Semantic Prediction Assistant Approach applied to Energy Efficiency in Tertiary Buildings

Iker Esnaola-Gonzalez [a,b,*], Jesús Bermúdez [b], Izaskun Fernandez [a], and Aitor Arnaiz [a]

[a] *IK4-TEKNIKER, Iñaki Goenaga 5, 20600 Eibar, Spain*
E-mail:{iker.esnaola, izaskun.fernandez, aitor.arnaiz}@tekniker.es
[b] *University of the Basque Country (UPV/EHU), Paseo Manuel Lardizabal 1, 20018 Donostia-San Sebastián, Spain*
E-mail: jesus.bermudez@ehu.eus

Abstract. Fulfilling occupants' comfort whilst reducing energy consumption is still an unsolved problem in most of tertiary buildings. However, the expansion of the Internet of Things (IoT) and Knowledge Discovery in Databases (KDD) techniques lead to research this matter. In this paper the EEPSA (Energy Efficiency Prediction Semantic Assistant) process is presented, which leverages the Semantic Web Technologies (SWT) to enhance the KDD process for achieving energy efficiency in tertiary buildings while maintaining comfort levels. This process guides the data analyst through the different KDD phases in a semi-automatic manner and supports prescriptive HVAC control strategies. That is, temperature of a space is predicted simulating the activation of HVAC systems at different times and intensities, so that the facility manager can choose the strategy that best fits both the user's comfort needs and energy efficiency. Furthermore, results show that the proposed solution improves the accuracy of predictions.

Keywords: Semantic Web Technologies, Knowledge Discovery in Databases, Energy Efficiency, Buildings

1. Introduction

Concerns over changing climatic conditions (i.e. global warming, depletion of ozone layer, etc.), energy security, and adverse environmental effects are growing among governments, researchers, policy makers, and scientists in developed as well as developing countries [75]. In order to meet the energy sustainability and minimize the climate change, the European Commission agreed a set of binding legislations inside the EU 2020 package. One of the spotlighted sectors re-garding this package is the building sector which, according to the UNEP (United Nations Environment Programme), consumes about 40% of global energy and is responsible for 36% of CO2 emissions in the EU. Therefore, efficient management of building energy plays a vital role and is becoming the trend for a future generation of buildings.

However, energy efficiency is not the only concern related with buildings. Since approximately 90% of people spend most of their time in buildings, feeling comfortable indoors is a must and poses a huge impact to preserve inhabitant's health, morale, working efficiency, productivity and satisfaction. As a conse-

---

*Corresponding author, e-mail: iker.esnaola@tekniker.es

quence, a system is needed which fulfils the occupants' expected comfort index whilst reducing energy consumption during the operation of a building. In this context, the expansion of the Internet of Things (IoT) and Knowledge Discovery in Databases (KDD) techniques will lead to both researching the reduction of such prominent impact and the improvement of comfort levels.

The achievement of energy efficiency while maintaining users' comfort in tertiary buildings is not a trivial question. There are many complementary ways to save and optimize energy use in buildings, but since temperature is the most important ambient parameter affecting electric load, forecasted indoor temperatures constitute a basic ingredient in energy efficiency plans [1].

Let us consider the following scenario. The facility manager of a given building seeks to establish an activation strategy for the HVAC (Heating, Ventilation and Air Conditioning) system, so that energy is used in the most efficient way while maintaining the optimal comfort[1] levels for the space occupants. In order to support prescriptive HVAC control strategies, space temperature predictions can be used as a base. That is, temperature of the space is predicted simulating the activation of HVAC system resources within the space at different times and intensities (e.g. activating all of them for four hours, activating half of them for six hours, etc.). Keeping this in mind, the facility manager requests a KDD process to obtain a temperature predictor model for the space under study.

The KDD can be understood as a five steps process leading to the extraction of useful knowledge from raw data [27], applicable for instance in decision support systems. The five steps can be summarized as follows:

1. Selection of datasets and subset of variables or data samples on which discovery will be performed.
2. Preprocessing tasks to ensure data quality and preparation for a subsequent analysis.
3. Transformation or production of a projection of the data to a form which data mining algorithms can work with and improve their performance.
4. Data mining by selecting the algorithm that best matches the user's goals and their application to search for hidden patterns.

5. Interpretation and evaluation of the results, patterns and models derived, in support of decision making processes.

This process can involve significant iteration and can contain loops between any two of the mentioned steps as can be seen in Figure 1.

Data analysts in charge of the KDD process are confronted with large, diverse and heterogeneous data. First of all, data related to the given space and its structural element properties including materials, heat transfer coefficients, and orientation of their boundaries. They also need to take into account information about sensors and actuators deployed in the building, their location, features and certainly their measurements. Likewise, data about weather conditions and weather forecasts for the building location are relevant. Furthermore, there is other information to consider such as the space occupancy, work schedule or human related organization. Under such circumstances where a deep energy efficiency and building domain knowledge is required, having insufficient domain expertise could make data analysts feel overwhelmed. Consequently, they typically resort to a trial and error approach searching for variables and tasks that could be confidently used to make accurate predictions. This is definitely an undesirable approach and it would be much more profitable to count with an assisted KDD process supported by technologies that enable the management of data semantics, data interrelationships, and knowledge representation.

Semantic Web Technologies (SWT) enable the explicit representation of knowledge both in human and machine understandable form. Moreover, SWT have been successfully used in the data integration as well as system interoperability tasks, and they enable the representation of expert knowledge obtained via knowledge elicitation processes. Once represented in semantic resources, SWT open a range of possibilities to exploit knowledge, such as full fledged data querying or further processing to infer new knowledge from implicitly hidden knowledge. Furthermore, if domain expert knowledge is adequately complemented with tools that support the assistance throughout the KDD process, its usability and exploitation capabilities will be at hand.

This paper presents the EEPSA (Energy Efficiency Prediction Semantic Assistant) process to address the aforementioned problematic scenario in the energy efficiency in tertiary buildings domain, leverag-

---

[1]Optimal comfort can be understood in many ways: a temperature that ranges between some given values, a temperature that varies less during a period of time, etc.
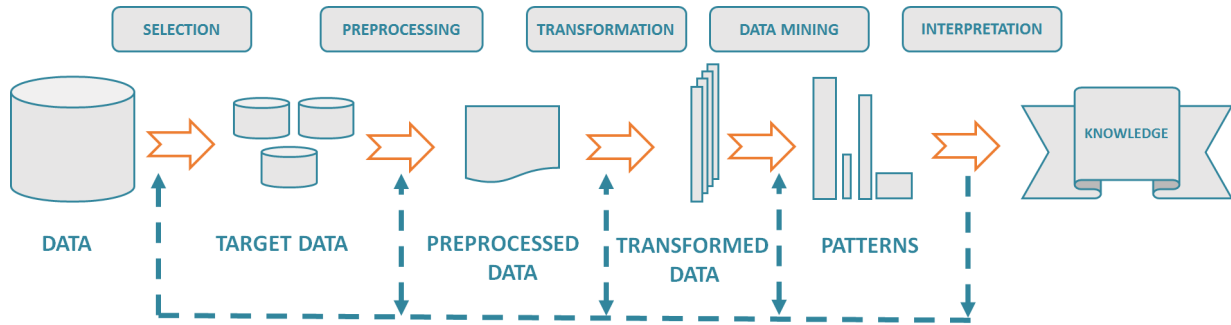
Fig. 1. An overview of the steps that compose the KDD Process proposed in [27].

ing Semantic Web Technologies such as ontologies, ontology-driven rules and ontology-driven data access.

