Semantic Enrichment of Association Rules Discovered in Operational Building Data

Petrova, Ekaterina Aleksandrova; Pauwels, Pieter

Published in:
Proceedings of the 37th CIB W78 Information Technology for Construction conference

DOI (link to publication from Publisher):
10.46421/2706-6568.37.2020.paper022

Publication date:
2020

Document Version
Publisher's PDF, also known as Version of record

Petrova E. and Pauwels P. (2020). "Semantic enrichment of association rules discovered in operational building data" In: *Proc. 37th CIB W78 Information Technology for Construction Conference* (CIB W78), São Paulo, Brazil, pp. 308-326. DOI: http://dx.doi.org/10.46421/2706-6568.37.2020.paper022

SEMANTIC ENRICHMENT OF ASSOCIATION RULES DISCOVERED IN OPERATIONAL BUILDING DATA

Ekaterina Petrova[1] and Pieter Pauwels[2]

Abstract: The advancements in Building Information Modelling, Building Monitoring Systems and machine learning have made the discovery of hidden insights and performance patterns in operational building data possible and highly accurate. Semantic web technologies play a fundamental role in terms of knowledge representation and provide the necessary infrastructure for reuse of the discovered insights. Such knowledge can be of particular importance to decision-making for building performance improvement, however, this requires patterns discovered with traditional data mining techniques to be attributed with semantics, so that they can be machine-interpretable and reusable. Using linked data-based crowdsourcing techniques for interpretation of building performance patterns enables the creation of knowledge graphs of building data, enriched with contextualized building performance insights. This paper presents a crowdsourcing mechanism that allows the semantic enrichment of building performance patterns through semantic annotation and classification. We discuss the results and the potential of linked building data graphs enriched with building performance insights.

Keywords: Semantics, Association Rule Mining, semantic data mining, linked data, building performance, crowdsourcing.

1 INTRODUCTION

With the emergence and establishment of Building Information Modelling (Borrmann et al. 2018; Sacks et al. 2018) the Architecture, Engineering and Construction (AEC) industry underwent a paradigm shift in the creation and use of information. The exponential generation of data throughout the building lifecycle and the advances in analytical approaches have augmented that shift even more by giving AEC practices the ability to make use of and reuse data in a structured way. Moreover, being able to discover valuable insights in the data (e.g. building data, simulation data, IoT data, etc.) makes it possible to cater to high-level decision making related to sustainability, energy efficiency, indoor environmental quality, occupant comfort, etc. (Fan et al. 2018a). Advanced knowledge discovery methods aid the extraction of high-level knowledge from low-level data (Fayyad et al. 1996). Such knowledge allows higher level analyses and has the potential to redefine the way buildings are designed by serving as an evidence base in performance-oriented design decision-making (Petrova et al. 2019).

However, to be useful and have an impact on decision-making, the insights discovered in data need to be transformed into actionable knowledge, which includes analytical efforts that require a lot more than identifying an analytical goal and selecting appropriate data mining algorithms (Fayyad et al. 1996). Crucial to knowledge discovery

[1] Assistant Professor, Department of the Built Environment, Aalborg University, Aalborg, Denmark, eap@build.aau.dk

[2] Associate Professor, Department of the Built Environment, Eindhoven University of Technology, Eindhoven, The Netherlands, p.pauwels@tue.nl

are the interpretation, contextualization, and enabling the reuse of building performance insights. Meaning is not explicit in insights discovered by data mining algorithms. Therefore, it needs to be attributed through semantic classification and annotation by domain experts, who can assess the value and meaning of the discovered building performance insights (e.g. frequent patterns, anomalies, association and sequential rules, etc.). Furthermore, to close the holistic information management cycle and enable reuse, the discovered knowledge has to be machine-readable and implementable in knowledge-based (decision support) systems.

In this regard, a reconciliation of statistical and symbolic Artificial Intelligence (AI) can provide the necessary combination of approaches to facilitate the above-mentioned objectives. Statistical methods have proven to be useful for discovering patterns, regularities or irregularities in data and symbolic representations excel at capturing the knowledge within a given domain explicitly, thereby allowing various forms of inference (Hoehndorf and Queralt-Rosinach 2017). As part of the statistical realm, machine learning approaches for knowledge discovery allow the extraction of valuable insights from the large datasets generated throughout the entire building life cycle. Semantic data modelling, linked data and web technologies (Berners-Lee et al. 2001; Bizer et al. 2009), on the other hand, have made it possible to represent the built environment formally, retrieve knowledge according to domain-specific requirements and reason about building performance (El-Diraby 2013; Pauwels et al. 2017).

Due to their proven ability to support decision-making, both approaches have independently received major attention in AEC. In depth research has been performed to identify how to transform raw data into building performance insights and make use of the multiplicity of collected, but usually rarely reused data. Included here are efforts aiming to define the various building data types (Petrova et al. 2019), as well as corresponding machine learning methods for data pre-processing, mining, visualization and use of discovered knowledge (Fan et al. 2015; Fan et al. 2018a; Fan et al. 2018b; Miller et al. 2018; D'Oca et al. 2018; Fan et al. 2019). Research has also shown that publishing data effectively, breaking up information silos, integration of data across domains and making data readable and understandable by both machines and humans is equally important. The latter is showcased at length in state of the art research contributions related to implementation of semantic web and linked data technologies in AEC (Curry et al. 2013; Pauwels and Terkaj, 2016; Pauwels et al. 2017; Rasmussen et al. 2017; Rasmussen et al. 2019; McGlinn et al. 2019). Finally, recent research also highlights a paradigm shift in the knowledge discovery and data mining community, which entails moving from mining raw data to mining the formalized knowledge directly (semantic data mining) (Lausch et al. 2015). In other words, the combination of symbolic and statistical approaches can enrich data mining processes with domain knowledge (Ristoski and Paulheim 2016) and facilitate knowledge discovery, representation and reuse, which cannot be achieved with any of these approaches alone.

Therefore, the main objective of the current research effort is to enable semantic enrichment of building performance insights discovered in operational building data in a machine-readable and reusable way. We look into how semantic annotations can be retrieved from domain experts and how they can be classified and encoded together with the building data and performance insights discovered with traditional data mining approaches to form a knowledge base. This is demonstrated for a use case in Denmark, for which frequent performance patterns and association rules have been discovered in indoor environmental quality sensor data streams. We compare the results with an alternative approach employing semantic sensor stream processing and frequent graph

pattern recognition and we discuss the implications. Finally, indicate how the created knowledge base can serve as an input to providing recommendations for evidence-based decision-making in performance-oriented design.

The paper starts by outlining the background and motivation of the research (Section 1). Section 2 then outlines the methodological approaches adopted in the study. We then proceed by presenting the results from the knowledge discovery and semantic data modelling efforts that provide the input for the creation of the initial knowledge base for decision support in performance-oriented building design (Section 3). Section 4 details the linked data-based crowdsourcing effort aiming to capture the domain expert interpretations of the discovered building performance insights, as well as their semantic annotations and classifications. Section 5 presents initial semantic sensor stream processing and frequent graph pattern analysis results, thereby showcasing an alternative approach pertaining to the semantic data mining domain. Finally, Section 6 presents concluding remarks.

2 METHODOLOGY

Machine learning approaches for knowledge discovery allow retrieving frequent and infrequent patterns (motifs and discords respectively), anomalies and association rules in operational building data. Included in this context is also the direct mining of formalized knowledge through the use of novel semantic data mining methods. In this article, we rely on results from a previously proposed method for combination of knowledge discovery (motif discovery and Association Rule Mining (ARM)) (Agrawal et al. 1993) in operational building data and semantic data modelling for knowledge representation of a performance enriched semantic building graph (Petrova et al. 2018; Petrova et al. 2019). Association rules indicate to what extent certain events (patterns) are related to, or are potentially caused by other events (patterns). Such associations can provide valuable insights into the buildings' behaviour. Capturing this information semantically together with other meaningful building data allows applying information retrieval techniques and ultimately- implementation of the discovered knowledge in a decision support system (Petrova et al 2019).

