.09.001.

[13] Markovic, M., Edwards, P., 2016. Semantic Stream Processing for IoT Devices in the Food Safety Domain, in: Martin, M., Cuquet, M., Folmer, E. (Eds.), SEMANTiCS 2016, Leipzig, Germany, CEUR-WS.org. URL: http://ceur-ws.org/Vol-1695/paper6.pdf.

[14] Markovic, M., Edwards, P., Kollingbaum, M.J., Rowe, A., 2016. Modelling Provenance of Sensor Data for Food Safety Compliance Checking, in: Mattoso, M., Glavic, B. (Eds.), IPAW 2016, McLean, VA, USA, Springer. doi:10.1007/978-3-319-40593-3_11.

[15] Marúdez-Edo, Elsaleh, T., Barnaghi, P.M., Taylor, K., 2017. IoT-Lite: a lightweight semantic model for the internet of things and its use with dynamic semantics. Personal and Ubiquitous Computing 21, 475–487. doi:10.1007/s00779-017-1010-8.

[16] Wiener, P. et al, 2016. BigGIS: A continuous refinement approach to master heterogeneity and uncertainty in spatio-temporal big data (vision paper), in: SIGSPATIAL 2016, Burlingame, California, USA, ACM. doi:10.1145/2996913.2996931.