The EEPSA process targets different KDD phases and each one poses its requirements. First of all, data needs to be semantically annotated with appropriated ontological terms. This semantic annotation is fundamental for enriching data, integrating heterogeneous data and representing it in a more domain-oriented way, as well as for enabling the improvement of the upcoming KDD phases. In the data selection phase the data analyst is assisted to decide which might be the most relevant variables for the matter at hand. Ontology-driven queries and inferencing capabilities support this task. The preprocessing phase intends to clean data from noise, missing values or inconsistencies to name a few. Ontology-driven rules help detecting such problematic data and classifying them according to their potential cause, as well as in proposing possible methods to fix them according to the established goal. The transformation phase generates additional knowledge in form of new attributes. Knowledge-driven rules, inferencing capabilities and external data sources are critical in this phase. All the enhancements in these phases could contribute to improve the robustness and performance of machine learning algorithms applied in the data mining phase and would ease the interpretation of the obtained results. Moreover, the proposed process is expected to be reusable in similar use cases of the same domain due to its high abstraction level.

Summarizing, the main contributions of this paper are the following:

– Description of a domain ontology that provides enough concepts and relations to express all the relevant information for the identified tasks and enables the representation of actionable expert knowledge.

– Outline of a process for assisting data analysts throughout a KDD process by leveraging SWT, with focus on some phases to show the mechanics of this proposal.
– A real-world evaluation of the proposed approach for illustrating the process.

The rest of this paper is structured as follows. Section 2 introduces the related work and analyses existing ontologies in the field. Section 3 presents the EEPSA ontology and the EEPSA process through the different KDD phases. Section 4 shows the application of this process on a real-world use case and evaluates and discusses obtained results. Finally the conclusions of this work are shown in section 5.

2. Related Work

2.1. KDD for Energy Efficiency in Buildings

KDD have traditionally been used to achieve energy efficiency in buildings such as in [33], where Artificial Neural Networks (ANN) and historic values have been used for short-time load forecasting in buildings. However, existing BMS (Building Management Systems) generally fail to fully optimize energy consumption in buildings. [35] states that current and forecasted information about events and weather (e.g. rain or snow) would help increasing the stability of the control systems minimizing energy consumption and increasing the occupants comfort. External meteorological conditions are used to improve the energy usage predictions in [4]. But not all external weather factors have the same impact in the energy consumption forecasting in buildings. In the use case presented in [51] for instance, effects of humidity and sun radiation had a less

significant impact in energy consumption, compared with the external temperature.

Related work in [50], [67] and [79] shows that not only external climatologic factors affect the energy use in buildings. Most modern buildings still condition rooms assuming maximum occupancy rather than actual usage. As a result, rooms are often over conditioned. [24] proposes different HVAC control strategies based on occupancy prediction of rooms. In a similar way [66] focuses on a better heating scheduling by predicting future occupancy. Wireless motion sensors and door sensors are used in [48] to infer occupants presence and activate or deactivate HVAC systems accordingly. [55] aims at developing predictive control strategies that use both weather and occupancy forecasts to limit peak electricity demand while maintaining high user comfort.

According to the related work shown in previous paragraphs, it has been proved that meteorological factors as well as occupancy of buildings have a significant impact both on the building energy consumption and comfort. The HVAC control strategies have also been deeply studied as a measure to achieve these two goals. However, the process of combining all these data sources into the KDD for exploiting them poses a big challenge. The research presented in this paper proposes the use of SWT towards a holistic approach to the improvement of the whole KDD process and obtained results.

2.2. Semantic Web Technologies for KDD

In the last years, advantages of semantic technologies for data understanding as well as for the data mining process itself have been highlighted in [42] and [60]. Furthermore, many approaches have proposed the use of Semantic Web data to enhance different KDD phases. Semantic Web Technologies address how one would discover the required data in today's chaotic information universe, how one would understand which datasets can be meaningfully integrated, and how to communicate the results to humans and machines alike.

According to [20], the Internet of Things (IoT) and Open Data are particularly promising in real time predictive data analytics for effective decision support, and the dynamic selection of Open Data and IoT sources for that purpose is the main challenge. Data quality is tackled in [28], [29] and [30], where data quality problems in Semantic Web data are identified by means of data validation rules. A review of the existing data quality work based on ontologies for the health domain is shown in [47]. In [62] desiderata and challenges for developing a framework for unsupervised generation of data mining features from Linked Data are identified. [43], [56] and [63] are examples of systems for enriching data with features that are derived from LOD (Linked Open Data). In [76] a feature-selection method based on ontology is proposed. The data mining environment RapidMiner [40] includes a LOD extension which provides a set of operators for augmenting existing datasets with additional attributes from open data sources [65]. In [53] semantic technologies are used to assist data scientists in selecting appropriate modelling techniques in the field of statistics or machine learning and building specific models as well as the rationale for the techniques and models selected. [38] presents an ontology to support the meta-learning for algorithm selection in the data mining, while in [6] one of the first intelligent discovery assistants is proposed. An overview of existing intelligent assistants for data analysis is provided in [68]. In [7] it has been noted that SWT can also have a potential impact in the Decision Support.

A detailed and extended survey on SWT within the KDD process can be found in [64]. The survey shows that, while many impressive results can be achieved already today, the full potential of Semantic Web Technologies for KDD is still to be unlocked.

Aforementioned work show that even though some initiatives apply SWT to improve a specific KDD phase, at the moment of writing this article no solution tackling the KDD process as a whole has been recognized. The research presented in this paper intends to be a preliminary effort towards that goal within the energy efficiency in tertiary buildings domain.

2.3. Existing Ontologies in the Field

BIM (Building Information Modelling) deals with the representation of functional and physical characteristics of a building [22]. That is, static information of a building element may be available and queryable in a BIM model; for example a door, its location, the material it is made of, and ocassionally, even when it was installed. But for instance, it is not possible to know whether the door is opened or closed in a given moment. This is why, in order to transform the building static data into live data, it is necessary to integrate information coming from IoT and sensing device network nodes. This data integration across several data sources can be obtained by adopting SWT. Further ap-

plications of SWT in this field are surveyed in [59]. All of them need conceptual foundation provided by ontologies.

Keeping this in mind, a brief summary of relevant ontologies of the current research domain is presented below. They cope with the building domain (ifcOWL, DogOnt, BOT), sensors and actuators domain (SSN, SAREF, FIESTA-IoT, IoT-O), and the weather domain (SmartHomeWeather). Other ontologies such as Semanco [49] or the Aemet Network of Ontologies [5] have also been analysed, but are not reflected in this paper. Some of the consulted surveys to identify these ontologies have been [23] and [44]. An interesting comparison between different IoT ontologies is also covered in [69]. The catalogues Linked Open Vocabularies [74] and LOV4IoT [34] have been used to search vocabularies covering desired concepts.

2.3.1. ifcOWL Ontology

IfcOWL ontology[2] provides an OWL representation of the Industry Foundation Classes (IFC) Schema which is the open standard for representing building and construction data. Using the ifcOWL ontology, one can represent building data in directed labelled graphs [58]. The graph model and the underlying web technology stack allows building data to be easily linked to material data, GIS (geographic information systems) data, product manufacturer data, sensor data, classification schemas, social data and so forth.

The ifcOWL ontology aims at supporting the conversion of IFC instance files into equivalent RDF files. It defines a faithful mapping of the IFC EXPRESS schema, which is the master schema for IFC models, and therefore replicates its object-oriented conceptualization, which has been found inconvenient for some practical engineering use cases (see [57]). Moreover, the ifcOWL conceptualization of some relationships and properties as instances of classes (i.e. *ifc:IfcRelationship*, *ifc:IfcProperty*) is counterintuitive to semantic web principles, that would expect OWL properties to represent them. A systematic transformation of this modelling issue has been presented in [19], producing the IfcWoD (IFC Web of Data) ontology, and some advantages of this semantic adaptation are claimed such as simplification of query writing, optimization of query execution and maximizing of inference capabilities. However, to the best of our knowledge, the IfcWoD ontology announced in that paper is not publicly available at the time of writing this article. In summary, the ifcOWL ontology is a necessary tool to incorporate IFC models to the semantic web infrastructure but is too complex for some use cases. IFC is used in construction industry and it rather focuses on building elements such as walls or doors, and their relations and geometries, with a granularity that is inconvenient for some scenarios. Furthermore, it is of secondary importance that an instance RDF file can be modelled from scratch using the ifcOWL ontology and an ontology editor.