To prepare for motif discovery and ARM, we first apply Symbolic Aggregate Approximation (SAX) (Lin et al. 2007) on the raw sensor data, which aims for dimensionality reduction and indexing with a lower bounding distance measure, i.e. the method allows reducing a large dataset to a smaller one, without losing the characteristics of the data. Motif discovery is then performed through identification of the Longest Repeated Substrings (LRS) within the SAX symbol sequences with a custom implementation of the Suffix Tree algorithm (Ukkonen 1995). Association rules between the identified frequent patterns are discovered through an implementation of the FP-growth algorithm, as both implementations are done by the help of the SPMF open-source data mining library. The output includes both the association rules, as well as their corresponding measures of interestingness, support and confidence, which show how frequently a rule appears throughout the dataset and how often it is found to be true (Agrawal et al. 1993). The association rules have thereafter been visualised for a better understanding of the correlational dependencies between the motifs and the sensor observations in which they have been discovered. Those visualisations serve as the main input to the semantic enrichment of the association rules, which is the main objective of this study.

As previously stated, to be useful, the discovered knowledge needs to be reusable, retrievable, machine-readable and integrated with other building data. Therefore, linked data techniques are used to represent the different datasets and discovered knowledge together. Home2020 was therefore modelled using the Linked Building Data (LBD) modelling principles and ontologies. More specifically, the building has been represented as a Resource Description Framework (RDF) graph by the use of the Building Topology Ontology (BOT) (Rasmussen et al. 2017). Furthermore, geographical location of the building is modelled through the use of geospatial ontologies, OpenStreetmap location and OpenWeatherMap, while sensor nodes and observations are added to the graph with the SOSA, SSN and OM (Units of Measure) ontologies. Data pertaining to heat consumption, domestic hot water use, use of appliances, HVAC system data, and HVAC design strategy for the building in accordance with the design brief requirements have also been added to provide the necessary context for the interpretation of the discovered performance patterns and rules. Occupant data has been modelled with the FOAF ontology. Finally, the discovered motifs and association rules are added to the semantic building graph by a custom "pattern" ontology (:ptn) specifically built for the purpose.

To be able to be interpreted and disambiguated, the discovered performance patterns have to be presented to domain experts in a way that allows contextualised knowledge to be continuously stored, retrieved, updated and reused. Therefore, we introduce a linked data-based crowdsourcing mechanism, which allows indoor environmental quality experts to contextualise the available building performance patterns and association rules by the use of semantic annotation tags and semantic classification.

Finally, as the above-described approach relies on traditional data mining techniques for performance pattern discovery and semantic modelling for representation of the results together with the available building data, we compare it to a direct semantic data mining approach using a frequent RDF graph pattern analysis method (Belghaouti et al. 2016).

3 Building Performance and Explicit Knowledge Bases

The way semantic graphs represent relations between buildings, locations, spaces, and other heterogeneous data enables the scaling and articulation of the discovered knowledge of how the existing building stock performs in a machine-readable form. Therefore, semantic graphs and ontologically demarcated data provide an infrastructure that allows knowledge disambiguation, contextualization and reuse through the rich, machine-readable semantic links between concepts. To enable building performance knowledge contextualization and demonstrate the value of semantics, the available building data needs to be treated in a way that allows capturing the evolution of the discovered knowledge over time. That includes the relation of the building performance insights to other relevant data in the AEC domain.

3.1 Knowledge Discovery and Semantic Data Modelling of Operational Building Data and Performance Insights for a Nearly Zero Energy Building

For this study, motifs and association rules have been discovered in indoor environmental quality sensor observations from a single family house located near the city of Aarhus, Denmark (Home2020), which was completed in 2017 and rated as nearly zero energy building (NZEB) according to the Danish energy labelling standard. The

collected data is from the period 01.12.2017 to 31.10.2018 and includes measurements of energy consumption for heating [MWh], ventilation system [kWh], control system [kWh], and kitchen appliances [kWh], as well as outdoor air temperature [◦C], return air temperature [◦C], return air relative humidity [%], hot water temperature [◦C], supply air temperature [◦C], ventilation speed [steps]. Both hot and cold water consumption [m3] are also monitored. This study focuses on the indoor environmental quality data, which includes temperature [◦C], CO2 [ppm], and relative humidity [%] observations for a bedroom, a living room and a kitchen. The measurement interval is five minutes.

As described in Section 2, the sensor observations are transformed into symbolic representations with SAX. That means that the sequence of all data points are replaced by a symbolic representation such as 3222222322223333..., with each SAX symbol representing an interval of data values (e.g. 2 = [22.86950723073572, 23.704365409749624] for the Temperature observations). As a result of the SAX transformation, the output dataset consists of sequences of symbolic representations per observed variable (Temperature, CO2, Relative Humidity) for each room per month. To enable motif discovery with the LRS algorithm, co-occurrence matrices are computed on the basis of the SAX representations to identify co-occurring SAX symbols on a monthly basis. The LRS algorithm then identifies the frequent repetitive patterns in the SAX symbol sequences (Fig. 1).

```
345555 - 3 - 13;103;130;
444333 - 5 - 78;167;196;504;559;
4445555 - 4 - 29;178;244;642;
44544 - 3 - 124;222;241;
455555556 - 3 - 14;246;598;
455556 - 4 - 31;131;180;644;
54433 - 4 - 62;224;363;432;
55544 - 6 - 107;160;191;217;361;636;
555666 - 10 - 133;147;182;251;309;382;603;621;646;690;
6555554 - 3 - 157;188;723;
66655 - 8 - 141;155;186;301;428;629;681;706;
6667666 - 3 - 137;297;386;
```

Figure 1: A set of LRS found in the SAX sequences of sensor observations (Petrova 2019)

Each motif is given an unique ID, which becomes the input for discovery of the association rules. Several hundreds of motifs and rules were discovered for each observed variable, room and month. Figure 2 presents a small excerpt of rules, as well as their constituting motifs and measures of interestingness. Essential here is the fact that not all performance patterns and rules will be interesting and present unknown novel and useful insights. Further contextualisation and interpretation are required to discover the rules with the highest level of novelty and value.

```
452 ==> 489 #SUP: 1 #CONF: 1.0
453 ==> 485 #SUP: 3 #CONF: 0.6
454 ==> 481 #SUP: 1 #CONF: 0.5
456 ==> 484 #SUP: 2 #CONF: 0.6666666666666666
457 ==> 488 #SUP: 1 #CONF: 1.0
459 ==> 481 #SUP: 1 #CONF: 0.5
459 ==> 488 #SUP: 1 #CONF: 0.5
482 ==> 460 #SUP: 1 #CONF: 0.5
460 ==> 482 #SUP: 1 #CONF: 0.5
460 ==> 485 #SUP: 1 #CONF: 0.5
457 488 ==> 378 #SUP: 1 #CONF: 1.0
378 488 ==> 457 #SUP: 1 #CONF: 0.5
```

Figure 2: An excerpt of the set of association rules obtained for the living room in Home2020 (Petrova 2019)

That may include considerations related to the combined effect of the support and confidence. Either way, it requires a domain expert to identify the strong and interesting rules that indicate novel building performance insights, thereby enriching them with semantics and transforming them from statistical output to actionable knowledge. The semantic enrichment of association rules requires a semantic data infrastructure that would allow the storage, retrieval, interpretation and reuse of the contextualised knowledge.

To allow the latter, all motifs and association rules have been modelled together with the available building data by the use of the PATTERN ontology, indicating their ptn:confidence, ptn:absoluteSupport, and ptn:relativeSupport measures. The modelled association rules (e.g. inst:associationRule_1) are linked to the sensor nodes they are related to with ptn:hasAssociationRule predicates. The constituting motifs for the association rules are represented as ordered lists of motifs for the left-hand side (ptn:LHS) and right-hand side (ptn:RHS) of each rule (Fig.3).

Figure 4 represents the resulting semantic building graph, which includes the available building data, system and occupant data, actuator and sensor data including data points for all observed variables, contextual data (geolocation and weather data), as well as the motifs and association rules discovered in the sensor data.

```
inst:associationRule_1
  rdf:type ptn:AssocationRule ;
  ptn:LHS (inst:Motif_45) ;
  ptn:RHS (inst:Motif_137) ;
  ptn:confidence "0.5"^^xsd:double ;
  ptn:absoluteSupport "1"^^xsd:double ;
  ptn:relativeSupport "0.5"^^xsd:double .

inst:motif_45
  rdf:type ptn:Motif ;
  ptn:SAXsequence "11122"^^xsd:string ;
  ptn:space inst:Kitchen ;
  ptn:month "8"^^xsd:string ;
  ptn:SAXsequenceFull (inst:SAXSymbol_91983cb8-4dd3-4544-a1fe-7
    b177e237bc0 inst:SAXSymbol_91983cb8-4dd3-4544-a1fe-7b177e237bc0
     inst:SAXSymbol_91983cb8-4dd3-4544-a1fe-7b177e237bc0 inst:
    SAXSymbol_41fadfdb-6560-4e96-9a7f-bc405f453452 inst:
    SAXSymbol_41fadfdb-6560-4e96-9a7f-bc405f453452 );
  ptn:observedVariable "CO2"^^xsd:string .

inst:SAXSymbol_36ef82d8-57c9-4e0a-a0bc-c1c66404b02b
  rdf:type ptn:SAXSymbol ;
  ptn:symbol "5"^^xsd:int ;
  ptn:lowerBound "645.651281059915"^^xsd:double ;
  ptn:upperBound "700.959674546294"^^xsd:double .
```

Figure 3: A snippet of the RDF graph with motifs and associated rules (Petrova 2019)

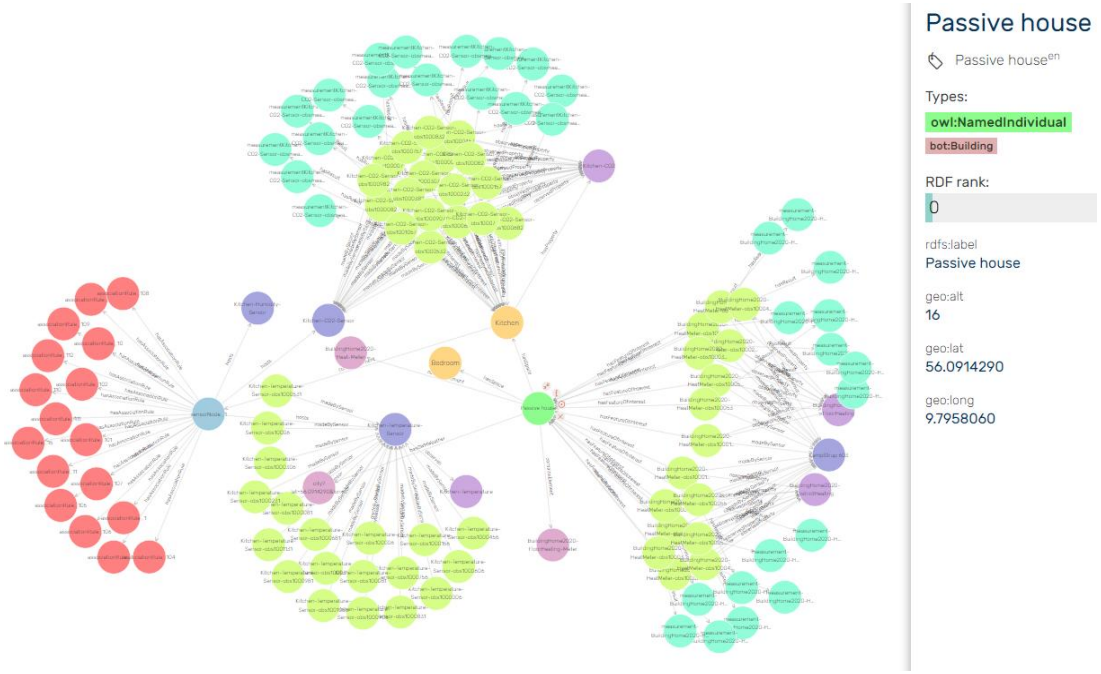

Figure 4: Semantic graph of Home2020 containing the building URIs and the related spaces, sensor nodes, occupants, sensor data and the motifs and association rules discovered in the data (Petrova 2019)

3.2 Visualization of Building Performance Patterns and Association Rules

The performance-enriched and contextualized semantic building graph allows dedicated information retrieval, and most importantly provides both the necessary infrastructure for capturing domain expertise and the input for interpretation and semantic enrichment of the association rules. To allow that, the knowledge embedded in the graph needs to be presented to the domain experts in a structured way, which requires a user interface and a data model that allows to store the meaning, be able to update it and embed it in the knowledge base for further reuse in evidence-based design processes. Therefore, to facilitate the process of knowledge interpretation, the association rules and corresponding patterns have been visualized to enable a better understanding. Figure 5 shows the visualization of association rule 453 ==> 485 #SUP: 3 #CONF: 0.6, which means that every three out of five times when pattern 453 appears throughout the dataset, pattern 485 also appears. The figure exemplifies the motifs with their ID and SAX sequences, the interval that the symbols are in, the observed variables in which they appear, the relationship between them and highlights the support measure. Such a visualization enables a much easier expert interpretation than the formal output of the algorithm as visualized in Fig. 2.

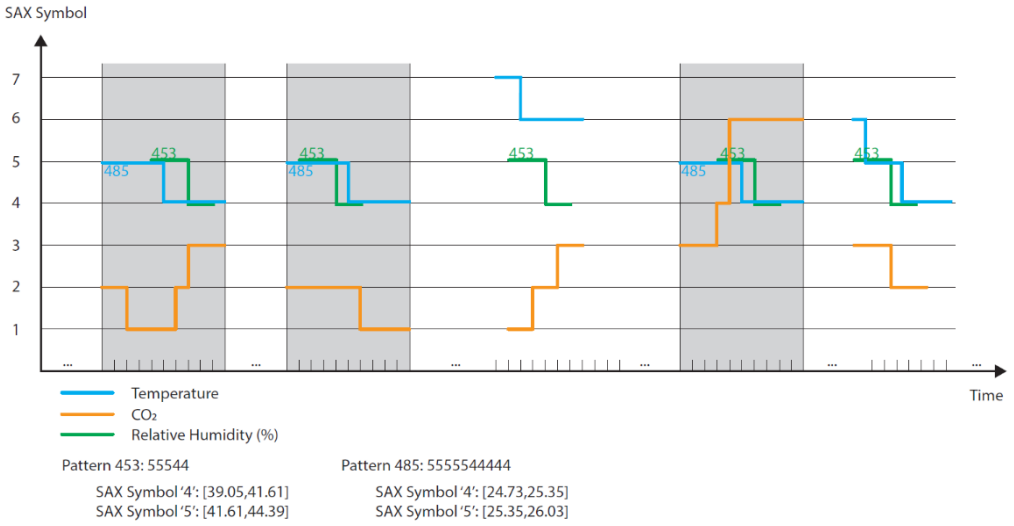

Figure 5: A visualization of an association rule discovered in indoor environmental quality data, the motifs that it contains of and their corresponding SAX sequences (Petrova 2019)

That way and for the given example, it is much easier to confirm that whenever the indicated interval sequence in Relative Humidity occurs, there is a 60% likelihood that the corresponding interval sequence in Temperature also occurs throughout the entire dataset. Being visualised, the association rules and motifs can then be semantically enriched through an appropriate data model that fits the structure of the knowledge base.