2.3.2. DogOnt Ontology

The DogOnt ontology[3] allows to formalize all the aspects of IDEs (Intelligent Domotic Environment) and it is designed with a particular focus on interoperation between domotic systems [8]. Mainly covering device, state and functionality modelling, it also supports device independent description of houses, including both controllable and architectural elements. DogOnt provides different reasoning mechanisms corresponding to different goals: to ease the model instantiation (by means of a set of auto completion rules), to verify the consistency of model instantiations, and to automatically recognize device classes starting from device functional descriptions.

However, building elements information such as measurements or insulation is not described in DogOnt. Observations made by sensing devices which are essential for a KDD process in the energy efficiency context, are not covered either.

2.3.3. BOT Ontology

Building Ontology Topology (BOT)[4] is a small ontology only covering core concepts of a building [61]. Proliferation of building domain ontologies raises interoperability issues unless appropriate ontology mappings are explicitly specified. Therefore, a first design principle for the design of BOT has been to keep a light schema that could promote its reuse as a central ontology in the Architecture, Engineering and Construction (AEC) domain. BOT covers the description of buildings, composed of storeys which have spaces that can contain and be bounded by building elements. These basic concepts and properties can be extended with concepts and properties from other ontologies covering the building domain, in such a manner that BOT serves as a shared vocabulary. Moreover, the W3C LBD (Linked Building Data) community group[5] is

---

[2]http://ifcowl.openbimstandards.org/IFC4_ADD2.owl

[3]http://elite.polito.it/ontologies/dogont.owl
[4]https://w3id.org/bot
[5]https://www.w3.org/community/lbd/

aiming to produce product ontologies that will extend from the *bot:Element* concept towards more specific building elements.

2.3.4. SSN Ontology

The Semantic Sensor Network (SSN) ontology[6] was developed by the W3C Semantic Sensor Networks Incubator Group (SSN-XG) and can describe sensors, accuracy and capabilities of such sensors, observations and methods used for sensing [13]. Also concepts for operating and survival ranges are included, as these are often part of a given specification of a sensor, along with its performance within those ranges. Finally, a structure for field deployment is included to describe deployment lifetime and sensing purpose of the deployed instruments. As part of the new SSN ontology, the scope is extended to actuation and sampling.

The initial SSN ontology was aligned with DOLCE ultra-lite (DUL) ontology[7] and built around a central Ontology Design Pattern (ODP) called Stimulus-Sensor-Observation (SSO) pattern, describing the relationships between sensors, stimulus, and observations.

The new SSN ontology follows a horizontal and vertical modularization architecture by including a lightweight but self-contained core ontology called SOSA[8] (Sensor, Observation, Sample, and Actuator) for its elementary classes and properties. In line with the changes implemented for the new SSN ontology, SOSA also drops the direct DUL alignment although an optional alignment can be achieved via the SSN-DUL alignment. Furthermore, similar to the original SSO pattern, SOSA acts as a central building block for the new SSN ontology but puts more emphasis on light-weight use and the ability to be used standalone.

The SSN ontology does not contain properties which can be measured by sensors. Neither is covered related material such as units of measurements of these properties, locations or hierarchies of sensor types, or time-related concepts. All this knowledge has to be modelled or imported from other existing vocabularies.

2.3.5. SAREF Ontology

The Smart Appliances REFerence (SAREF) ontology[9] is a shared model of consensus that facilitates the matching of existing assets in the smart appliances domain [16]. The ontology is based on the fundamen-

tal principles of reuse and alignment of concepts and it also provides building blocks that allow separation and recombination of different parts of the ontology depending on specific needs.

SAREF enables modelling devices and sensors in terms of functions, states and services they provide. Nevertheless, the ontology does not address the description of the observation in an interoperable manner to ease further tasks such as reasoning. It provides the link to the FIEMSER[10] data model covering building-related concepts but this knowledge is not enough to describe building elements and their features.

SAREF4BLDG ontology[11] presents an extension of SAREF for the building domain based on the IFC standard. It is limited to the description of devices and appliances within the building domain, so building elements and their features are not covered. However new classes such as buildings, spaces and the physical objects are described.

2.3.6. FIESTA-IoT Ontology

FIESTA-IoT Ontology[12] aims to achieve semantic interoperability among heterogeneous test beds [3]. Ontology reusing and ontology mapping methodologies guided the design of this ontology. Ontologies and taxonomies, such as SSN ontology, M3-lite ontology[13] (a lite version of M3 ontology), Basic Geo WGS84 vocabulary[14], IoT-lite ontology[15], OWL-Time ontology[16], and DUL ontology have been reused to build FIESTA-IoT.

Despite sensing devices are deeply described and covered, tagging and actuating devices are not at the same level. Furthermore, even though the smart building domain is described, building elements and its features are not.

2.3.7. IoT-O Ontology

IoT-O ontology[17] is a core-domain modular IoT ontology proposing a vocabulary to describe connected devices and their relation with their environment [69]. It is intended to model knowledge about IoT systems and to be extended with application specific knowledge. It has been designed in separated modules to fa-

---

[6]https://www.w3.org/ns/ssn

[7]http://www.ontologydesignpatterns.org/ont/dul/DUL.owl

[8]https://www.w3.org/ns/sosa/

[9]http://ontology.tno.nl/saref.owl

[10]https://sites.google.com/site/smartappliancesproject/ontologies/fiemser.ttl

[11]https://w3id.org/def/saref4bldg

[12]http://ontology.fiesta-iot.eu/ontologyDocs/fiesta-iot.owl

[13]http://ontology.fiesta-iot.eu/ontologyDocs/m3-lite.owl

[14]http://www.w3.org/2003/01/geo/wgs84_pos#

[15]http://purl.oclc.org/NET/UNIS/fiware/iot-lite#

[16]http://www.w3.org/2006/time#

[17]http://homepages.laas.fr/nseydoux/ontologies/IoT-O.owl

cilitate its reuse and/or extension. It consists of five different modules:

- A sensing module, based on SSN ontology.
- An acting module, based on SAN (Semantic Actuator Network) ontology[18].
- A service module, based on MSM (Minimal Service Model)[19].
- A lifecycle module, based on a lifecycle vocabulary and an IoT-specific extension.
- An energy module, based on PowerOnt [9].

The building information is described reusing DogOnt concepts, but information regarding building elements or their features is not covered.

2.3.8. SmartHomeWeather Ontology

Smart Home Weather[20] is an OWL ontology that covers both the weather data and the concepts required to perform weather-related tasks within smart homes [70]. Apart from concepts such as weather phenomena and states that can be used to model external climatic condition, this ontology covers near future weather forecasting, making it suitable to use in a smart home scenario.

2.3.9. Discussion

The ontologies presented in this section cover different topics considered by our domain of discourse. Moreover, they overlap to a greater or lesser extent in some of their parts. However, none of them meet all the EEPSA process requirements by themselves. Therefore, we propose to fill the gap between the state of the art ontology offer and the identified requirements of the EEPSA process with the production of a proper ontology that covers the needed terminology, following the good practices of modularity and reuse. The decision for reusing all or parts of any of them in the ontology supporting the EEPSA process, was taken on the basis of a conceptual agreement with the requirements, axiomatic richness relating their terms, simplicity of the structure to facilitate querying, popularity of the ontology to improve interoperability, and documentation accessibility to facilitate new users. Reusing parts from one ontology prevents the reuse of parts of others to avoid redundancy issues. For instance, reusing *bot:Building* and *bot:Element* from BOT, prevents from using their

equivalents *ifc:Building* and *ifc:BuildingElement* from ifcOWL. However, it is essential to explicitly express and maintain those equivalence mappings with related ontologies, as well as some other ontological relationships. For example, *bot:Element rdfs:subClassOf saref4bldg:PhysicalObject* and *bot:hasSpace rdfs:subPropertyOf saref4bldg:haSpace* from SAREF4BLDG. Only parts of some of them will be reused, and therefore a preliminary mapping process will be necessary to interoperate with datasets using the other vocabularies.