4 Capturing Domain Expertise - The Effect of the Crowd on the Semantic Enrichment of Association Rules

To achieve the semantic enrichment of association rules, this study aims to combine the powerful pattern recognition capability of machines with the domain expertise of humans. It is hereby important to distinguish between domain expertise or knowledge with regards to formal ontologies for semantic representation of data, and domain knowledge in terms of the human expertise required to interpret building performance patterns.

In this study, both concepts are applied accordingly, as ontologies are used for knowledge representation and storing in the semantic graphs and human expertise is harvested for the semantic enrichment of the association rules. The provided human domain expertise is also mapped to a formal ontology and added to the semantic graph. Figure 6 presents the architecture of the intended knowledge capture system and the interaction between the experts and the knowledge base during the interpretation and semantic enrichment of the association rules.

Fundamentally, the defined semantic enrichment approach and the corresponding system architecture rely on the concept of the "crown truth" and the notion that collecting annotations of the same objects of interpretation across a crowd reduces subjectivity, provides more meaningful representations and much more reasonable interpretations. In other words, the semantic enrichment system relies on the dominance of the human domain expertise, which is solicited from an expert crowd. The expert crowd in this case consists of indoor environmental quality experts with various levels of

expertise, years of experience, area of expertise (thermal, visual, acoustic, atmospheric), etc. All association rules are stored in the knowledge base and can be retrieved as soon as a domain expert logs in and activates their profile. For each association rule, each expert can define new meaning by annotation, perform classification with semantic tags or review existing interpretations by upvoting or updating. That input gets stored in the semantic graph together with a reference to the Uniform Resource Identifier (URI) of the corresponding domain expert who provided the input. Under the effect of the crowd, the association rules with highest level of interestingness and usefulness become visible, including annotations that would allow retrieval of the semantically rich building performance metrics. The following section will, therefore, define the technical aspects of the outlined expert crowd-centric semantic enrichment mechanism.

4.1 Crowdsourcing Building Performance Patterns

Crowdsourcing as an approach responds to the above-described notion of the crowd truth and provides an opportunity to capture collective intelligence and knowledge that are otherwise dispersed (Schenk and Guittard 2011). As a result, crowdsourcing has received major attention in various domains, e.g. image recognition, fabrication, design (Xiang et al. 2018). That also applies to the Semantic Web domain, where crowdsourcing techniques have been used for semantic annotation, ontology engineering, knowledge base curation and linked data quality assurance (Sack 2014; Sarasua et al. 2015). AEC research demonstrates the implementation of crowdsourcing for BIM-based construction material libraries through annotation of site photo logs (Han and Golparvar-Fard 2017), annotation of construction workers on site (Liu and Golparvar-Fard 2015), co-creation of infrastructure as-built BIM models and infrastructure maintenance (Consoli and Reforgiato 2015).

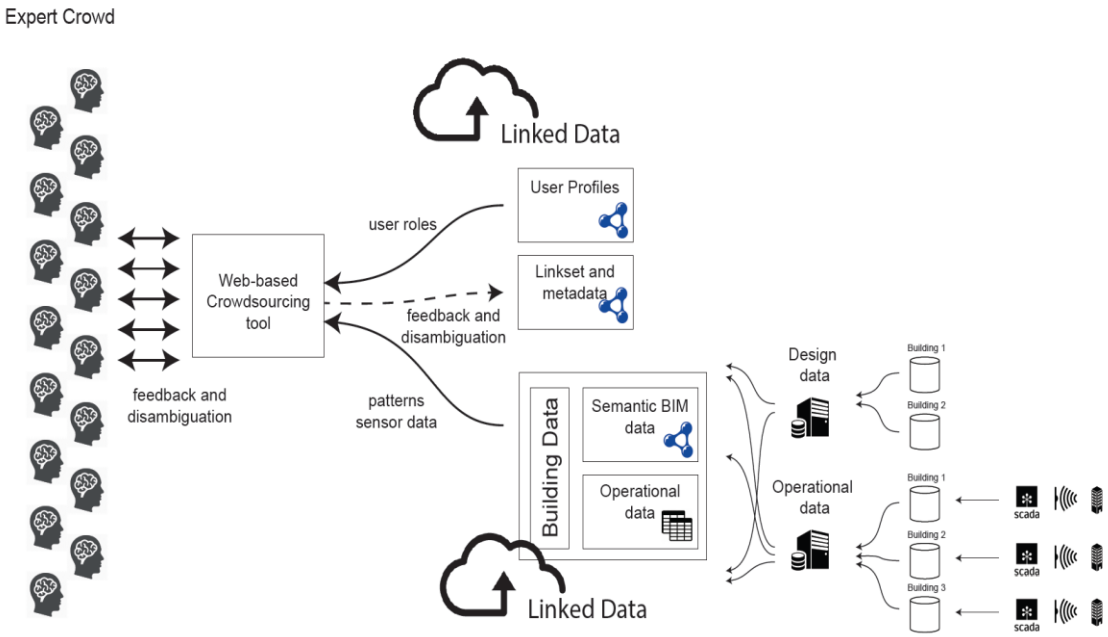

Figure 6: System architecture of the proposed crowdsourcing system (Petrova 2019)

The adopted type of crowdsourcing platform in this research effort is what is defined as “Information Pooling”, i.e., based on additive aggregation of distributed information

and aiming to integrate diverse opinions, assessments, predictions or other kinds of information from contributors (Blohm et al. 2018). Using that principle, the expert annotations are collected through the crowdsourcing platform and stored directly in the semantic graph. Domain experts are hereby modelled using the FOAF ontology, whereas their input is modelled according to the schema.org ontologies, which provide an opportunity to use Review and Commenting mechanisms. In this case, expert Reviews and Comments are linked directly to the schema:CreativeWork class. In addition, the schema:Person class can also be used for defining the human experts.

Alternatively, the Review ontology can also be used, however, schema.org provides more flexibility and dimension to the linked data-based crowdsourcing effort, as it allows storing votes (e.g. schema:upvoteCount). Furthermore, Reviews, Comments, and CreativeWorks can be combined and further enriched by adding metadata to each of them (agent, about, dateCreated, text, etc.). That allows a much bigger flexibility in terms of semantic annotation, tagging and adding of descriptions for further clarification of the expert interpretations. This principle and resulting data model is depicted in Fig. 7.