The suite of imported modules by the EEPSA ontology[21] includes the tailor made bim4EEPSA module[22], which was devised to describe buildings and their spaces; the SSN ontology to cover sensing and actuating devices; the measurements4EEPSA module[23], which was composed in order to cover measurement related concepts; the OWL-Time ontology to describe time-related concepts; and Basic Geo Vocabulary to represent spatially located things. In the next section, the EEPSA ontology is presented, along with the rationale behind the decisions made for those modules selection and how they were properly customized for covering specific topics of the requirements. Furthermore, the EEPSA process supported by the ontology is detailed.

3. EEPSA in KDD Support

When following the EEPSA process the data analyst utilizes some off-the-shelf tools and others which are specifically designed. For the semantic annotation phase the data analyst counts on an ontology-driven editing framework to manually edit models and also semi-automatic tools to provide annotated data from data repositories, such as platforms to map relational databases to RDF data, or data wrangling tools for more unstructured data. The EEPSA framework will provide domain experts with facilities to design and upload parameterized queries and rules that will be properly stored and later offered to data analysts as pre-defined solutions to different tasks in the aforementioned phases. The analyst interacts freely with the EEPSA framework by accessing and managing data through the incorporated facilities.

---

[18]https://www.irit.fr/recherches/MELODI/ontologies/SAN.owl
[19]http://iserve.kmi.open.ac.uk/ns/msm/msm-2014-09-03.rdf
[20]https://www.auto.tuwien.ac.at/downloads/thinkhome/ontology/WeatherOntology.owl

[21]https://w3id.org/eepsa
[22]https://w3id.org/bim4eepsa
[23]https://w3id.org/measurements4eepsa

Next, the EEPSA ontology is presented. Afterwards, the EEPSA ontology's support in the EEPSA process through the different KDD phases is explained.

3.1. The EEPSA Ontology

Following best practices for ontology design, a set of competency questions were identified in order to establish the ontology requirements. A glossary of terms extracted from those competency questions and their answers were used to look for ontological and non-ontological resources to be considered in the ontology design. In the energy efficiency in buildings domain, there are three main areas of discourse: (i) the space in which the energy efficiency is going to be performed, (ii) the devices deployed in it, and (iii) the data gathered and actuations made by those devices.

Regarding the buildings and spaces area, it must be kept in mind that tertiary buildings have spaces with specific features that may be different from the typically rather small rooms in residential buildings. Therefore, characteristics that are specific for the tertiary buildings have to be covered. Environmental context of the space has to be described, such as the location and orientation. The physical structure and building element properties such as surface area or thermal mass are relevant and need to be described. Tertiary buildings may house many different activities, which need to be described as well as other related concepts like capacity and occupancy rate. Furthermore, the building and spaces area needs to add expert knowledge related with the energy efficiency; for instance, the causality relationships between different environmental conditions. Concerning the sensing and actuating devices deployed within a space, the EEPSA process needs to describe its specifications and functionalities. It needs to describe the type of device too (e.g. motion sensor or window blind actuator), as well as the properties they observe or act on, and contextual information like their location and orientation. Last but not least, the area concerning the data gathered by these devices needs to represent their measurements. Alongside with them, instant of time when these are completed and their values need to be described. Furthermore, a whole coverage of outliers, their potential causes and possible solutions are also required.

Among others, the ontologies presented in the Related work section were assessed and, finally, parts of some of them were reused or re-engineered. The EEPSA ontology has been designed by dividing it in loosely coupled, self-contained components [17],

which facilitates its development and maintenance as well as reuse by imports and controlled extension of parts of the ontology. Figure 2 shows an excerpt of relevant classes and properties.

Since the EEPSA process may be used by non-experts in the building domain, there is a need to describe buildings and spaces in which energy efficiency is aimed in a rather simple way. Looking at the AEC (Architecture, Engineering and Construction) domain ontologies, overlapping concepts can be easily discovered and a neglected integration of them would produce redundancy problems as discussed in Section 2. Therefore, a single vocabulary covering basic building concepts and physical structures in a fairly simple manner was devised. Towards that purpose, top level concepts of the SAREF4BLDG ontology were taken into account, but finally the BOT ontology was preferred. That decision was based on the clean and simple conceptualization of the BOT basic concepts *bot:Building*, *bot:Space*, and *bot:Element*, in addition to a proper alignment with the SAREF4BLDG ontology, explicit links to ifcOWL ontology terms, and a well explained documentation. However, specialized subclassing structure below the *bot:Space* or *bot:Element* classes is not developed and therefore, it was needed to be extended to meet the EEPSA process requirements.

Several ontologies were assessed for the description of spaces. DogOnt ontology targets residential buildings but, although they could resemble tertiary buildings, service, heating and energy demands are different [73]. Furthermore, tertiary buildings are considerably more heterogeneous, encompassing hospitals, schools, restaurants and lodgings [72]. Therefore the EEPSA ontology needs to offer more generic spaces than *dogont:Bedroom* or *dogont:LivingRoom*. Moreover, coverage of building elements in the DogOnt ontology is not as broad as needed for the EEPSA process, even though entire buildings can be represented by extending it through subclassing of *dogont:BuildingEnvironment* and through the definition of proper relationships [8].

IfcOWL represents the IFC open standard for building and construction data. It is mainly designed for the construction industry and, as a result, it is not well suited to space modelling as needed by the EEPSA process. However, ifcOWL presents a comprehensive collection of property sets (known as PSETs) for describing building, spaces and building elements features. Following the semantic transformations proposed in [19], some of those properties (for instance,
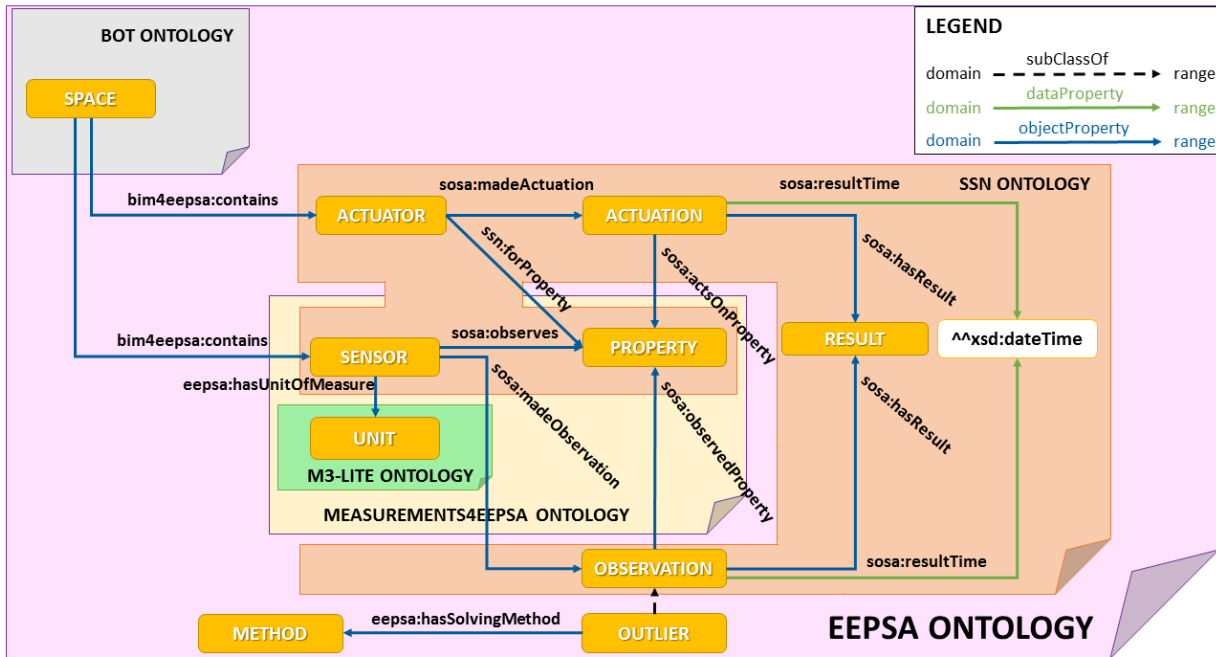
Fig. 2. An overview of relevant classes and properties in EEPSA ontology.

PSET Building Common) were re-engineered and used by the EEPSA ontology to describe specific spaces such as those located at an underground storey (*eepsa:BelowGroundLevelSpace*). This re-engineering method provides domain experts a flexible procedure for extending the EEPSA ontology. Moreover, this method improves interoperability since parts of if-cOWL models could be automatically translated to EEPSA models applying the simplification processes explained in [57].

It was already noted in section 2 that the W3C LBD community group is aiming to produce PROD-UCT ontologies that extend from *bot:Element* class towards more specific element classes, but those ontologies are not available at the time of writing this article. Therefore, in order to cover building structures, BOT classes were extended with some other generic classes. For instance, *bim4eepsa:Door*, *bim4eepsa:Wall*, and *bim4eepsa:Window* were defined as subclasses of *bot:Element*. Furthermore, *bim4eepsa:WeatherStation* was defined as subclass of *bot:Building*. All those axioms were gathered in a module named bim4EEPSA shown in Figure 3, which is imported into the EEPSA ontology. Notice that this modular design allows to easily change this building-related hierarchy replacing the imported module.

Furthermore, the EEPSA ontology encodes expert knowledge that represents causality relationships among different variables, and also includes the definition of queries and rules for the EEPSA process. For instance, the object property *eepsa:isAffectedBy* relates spaces to climatic variables that affect their environmental conditions. An individual of class *eepsa: NaturallyEnlightenedSpace* (a space containing a skylight or an external window, defined in Listing 1, in the Appendix A) will have its indoor temperature affected by the variable *m3-lite:SolarRadiation*, while this same variable will have nearly no effect in an individual of class *eepsa:BelowGroundLevelSpace*.

Regarding sensing and actuating devices deployed in buildings, the latest SSN ontology was selected due to its well founded design and careful documentation, in addition to its wide recognition. For instance, sensors are described with *sosa:Sensor* and actuators with *sosa:Actuator*. Since the SSN ontology does not cover types of sensors or actuators, observable or actuatable properties, units of measurements or orientation of devices, the EEPSA ontology imports the module measurements4EEPSA. This module is composed of a set of subclasses of *sosa:Sensor*, *ssn:Property* and *qudt:Unit* (and their corresponding properties among others) from the M3-Lite ontology. The Local-
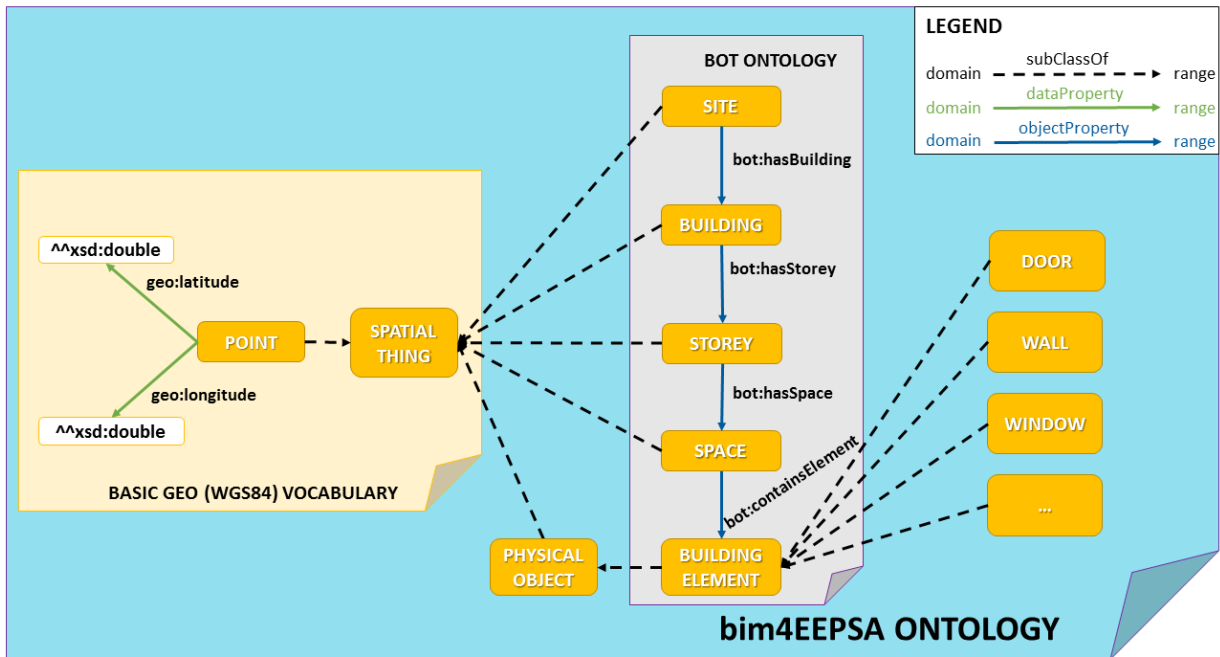
Fig. 3. An overview of the bim4EEPSA ontology's main classes and properties.

ity Module Extractor[24] tool [14] was used for automatically extracting proper subclasses to be reused. Some of those classes were extended with additional ones to improve coverage such as observable properties *m4eepsa:SpaceOccupancy* and *m4eepsa:WaterFlow* or object orientations like *m4eepsa:EastOrientation*. Furthermore, the EEPSA ontology introduces the property *eepsa:hasDataSource* to link properties to data sources where those observable properties can be retrieved from when they are not measured by sensors.

Concerning measurements and actuations made by devices, sensing device measurements are represented as individuals of *sosa:Observation* and actuations made by actuating devices as individuals of *sosa:Actuation*, and so reusing terms from the SSN ontology. Time instants when their actions are made are represented with data property *sosa:resultTime*, whereas their value is represented with the *sosa:Result* class. In the EEPSA ontology a class *eepsa:Outlier* is defined as subclass of *sosa:Observation* in order to represent observations that do not conform to the expected behaviour. A hierarchy of outlier types are defined as subclasses, classifying outliers according to their potential cause. These subclasses will be popu-

lated with outliers detected in the Preprocessing phase of the EEPSA process, such as those caused by the rain (*eepsa:OutlierCausedByRain*) or by a device malfunction (*eepsa:OutlierCausedByDeviceError*). Outliers can occur for various reasons and understanding them might help determining what action to perform. Factors that may affect sensors are represented with the property *eepsa:susceptibleToOutliersCausedBy*. Furthermore, each outlier type class is linked to a proposed method to offset the problem, by means of property *eepsa:hasSolvingMethod*. For example, a temperature outlier caused by a sensing device heated by direct sunlight (*eepsa:OutlierCausedBySolarRadiation*) is linked to two recommended solution methods: *eepsa:DeviceRelocation*, which recommends to relocate the device to an adequate place where it is not exposed to direct sun and *eepsa:DeviceShelter*, recommending to shield the device with a Stevenson Screen or a similar instrument to cover it from direct heat radiation. Following any of these advices should prevent the device from getting heated by direct sunlight and consequently, from measuring erroneous observations.

This EEPSA Ontology provides the necessary conceptual terminology and support for all the KDD steps as detailed in the next sections.