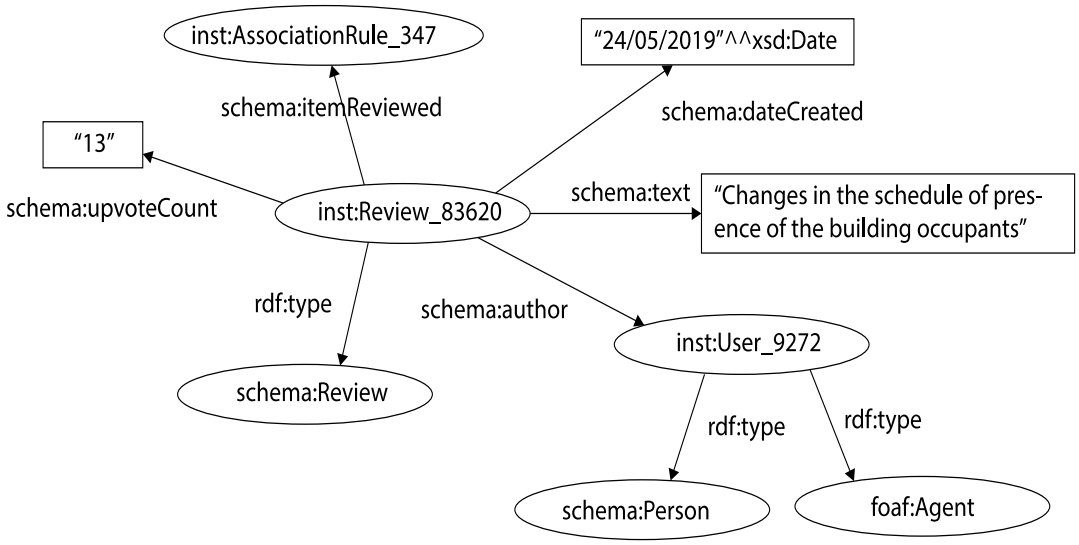

Figure 7: Data model for semantic annotation and interpretation of association rules (Petrova 2019)

4.2 Semantic Annotation and Classification of Association Rules

So far, the described approach allows the domain expertise to be input and stored in the semantic graph in the form of reviews and descriptions defined by the human experts. However, it does not provide semantically definitive tags or classifications, which are necessary for information retrieval. Thus, semantic tags have been further created so that the provided reviews and descriptions could be formally annotated and classified. The tags have been created on the basis of the most usual causes of any frequent performance patterns appearing in sensor observations from buildings. Typically, such patterns are related to dynamic parameters that influence building performance directly. The semantic tags for annotation of association rules are, therefore, identified by the most probable reason for the occurrence of the patterns and defined as (1) external

conditions, (2) occupant behaviour, (3) system performance, (4) design and (5) construction. Thus, expert input is classified and annotated with these tags.

Furthermore, the crowdsourcing system allows adding previously undefined subtags, should such be deemed necessary by the domain experts for clarification of a rule. Figure 8 illustrates the principle behind the semantic tagging and classification. Naturally, the data model has to be implemented in an application, which the expert crowd can interact with to complete the semantic enrichment of the association rules. However, the development of such an application and user interface is beyond the scope of this paper. Finally, even with the semantic annotation and classification in place, it is still not possible to assess the value of the semantically enriched association rules. In other words, further input is required to filter interesting association rules that point to abnormal or unexpected patterns and exclude the expected dependencies. To achieve that, we rely on the previously mentioned Upvote option as provided by the schema.org ontology (schema:upvoteCount), thereby allowing the domain experts to perform Input (Annotation)- Review- Upvote cycles and enrich association rules, but also indicate a level of interestingness that is based on both statistical measures and expertise. Figure 9 depicts the proposed crowdsourcing mechanism.

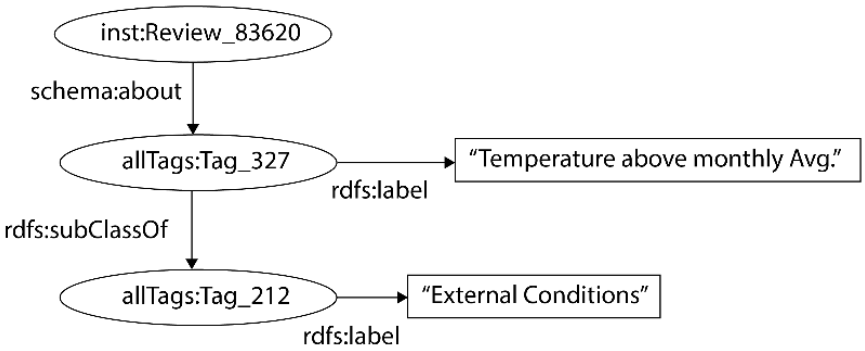

Figure 8: Semantic tags for classification of expert reviews (Petrova 2019)

5 Direct Semantic Stream RDF Graph Pattern Discovery

As seen so far, ARM is an effective method of discovering frequent patterns in building performance. The described crowdsourcing approach can also be effective in the semantic enrichment and interpretation of the discovered rules. However, it has to be acknowledged that the sole reliance on the users' intervention can be time-consuming and error-prone, especially in the cases of large amounts of data. Also, the measures of interestingness (confidence and support) consider the knowledge at instance-level and any available knowledge at schema level is disregarded, which may have a negative impact on the actual interpretation. In this regard, several studies suggest RDF stream processing as an alternative, i.e. converting the raw sensor data streams into RDF streams and use semantic data mining approaches on the resulting graph to identify association rules. Therefore, we further look into the RDF stream processing and graph pattern recognition method and discuss to what extent it could be compared to the previously described approach.

5.1 RDF Stream Processing and RDF Graph Pattern Recognition

Several researchers state that to enable stream processing, we should move from storing semantic data in batches and querying it ("one-time semantics") to using query languages with streaming extensions to perform continuous queries on the semantic data streams ("continuous semantics") (Della Valle et al. 2009; Calbimonte et al. 2012). In that relation, the main steps to publishing sensor data as RDF streams have also been defined and include conversion from sensor data streams to RDF streams, storing the resulting RDF streams, and linking them with other relevant datasets. That requires the selection of relevant ontologies, defining an appropriate mapping language for conversion, selection of continuous query languages and choosing relevant datasets to link to (Llanes et al. 2016) (Fig.10).

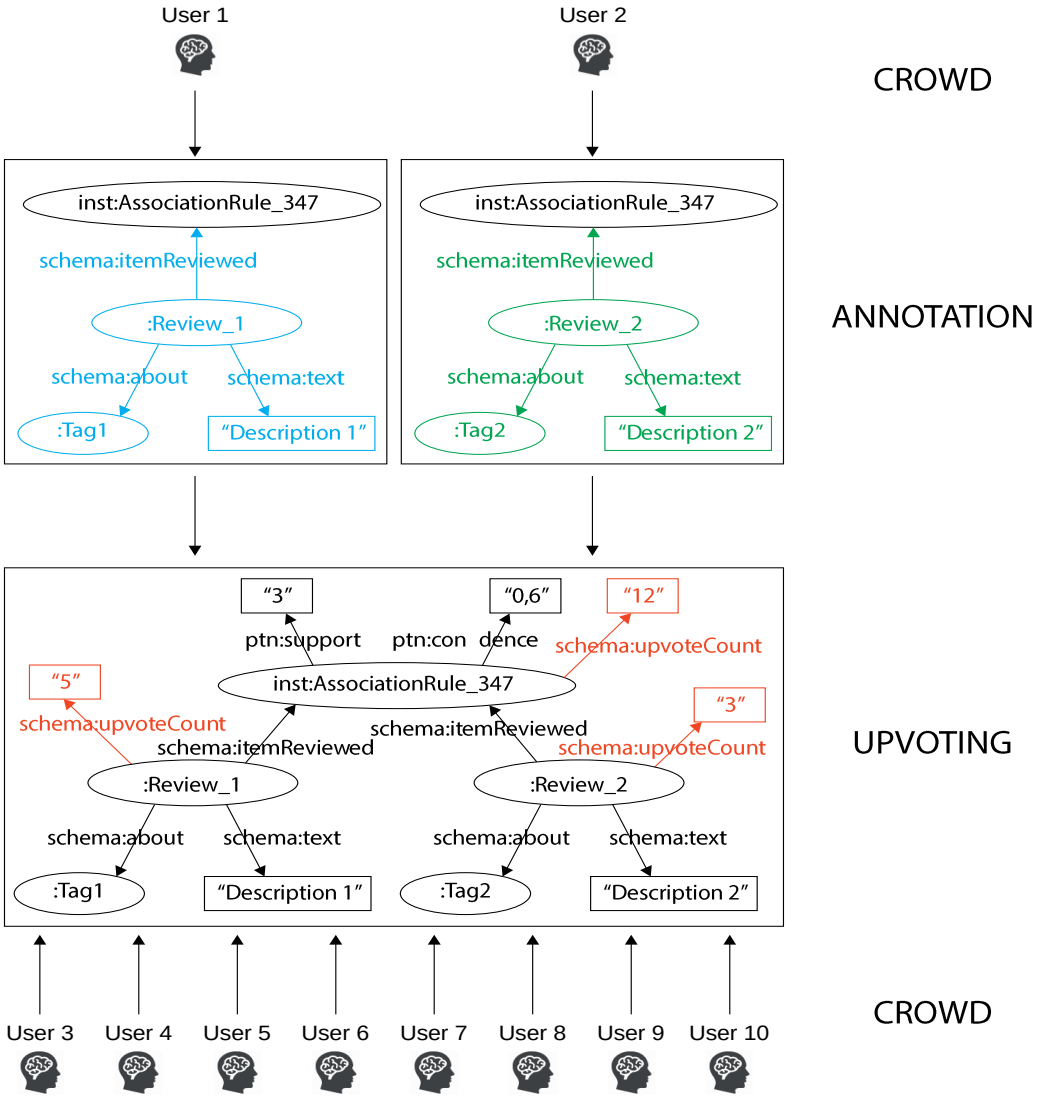

Figure 9: A snippet from the semantic graph containing the expert annotations and reviews of discovered association rules and the crowdsourcing process (Petrova 2019)

To demonstrate the principle of pattern recognition within the RDF graph structure, we employ a method for frequent RDF graph pattern detection in semantic data streams, which relies on the graph predicates (Belghaouti et al. 2016).

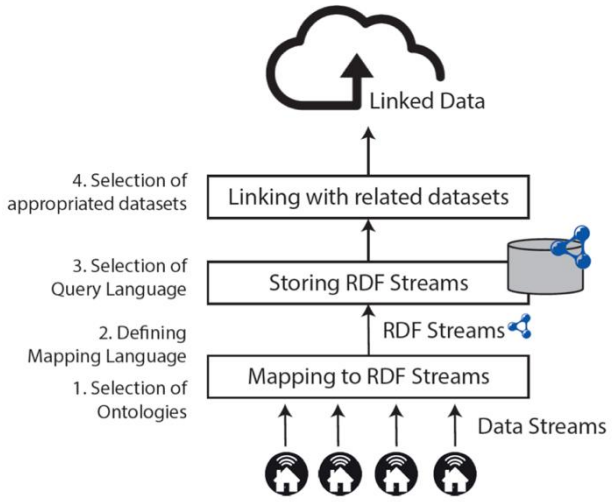

Figure 10: Process of publishing sensor data streams as RDF streams

Each graph in the stream data can be represented as a directed star graph as shown on the left side in Fig. 11. The proposed method relies on the fact that the streams are represented according to particular ontologies, which means that most streams will be relatively uniform and expose a very frequent RDF graph structure.

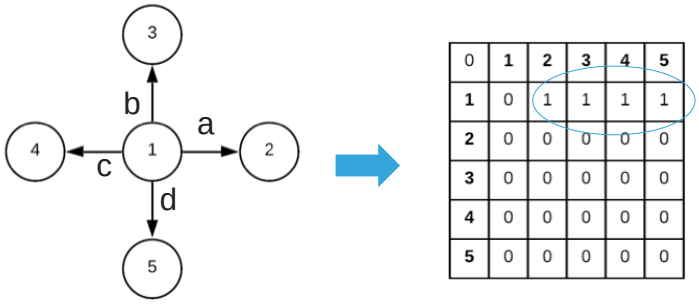

Figure 11: A directed star graph as represented in a RDF stream and the corresponding adjacency matrix based on the graph predicates, based on Belghaouti et al. (2016)

According to Belghaouti et al. (2016), RDF graphs could then be represented using adjacency matrices, however, such an approach would not be fully efficient and is not suitable for RDF graphs as it will result in a very sparse matrix (right in Fig.11). Therefore, it is proposed to reduce the associated adjacency matrix to a bit vector (Fig.12).

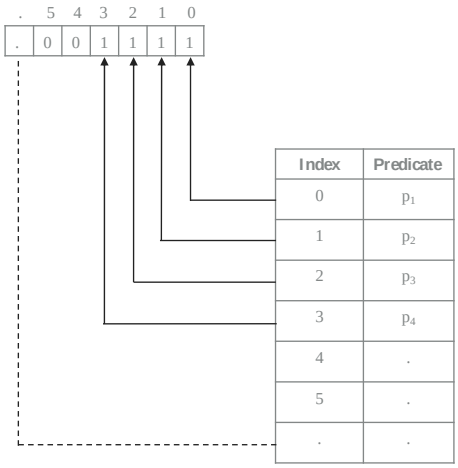

Figure 12: Principle of construction of the graph pattern using a bit vector based on the graph predicates, based on Belghaouti et al. (2016)

Based on the bit vector it is then possible to construct the Predicates Hash Table (PHT), which contains all the detected predicates in RDF graphs of an input stream and holds them, as well as the Graph Hash Table (GHT), which detects all the RDF graph patterns and holds them. In theory, that makes it possible to observe the evolution of the stream and the detected frequent RDF graph patterns. However, since the predicates are identical, the only change will be in the numerical value of the observations, which the graph pattern recognition method cannot account for. To demonstrate the detection of the graph patterns, we consider again the building graph of Home2020 illustrated in Figure 4 and extracted star shaped graphs storing the Relative Humidity and Temperature observations from the indoor environmental quality data stream from the kitchen. As seen from the examples, the RDF graph structure is rather consistent, also in terms of ontological representation, which is consistent with the outlined in (Belghaouti et al. 2016). The observed differences stem from the type of the observed variables and their corresponding units of measure.

```
inst:Kitchen-Humidity-Sensor-obs1132308
                rdf:type sosa:Observation ;
                sosa:hasFeatureOfInterest inst:Kitchen ;
                sosa:hasResult inst:measurementKitchen-Humidity-Sensor-obsmeas1132308 ;
                sosa:madeBySensor inst:Kitchen-Humidity-Sensor ;
                sosa:observedProperty inst:Kitchen-Humidity ;
                sosa:resultTime "22/01-2018 10:35:45"^^xsd:dateTime .

inst:measurementKitchen-Humidity-Sensor-obsmeas1132308
                rdf:type om:Measure ;
                om:hasNumericalValue "43.0"^^xsd:double ;
                om:hasUnit om:percent .


inst:Kitchen-Temperature-Sensor-obs2913631
                rdf:type sosa:Observation ;
                sosa:hasFeatureOfInterest inst:Kitchen ;
                sosa:hasResult inst:measurementKitchen-Temperature-Sensor-obsmeas2913631 ;
```

```
                sosa:madeBySensor inst:Kitchen-Temperature-Sensor ;
                sosa:observedProperty inst:Kitchen-Temperature ;
                sosa:resultTime "14/04-2018 22:15:45"^^xsd:dateTime .

inst:measurementKitchen-Temperature-Sensor-obsmeas2913631
                rdf:type om:Measure ;
                om:hasNumericalValue "24.0"^^xsd:double ;
                om:hasUnit om:degreeCelsius .
```

Following the described methodology, we can then construct the bit vector of the graph and identify the repetitive graph pattern. That is hereby demonstrated with the Temperature observation (Fig.13). As seen in the single example below, the detected frequent RDF pattern is rather different in nature from the performance patterns identified and interpreted earlier. While the graph pattern presents a variety of semantically rich and highly contextual data, it does not contain any explicit semantics related to building performance behaviour. As it is identified based on the graph predicates, the presented graph pattern could provide information about the evolution of the stream over time if predicates change, but any behavioural insights, correlations between performance variables or causations need to be discovered in alternative ways.