---

[24]https://www.cs.ox.ac.uk/isg/tools/ModuleExtractor/

3.2. Semantic Annotation

A preliminary phase to a KDD process assisted by SWTs consists in annotating data with terms selected from appropriate ontologies and thus providing them with semantics. In the EEPSA process context, semantic annotation of data means to construct an RDF model of the data, giving identifying URIs to resources and inter-relating them using ontology terms. When linking or mapping raw data to existing ontologies or vocabularies a better representation of data is achieved, structuring it and setting formal types and relations among them. Data integration is also achieved [52], and additional background knowledge can be added to the dataset. Furthermore, the resulting dataset improves semantic interoperability [54], providing both human and machines with a shared meaning of terms. This increases the dataset value and the potential to improve the upcoming KDD phases. In addition to the aforementioned integration and interoperability advantages, the resulting data is more domain-oriented than the original source, and makes the solution more application-independent. Consequently, after the Semantic Annotation phase, there is no need for the data analyst to be aware of the structure of the underlying raw data.

The semantic annotation task can be performed by manually editing an RDF model with the help of an adapted graphical user interface (GUI) or a data wrangling tool, or else with a properly programmed automatic middleware. In this phase, all data regarding the building, space and its features, sensing and actuating devices, and their corresponding measurements/actuations are semantically annotated with the selected terms from their corresponding domain ontologies gathered in the EEPSA ontology. Note that the EEPSA ontology, which is the main contribution of this paper to this phase, is designed to favour the reuse of well-known ontologies and therefore facilitates the eventual transformation of models annotated with terms of diverse ontologies to models annotated with the EEPSA ontology. Whether the annotated data is stored natively as RDF or viewed as RDF via middleware, SPARQL queries will be later used to access data across diverse data sources.

Summarizing, after semantically annotating data based on terms contained in the EEPSA ontology, data integration, interoperability and independence from original source are improved. Moreover, this semantic annotation enables the upcoming EEPSA process phases towards the goal of improving the energy efficiency.

3.3. Data Selection

This is the first phase of a typical KDD process. Relevant datasets and subsets of variables that will form the data input for machine learning algorithms are selected. To that end, the data analyst has to understand the data itself: which is the knowledge captured in it, and which is the additional knowledge that can be extracted from it. However, this step is often not trivial and in most cases, domain-specific knowledge is needed to successfully complete it.

Existing work focuses on the use of tools and approaches to visualize and explore LOD to understand data [15]. However, no relevant work that supports the data analyst in the data selection phase has been spotted. In the EEPSA process, SWT are used to support the data analyst choosing the most relevant datasets and variables related with the energy efficiency problem at hand.

Once the target building space is semantically annotated (Semantic Annotation phase) and thanks to the knowledge captured in the form of OWL axioms in the EEPSA ontology, a reasoner classifies the space into one or several space types, and moreover infers that it might be affected by some specific variables (which in the EEPSA ontology are represented with subclasses of *ssn:Property*). For example, a space with windows towards the outside, is a naturally enlightened space (*epsa:NaturallyEnlightenedSpace*) and due to the axioms:

```
NaturallyEnlightmentSpace SubClassOf
(isHighlyAffectedBy value 'Cloud Cover Quantity
Kind') and (isHighlyAffectedBy value 'Solar
Radiation Measurement, PAR Measurement
(Photosynthetically Active Radiation)')
and (isHighlyAffectedBy value 'Sun Position Direction')
and (isHighlyAffectedBy value 'Sun Position Elevation')
```

the reasoner infers that the space's indoor temperature may be affected by variables such as sun radiation and sun position elevation, among others. Consequently, in the EEPSA process' Data Selection phase, the data analyst will get to know, in an automatic way, which variables might be relevant for the target space even though not being an expert in the domain.

After having suggested which variables are the most relevant ones for the task at hand, the data analyst needs to know which of them are being collected by the devices or other mechanisms deployed on the space and which are not. This can be obtained by instan-

tiating and running a parameterized and pre-defined SPARQL query (see Listing 2, in the Appendix A) available in the EEPSA framework over the semantically annotated data.

Summarizing, the EEPSA process uses OWL inferences to assist the data analyst in classifying the space at hand and suggesting variables affecting it. Furthermore, parameterized SPARQL queries are also provided in order to extract more relevant information (for instance, to know whether those variables are being collected by devices deployed in the space or not). This paper has presented some illustrating examples of reasoning tasks and semantic technology resources (SPARQL queries) to assist data analysts deciding which data may be relevant.

The next phase deals with preprocessing the collected data in order to ensure their quality.

3.4. Preprocessing

Today's real-world datasets are highly susceptible to noisy, missing, and inconsistent data due to their typically big size and their likely origin from multiple, heterogeneous sources [37]. These factors have a direct impact in the data quality and low quality data will lead to low quality mining results. This is why it is important to ensure data quality in KDD processes. There are several data preprocessing techniques to increase data quality (e.g. filtering, outlier detection and missing data treatments), which can consequently improve the accuracy and efficiency of data mining algorithms. Moreover, these techniques are not mutually exclusive and may be applied together.

3.4.1. Outlier Detection

Outliers are data objects that stand out amongst other data objects and do not conform to the expected behaviour in a dataset [45]. In addition, outliers can worsen data quality, complicate the knowledge extraction process and lead to wrong conclusions. The process of finding those data objects in a dataset is known as Outlier Detection and it is an essential task in a wide range of domains including fault detection in safety critical systems, intrusion detection for cyber-security and data monitoring in WSNs (Wireless Sensor Networks). This process has been a widely researched topic for many years and there has been an abundance of work from statistics, geometry, machine learning, database, and data mining communities. There are many outlier detection methods divided into groups such as model-based, distance-based or density-based,

according to their assumptions regarding normal data objects versus outliers. Further information regarding these and other outlier detection methods can be found in [11] and [39].

Outliers can occur for various reasons and understanding their provenance helps to determine what actions to take after detecting them. In some cases the aim might be to isolate the outlier and act on it (e.g. fraud detection in credit cards) while in others, outliers are filtered out to avoid inaccurate results (e.g. data analytics). However, identifying the potential cause of outliers still remains an unsolved challenge in most cases: it is not always straightforward and it may become an arduous task. There are also challenging scenarios where a data object may be considered an outlier in one context (e.g. $40°$C measurement is an outlier for a winter day in the north of Spain), but not an outlier in a different context (e.g. $40°$C measurement is not an outlier for a summer day in the south of Spain). With regards to WSNs, which are essential components to capture building conditions, several factors make them prone to outliers due to their particular requirements, dynamic nature and resource limitations [26]. Apart from these factors, WSNs are also context dependent, so that results obtained after applying conventional techniques might be skewed.

Although being an often studied topic, outlier detection has not received sufficient attention from the Semantic Web Community. In [77] a domain ontology has been used to support the outlier detection based on a statistical method. In [31] segment outliers and unusual events are detected in WSNs combining statistical analysis and domain expert knowledge captured via ontology and semantic inference rules. That approach determines whether the sensor collects suspicious data or not by calculating its similarity with neighbours. To the extent of our knowledge, this proposal is one of the few works where Semantic Technologies have a direct role in outlier detection methods. However, it may not be applicable to isolated nodes where there are no nearby sensors to compare their similarity. Furthermore, the identification of the potential cause of outliers is not tackled in that approach.

We believe that the role of SWT in Outlier Detection tasks could be more important and could have a prominent impact. Not only improving the outlier detection, but most importantly in the assistance of data analysts during this process and spotting the potential cause of outliers. This is why the EEPSA process proposes the SemOD (Semantic Outlier Detec-

tion) Framework [25], which focuses on contributing in these issues.

The SemOD Framework is based on domain and expert knowledge expressed in the EEPSA ontology to identify circumstances that make sensors susceptible to errors. Each of these circumstances has been assigned a method (SemOD Method) in which constraints that describe outliers are generated. These constraints are generated in a (semi)automatic way following purposely defined steps and using a set of facilities, guided by the EEPSA ontology axioms. These resources have been designed by experts in a way that no previous knowledge regarding the domain or semantic technologies are required to take advantage of them. The data analyst is then assisted to make use of these methods to generate a SPARQL query (SemOD Query) which retrieves measurements made under a certain circumstance that makes them presumably outliers.