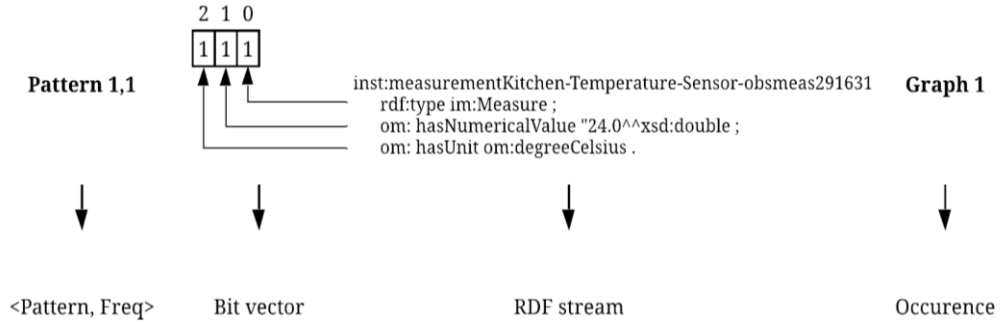

Figure 13: RDF graph pattern detection in indoor environmental quality data

6 Conclusion

The rapidly increasing amount of monitored building data allows using novel statistical and symbolic AI approaches for discovery of valuable knowledge in building performance. Such methods have come a long way in processing vast amounts of data, finding patterns and relationships and predicting trends, thereby enhancing human decision-making in the building performance improvement domain. However, regardless of how powerful pattern recognition, knowledge representation or information retrieval techniques are applied, the interpretation of the discovered building performance patterns is in the hands of the domain experts, who usually need to apply domain expertise to interpret their meaning and implications on the overall building performance. Data analytical output usually does not convey any explicit semantics and its value is dependent on the contextualization and interpretation stemming from domain expertise.

This paper approaches this issue with a novel methodology for semantic enrichment of discovered frequent repetitive patterns and association rules in monitored indoor environmental quality data from a passive house in Denmark. By applying motif discovery and Association Rule Mining, we obtain patterns and rules that are

represented and stored in a semantic building graph together with other available building data by the use of several domain ontologies. For semantic enrichment, interpretation and contextualization, we devise a linked-data based crowdsourcing mechanism, which captures domain expertise in the form of semantic annotation and classification. That results in a comprehensive knowledge base that stores not only building data and indoor environmental quality sensor observations, but also building performance patterns and their meaning. Such a knowledge base can be of high value in evidence-based design processes and building performance assessment and improvement. Furthermore, we compare the applied approach, which combines knowledge discovery and semantic data modelling to a direct RDF graph pattern mining approach to assess feasibility and potential.

With regards to the latter, several important observations need to be addressed. First and foremost, while machine learning approaches for Association Rule Mining are rather effective for detecting frequent patterns, there are several manual steps related to data treatment and parameter selection. Moreover, a traditional data mining approach such as this one analyses the data in batches and the discovered knowledge is, therefore, only locally valid for the dataset in question. Second, while Semantic Web and linked data technologies allow representing and storing the discovered knowledge, but the value of the discovered performance patterns and rules lies in their meaning, which is not explicit, unless interpreted by a domain expert. And while the presented crowdsourcing mechanism provides a solution to that, it still has to be acknowledged that the contributions of the expert crowd used for interpretation and semantic enrichment of the rules may vary. That means that an additional validation layer may be necessary. It has to be noted, that even though the semantically enriched performance patterns are stored in the graph and can be retrieved, they do not provide direct solutions in terms of, for instance, design decision support or building performance optimization. They serve merely as an evidential layer to human decision-making.

Finally, being based on the graph predicates, the demonstrated frequent RDF graph pattern detection method could provide an insight about the evolution of semantic sensor data streams based on the graph predicates, but does not provide any actual building performance insight, as the pattern recognition is solely based on the graph structure and not on the numerical values of the sensor observations. Therefore, future research in that direction may rely on graph alignment techniques to harvest the benefits of both methodologies.

The proposed crowdsourcing mechanism can be of high value to engineering practice in several ways. By combining both knowledge discovery and semantic data modelling approaches, the crowdsourcing platform establishes the missing link between machine learning output and the human domain knowledge necessary for its interpretation. That enables the reuse of highly valuable and complex engineering knowledge and can also serve as an educational mechanism for understanding dynamics in building performance and indoor environmental quality parameters. The system also enables the creation of a feedback loop between building operation and design and helps practitioners learn from the behaviour of the existing building stock, engage with the streaming sensor data, understand it and bring out the value in it.

7 References

Agrawal, R., Imielinski, T., and Swami, A. (1993). Mining association rules between sets of items in large databases. In: *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, Washington, USA, pp. 207-216, https://doi.org/10.1145/170035.170072.

Belghaouti, F., Bouzeghoub, A., Kazi-Aoul, Z., and Chiky, R. (2016). FreGraPaD: Frequent RDF Graph Patterns Detection for semantic data streams. In: *2016 IEEE Tenth International Conference on Research Challenges in Information Science (RCIS)*, pp. 1-9, https://doi.org/ 10.1109/RCIS.2016.7549333.

Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The Semantic Web. *Scientific American*, 284, pp. 34–43.

Bizer, C., Heath, T., and Berners-Lee, T. (2009). Linked data - the story so far. *International Journal on Semantic Web and Information Systems,* 5(3), pp. 1-22, https://doi.org/10.4018/jswis.2009081901.

Blohm, I., Zogaj, S., Bretschneider, U., and Leimeister, J. (2018). How to manage crowdsourcing platforms effectively? *California Management Review*, 60(2), pp. 122–149, https://doi.org/10.1177/0008125617738255.

Borrmann, A., König, M., Koch, C., and Beetz, J. (2018). *Building Information Modeling: Technology foundations and industry practice.* 1 ed. Springer.

Calbimonte, J.-P., Jeung, H., Corcho, O., and Aberer, K. (2012). Enabling query technologies for the semantic sensor web. *International Journal on Semantic Web and Information Systems*, 8(1), pp. 43–63, https://doi.org/10.4018/jswis.2012010103.

Consoli, S. and Reforgiato, R. (2015). An urban fault reporting and management platform for smart cities. In: *Proceedings of the 24th International Conference on World Wide Web*, Florence, Italy, pp. 535–540, https://doi.org/10.1145/2740908.2743910.

Curry, E., O'Donnell, J., Corry, E., Hasan, S., Keane, M., and O'Rian, S. (2013). Linking building data in the cloud: Integrating cross-domain building data using linked data. *Advanced Engineering Informatics*, 27, pp. 206-219, https://doi.org/10.1016/j.aei.2012.10.00.

Della Valle, E., Ceri, S., van Harmelen, F., and Fensel, D. (2009). It's a streaming world! Reasoning upon rapidly changing information. *IEEE Intelligent Systems*, 24(6), pp. 83–89, https://doi.org/10.1109/MIS.2009.125.

D'Oca, S., Hong, T., and Langevin, J. (2018). The human dimensions of energy use in buildings: A review. *Renewable and Sustainable Energy Reviews*, 81, pp. 731 – 742, https://doi.org/10.1016/j.rser.2017.08.019.

El-Diraby, T. E. (2013). Domain ontology for construction knowledge. *Journal of Construction Engineering and Management*, 139(7), pp. 768-784, https://doi.org/10.1061/(ASCE)CO.1943-7862.0000646.

Fan, C., Xiao, F., Madsen, H., and Wang, D. (2015). Temporal knowledge discovery in big BAS data for building energy management. *Energy and Buildings*, 109, pp. 75–89, https://doi.org/10.1016/j.autcon.2014.12.006.

Fan, C., Xiao, F., Li, Z., and Wang, J. (2018a). Unsupervised data analytics in mining big building operational data for energy efficiency enhancement: A review. *Energy and Buildings,* 159, pp. 296–308, https://doi.org/10.1016/j.enbuild.2017.11.008.

Fan, C., Sun, Y., Shan, K., Xiao, F., and Wang, J. (2018b). Discovering gradual patterns in building operations for improving building energy efficiency. *Applied Energy,* 224, pp. 116 – 123, https://doi.org/10.1016/j.apenergy.2018.04.118.

Fan, C., Xiao, F., Yan, C., Liu, C., Li, Z., and Wang, J. (2019). A novel methodology to explain and evaluate data-driven building energy performance models based on

interpretable machine learning. *Applied Energy*, 235, pp. 1551-1560, https://doi.org/10.1016/j.apenergy.2018.11.081.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, 17, pp. 37–54, https://doi.org/10.1145/240455.240463.

Han, K. and Golparvar-Fard, M. (2017). Crowdsourcing BIM-guided collection of construction material library from site photologs. *Visualization in Engineering*, 5(14), https://doi.org/10.1186/s40327-017-0052-3.

Hoehndorf, R. and Queralt-Rosinach, N. (2017). Data Science and symbolic AI: Synergies, challenges and opportunities. *Data Science,* 1, pp. 27–38, https://doi.org/10.3233/DS-170004.

Lausch, A., Schmidt, A., and Tischendorf, L. (2015). Data mining and linked open data new perspectives for data analysis in environmental research. *Ecological Modelling*, 295, pp. 5–17, https://doi.org/10.1016/j.ecolmodel.2014.09.018.

Lin, J., Keogh, E., Wei, and Lonardi, S. (2007). Experiencing SAX: a novel symbolic representation of time series. *Data Mining and Knowledge Discovery*, 15, pp. 107–144, https://doi.org/10.1007/s10618-007-0064-z.

Liu, K. and Golparvar-Fard, M.: Crowdsourcing construction activity analysis from jobsite video streams. *Journal of Construction Engineering and Management*, 141(11), https://doi.org/10.1061/(ASCE)CO.1943-7862.0001010.

Llanes, K., Casanova, M., and Lemus, N. (2016). From sensor data streams to linked streaming data: A survey of main approaches. *Journal of Information and Data Management*, 7(2), pp. 130–140.

McGlinn, K., Wagner, A., Bonsma, P., McNerney, L., and O'Sullivan, D. (2019). Interlinking geospatial and building geometry with existing and developing standards on the web. *Automation in Construction*, 103, pp. 235–250, https://doi.org/10.1016/j.autcon.2018.12.026.

Miller, C., Nagy, Z., and Schlueter, A. (2018). A review of unsupervised statistical learning and visual analytics techniques applied to performance analysis of non-residential buildings. *Renewable and Sustainable Energy Reviews*, 81, pp. 1365 – 1377, https://doi.org/10.1016/j.rser.2017.05.124.

Pauwels, P. and Terkaj, W. (2016). EXPRESS to OWL for construction industry: towards a recommendable and usable ifcOWL ontology. *Automation in Construction*, 63, pp. 100-133, https://doi.org/10.1016/j.autcon.2015.12.003.

Pauwels, P., Zhang, S., and Lee, Y.-C. (2017). Semantic web technologies in AEC industry: a literature review. *Automation in Construction*, 73, pp. 145–165, https://doi.org/10.1016/j.autcon.2016.10.003.

Petrova, E., Pauwels, P., Svidt, K., and Jensen, R.L. (2018). In Search of Sustainable Design Patterns: Combining Data Mining and Semantic Data Modelling on Disparate Building Data. In: I. Mutis & T. Hartmann (eds.) *Advances in Informatics and Computing in Civil and Construction Engineering*, pp. 19-27, Springer.

Petrova, E. (2019). *AI for BIM-based sustainable building design: Integrating knowledge discovery and semantic data modelling for evidence-based design decision support.* Ph.d.-serien for Det Ingeniør- og Naturvidenskabelige Fakultet, Aalborg Universitet, Aalborg Universitetsforlag.

Petrova, E., Pauwels, P., Svidt, K., and Jensen, R.L. (2019). Towards Data-Driven Sustainable Design: Decision Support based on Knowledge Discovery in Disparate Building Data. *Architectural Engineering and Design Management,* 15 (5), pp. 334-356, https://doi.org/10.1080/17452007.2018.1530092.

Rasmussen, M.H., Pauwels, P., Hviid, C.A., and Karlshøj, J. (2017). Proposing a central AEC ontology that allows for domain specific extensions. In: F. Bosche, I. Brilakis, R. Sacks (eds.), *Proceedings of the Joint Conference on Computing in Construction*, Heraklion, Crete, Greece, pp. 237-244, https://doi.org/10.24928/JC3-2017/0153.

Rasmussen, M.H., Lefrançois, M., Pauwels, P., Hviid, C.A., and Karlshøj, J. (2019). Managing interrelated project information in AEC Knowledge Graphs. *Automation in Construction*, 108, 102956, https://doi.org/10.1016/j.autcon.2019.102956.

Ristoski, P. and Paulheim, H. (2016). Semantic web in data mining and knowledge discovery: A comprehensive survey. *Web Semantics: Science, Services and Agents on the World Wide Web*, 36, pp. 1–22, https://doi.org/10.1016/j.websem.2016.01.001.

Sack, H. (2014). Crowdsourcing for evaluation and semantic annotation. In: A. Bernstein, J. Leimeister, N. Noy, C. Sarasua, E. Simperl (eds.), *Crowdsourcing and the Semantic Web*, Dagstuhl Publishing, Germany, pp. 43–44, https://doi.org/10.4230/DagRep.4.7.25.

Sacks, R., Eastman, C., Lee, G., and Teicholz, P. (2018). *BIM Handbook: A guide to Building Information Modeling for owners, designers, engineers, contractors and facility managers*. 3 ed. Wiley, Hoboken, NJ, USA.

Sarasua, C., Simperl, E., Noy, N., Bernstein, A., and Leimeister, J. (2015). Crowdsourcing and the semantic web: A research manifesto. *Human Computation*, 2(1), pp. 3–17, https://doi.org/10.15346/hc.v2i1.2.

Schenk, E. and Guittard, C. (2011). Towards a characterization of crowdsourcing practices. *Journal of Innovation Economics*, 7, pp. 93–107, https://doi.org/10.3917/jie.007.0093.

Ukkonen, E. (1995). On-line construction of suffix trees. *Algoritmica*, 14(3), pp. 249-260, https://doi.org/10.1007/BF01206331.

Xiang, W., Sun, L., You, W., and Yang, C. (2018). Crowdsourcing intelligent design. *Frontiers of Information Technology & Electronic Engineering*, 19(1), pp. 126–138, https://doi.org/10.1631/FITEE.1700810.