For example, when exposed to the sun, the glass of a temperature sensor can heat up and reach a much higher temperature than it really is, which is a circumstance for generating outliers. A SemOD Method for detecting outliers caused by this circumstance, firstly offers a constraint pattern describing a sensor's sun exposure times as presented in Listing 3 (in the Appendix A). Then, to fill this constraint pattern, the SemOD Method obtains values asserted in the ontology by means of the SPARQL query pattern shown in Listing 4 (in the Appendix A). This query is parameterized by the wild card $OBJECT, which will be replaced with the corresponding sensor's URI. Then, the instantiated constraints have to replace the wild card $PREVIOUSLY_GENERATED_CONSTRAINTS in the FILTER clause of the predefined SemOD Query pattern shown in Listing 5 (in the Appendix A). These constraints also need to be casted into their corresponding data types. Moreover, the graph where the query is going to be performed needs to be specified in the FROM clause, replacing the $RDF_GRAPH wild card, and $PROPERTY wild card also needs to be specified with the corresponding variable's URI. Finally, the SPARQL query is generated. When executed, it obtains the observations suspected to be outliers and they are asserted as individuals of class *eepsa:OutlierCausedBySolarRadiation*. Therefore, not only are outliers detected, but also they are classified according to their potential cause in their corresponding subclass of *eepsa:Outlier*. Listing 6 (in the Appendix A) shows an excerpt of the SPARQL query (SemOD Query) generated to detect outliers caused by sun radiation. Further details of the SemOD Framework can be found in [25].

3.4.2. Missing Values Imputation

Missing Data or Missing Values are one of the most relevant problems in data quality nowadays. They are common in different domains ranging from medical research [21] to social sciences [2]. Sensors are no exception and usually suffer from missing values caused by several reasons like a communication malfunction [36]. Furthermore, many problems like the introduction of a substantial amount of bias and the complication of handling and analysis of data can arise due to the missing values. One of the most common solutions to handle missing values is the imputation, a process that replaces missing data with substituted values. There are multiple imputation methods and depending on the characteristics of the missing values (e.g. duration of missing values period) some of them may provide better outcomes than others.

We consider that SWT could play an important role in the imputation of missing values. Expert knowledge could be elicited, which would in turn allow the classification of missing values according to their characteristics and assist the data analyst suggesting the most suitable imputation methods [32]. This should be further studied in future work.

In summary, the Preprocessing phase in the EEPSA process provides the data analyst with a framework that facilitates the generation of SPARQL rules to detect outliers within the current dataset and classify them according to their potential cause. OWL inferences are also used to propose methods to solve outliers according to their cause and avoid them in the future. Those measures are expected to ensure data quality, which has an effect on data mining algorithms' performance.

Once the current data is preprocessed and its quality is ensured, the next step in the KDD process is the Transformation phase.

3.5. Transformation

In this stage, a projection of the data is produced in a form that data mining algorithms can accept as input. Amongst all the possible tasks in the Transformation phase (e.g. feature extraction), the EEPSA process focuses on the feature generation task.

The vast majority of existing feature generation solutions such as [12], [56] and [43] choose a general knowledge base like DBpedia or YAGO to obtain

property values about the mapped entities and generate new attributes. This approach is considered to only partially exploit SWT capabilities, therefore other alternatives are proposed: the generation of new features from domain-specific knowledge bases and the inference of new features based on existing data.

For cases where a concrete variable is not being collected in the target space, captured knowledge in the EEPSA ontology lets the data analyst know which alternative data sources are available for that variable. For example, a space with bad insulation (*eepsa:BadInsulatedSpace*) might be affected by outdoor humidity among other variables. If there is no sensing device observing it (which can be determined with the SPARQL in Listing 2, in the Appendix A), a reasoner infers that relevant data values for that variable can be retrieved from a nearby weather station.

Nowadays, with the advent of (Linked) Open Data, there are many trustworthy third-party repositories containing valuable information. In the energy efficiency in buildings scenario, where it has been proved that external meteorology affects the energy consumption, weather services enable the possibility of increasing datasets value with specific knowledge. In most cases, weather services information may be accessible in Open Data repositories, but they are rarely offered in RDF Stores. Therefore, there is a need to develop a process to that end. Since weather stations' data may have heterogeneous structures depending on the agency they are controlled by, it is infeasible to propose a generic process applicable to all of them. As a starting point, an ETL (Extract, Load, Transform) process has been defined for weather stations regulated by Euskalmet (Basque Meteorology Agency) and the observations they measure. This process extracts data from Open Data Euskadi (the Basque Open Data portal), annotates them semantically based on the EEPSA ontology using the JENA framework[25], and makes them publicly available[26] in a Virtuoso Open Source version 07.20.3217 Server[27]. The data analyst may have access to this data via SPARQL queries to generate the new meteorological variables needed. A similar ETL process is expected to be developed for weather stations controlled by AEMET (Spanish Meteorological Agency), which extend beyond the Basque Country to the whole Spanish territory.

However, there are variables that cannot be obtained from third party data sources. For some of those cases, an alternative is expected to be offered as part of a future work. For example, indoor illuminance approximate values for sensing devices located in spaces with windows next to the outside (*eepsa:NaturallyEnlightenedSpace*) can be derived from the sky's cloud cover, sun elevation and direction information. Expert knowledge is expected to be modelled in the form of rules so that, depending on the values of the cloud cover and sun position a reasoner can infer the approximate illuminance value for the sensing device. For example, when there were no clouds and the sun were in a particular point (i.e. a point where its light hit the sensing device through the window), the rule would determine a higher illuminance value than at night (when there were no sun).

The proposed feature generation task has to be performed as many times as demanded by the number of variables to generate. The goal is to get the variables previously suggested in the Data Selection phase towards the improvement of the upcoming Data Mining phase. Retrieved or inferred data is considered to have a minimum quality, so preprocessing tasks should not be necessary afterwards.

Summarizing, the current EEPSA process uses OWL inferences to identify sources of information where certain variables can be retrieved from.

3.6. Data Mining

This is the phase where intelligent methods such as machine learning algorithms are applied to extract knowledge. Data analysts will try to make the best predictions to achieve energy efficiency in the target space. For that purpose, data enhanced in previous phases has to be retrieved and integrated in the data analysis environment, mainly by means of SPARQL queries.

3.7. Interpretation

Interpreting results obtained from the data mining phase is not always a straightforward task. Many times, even being an expert in the domain is not enough to understand the results. If underlying semantics of data is not correctly interpreted, results may not be as precise and consistent as they can be [46].

In [18] and [71] Linked Open Data has been proposed as a source of additional information to support the interpretation of the data mining method re-

---

[25]http://jena.apache.org/

[26]All data has been provided by Open Data Euskadi and Euskalmet.

[27]http://193.144.237.227:8890/sparql

sults. However, an effective decision-making must result from reasoning and analysis of knowledge, and must also take into account the experience and expertise of decision-makers. The EEPSA ontology is intended to be extended with this knowledge in further stages of the research, in order to contribute in the Interpretation phase. In any case, thanks to the Semantic Annotation phase, data is enriched so that additional information about the domain can be brought, which contributes to an easier and more effective results interpretation.

4. Experiments and Results

The feasibility of the EEPSA process is tested in the IK4-TEKNIKER building, a technological centre constituted as a not-for-profit foundation located in Eibar (Basque Country, Spain). The scenario on which the EEPSA process is applied to is the second floor of this building (from now on referred to as Open Space) shown in Figure 4. It is a single large room without walls that acts as an office where over 200 people work on a daily basis. As regards the usual work schedule, Monday to Thursday is split-shift and Fridays have reduced working hours.

A service is needed for suggesting the facility manager when HVAC systems have to be activated in the Open Space in order to reach a minimum comfort temperature of 23°C at 08:00 a.m. (when the workday starts). The HVAC control strategy needs to be efficient from an energy expense point of view too. The EEPSA process is applied to meet the facility manager's requirements.

The Open Space is equipped with sensing devices developed in the European FP-7 Tibucon project[28] that observe temperature, humidity and illuminance at five minutes intervals. There are three Tibucon devices located indoors and one located outdoors[29]. The Open Space is also equipped with eight AHUs[30] (Air Handling Units) and collected information is simplified to whether the HVAC system is activated or not.

A baseline model has been developed without the support of the EEPSA process. This baseline model's

results are compared with those obtained after applying the EEPSA process (see Section 4.3), to observe if they have improved and to what extent. Data spanning six months from 31st January 2016 to 1st August 2016 was sampled hourly. Around 20% of data in this period was not measured due to external problems and in many circumstances, temperature sensing devices measured unlikely high temperature values.

The following section details the application of the EEPSA process in the Open Space.

4.1. The EEPSA on the Loop

The first phase of the EEPSA process is the Semantic Annotation phase. As previously stated, in an energy efficiency in buildings problem, there are three main information sources to be annotated: (i) the space in which the energy efficiency is going to be performed, (ii) the devices deployed in it, and (iii) the information gathered by those devices.

In order to represent the Open Space, first of all an individual of class *bot:Building* was created to represent the IK4-TEKNIKER building (*eepsa:ik4tekniker*) in which it is contained. Then, the *eepsa:floor2* was created as an individual of class *bot:Storey*, and related with the building by means of the property *bot:hasStorey*. The individual *eepsa:openSpace* belonging to class *bot:Space* is related with *eepsa:floor2* by the property *bot:hasSpace*. Building elements of the Open Space are represented with individuals of classes such as *bim4eepsa:Door* or *bim4eepsa:Window* and are lined by the property *bot:containsElement*. Sensors and actuators within the Open Space (including the Tibucon sensing device located outdoors) are represented with *sosa:Sensor* and *sosa:Actuator* classes. A simplified RDF representation of the Open Space[31] is available at Listing 7 in the Appendix A.

All data regarding deployed devices and their gathered observations are stored in a PostgreSQL Database. In order to semantically annotate this data with the EEPSA ontology, the Ontop tool[32] is used. Ontop is an OBDA (Ontology-Based Data Access) tool which enables mappings between relational DB and an ontology [10]. It also enables to build a semantic layer, so that data can be queried with the SPARQL language while staying available as relational DB. Mappings can be implemented using the Ontop Protégé plugin. Nev-

---

[28]http://www.tibucon.eu/

[29]A sample of data gathered by Tibucon devices is available at http://193.144.237.227:8890/DAV/home/dba/DataSample.csv

[30]Air Handling Unit is an HVAC system component used to regulate and circulate air. There may be more than one AHUs associated to a single HVAC system, usually in charge of conditioning a specific space or zone.

[31]The representation of the Open Space is not contained in the EEPSA ontology, as it is an instance of a Building Space.

[32]http://ontop.inf.unibz.it/

Fig. 4. IK4-TEKNIKER building's Open Space.

Table 1

Closest Euskalmet weather stations to IK4-TEKNIKER building measuring outdoor temperature (results obtained after executing SPARQL query shown in Listing 8 the 20/07/2017).

stationID	stationName	distanceToBuilding
"C075"	"Eitzaga"	5.86976
"C0D3"	"Aixola (Embalse)"	6.91178
"C078"	"Altzola (Deba)"	8.17392
"C0BE"	"Berriatua"	13.2363
"C074"	"Elorrio"	13.7465

ertheless, inference capabilities offered by Ontop tool are not enough to meet the EEPSA process' needs. Therefore, RDF assertions derived from mappings are dumped and stored in a Virtuoso server 07.20.3217 version, running on an Ubuntu 14.04 Server. This RDF store is private due to the sensitiveness of data.

Once the Open Space itself, the deployed devices and their observations are semantically annotated, the upcoming phase is the Data Selection phase. In order to make predictions as accurate as possible, variables affecting indoor conditions of the Open Space have to be identified. According to what is inferred[33] from the EEPSA ontology class definitions, the Open Space is an adjacent to the outside

(*eepsa:AdjacentToOutsideSpace*) and naturally enlightened (*eepsa:NaturallyEnlightenedSpace*) space. As a result of the definition of these space subclasses, it is inferred that Open Space's indoor temperature might be affected by the following variables:

– m4eepsa:IndoorRelativeHumidity
– m4eepsa:IndoorTemperature
– m4eepsa:OutdoorRelativeHumidity
– m4eepsa:OutdoorTemperature
– m4eepsa:SpaceOccupancy
– m3-lite:CloudCover (*)
– m3-lite:SolarRadiation (*)
– m3-lite:SunPositionDirection (*)
– m3-lite:SunPositionElevation (*)
– m3-lite:WindSpeed (*)

---

[33]All inferences are made using a HermiT 1.3.8.413 reasoner.

However, after executing the SPARQL query defined in Listing 2 (in the Appendix A), it is concluded that not all of these variables are being observed in the Open Space. Namely, the variables with an asterisk (*) are not being observed. Since not all variables affecting energy consumption in the Open Space are collected, predictions may not be as accurate as they could be. Therefore, upcoming phases of the EEPSA process prepare data towards the improvement of these predictions.

The Preprocessing phase deals with ensuring quality of available data, and the EEPSA process does so with the proposed SemOD Framework. The resulting SPARQL query generated after using the SemOD Framework (shown in Listing 6, in the Appendix A), was applied on the observations gathered in the Open Space. Results (which are further analysed in Section 4.3) showed that the outdoor device suffers from 1,253 outliers. This, together with the missing values the dataset had, was considered as a low quality dataset by the data analysts in charge of the problem. Since low quality data may lead to low quality results, it was decided that the information provided by this device (outdoor temperature of the Open Space) should be retrieved from a higher quality data source. This matter is tackled in the next step.

Within the Transformation phase, the EEPSA process focuses on the feature generation task in order to obtain variables affecting energy consumption of a space. Even though this task is intended for variables that are not currently being measured, it can also be used for variables that are being observed but for certain reason (e.g. inconsistent or noisy data) need to be generated. In the Open Space, as previously stated, the outdoor temperature was considered as a low quality dataset due to its outliers and missing values, so it was decided to generate its values in this phase. Owing to the EEPSA ontology's OWL axioms, a reasoner inferred that the outdoor temperature could be obtained from a weather station.

The first step was to check if there were any weather stations measuring outdoor temperature nearby the Open Space. To do so, a data analyst executed the GeoSPARQL query shown in Listing 8 (in the Appendix A) in the aforementioned Virtuoso SPARQL endpoint containing Euskalmet weather stations information[34]. The execution of this query returned a set of weather stations measuring outdoor temperature,

sorted by proximity to the Open Space, as shown in Table 1. However, it is not compulsory for the data analyst to choose the closest weather station. Other factors than the distance can influence on the election of one or another weather station, for instance the altitude where the sensing device is deployed. This information is also represented and can be queried. After comparing Open Space's outside temperature with temperatures observed by nearby weather stations, it was concluded that Eitzaga was the most suitable one due to its conditions similarity.

Once the data analyst decided which was the weather station chosen to retrieve the data, a parameterized SPARQL query was performed over the same endpoint. This time, the data analyst needed to determine the weather station, the variables and the time span to retrieve the needed information. For the Open Space use case, the SPARQL query was set with the variable outdoor temperature, the weather station Eitzaga and the time span between 31st January 2016 and 1st August 2016. The query returned the outdoor temperature values measured in the Eitzaga between the 31st January 2016 and 1st August 2016.

Looking at the results obtained after applying the SPARQL Query in Listing 2 (in the Appendix A) during the Data Selection phase, it was observed that another variable that was not being collected but affected the Open Space was the Wind Speed. This variable can also be retrieved from a weather station, so the same process as for outdoor temperature was followed.

After repeating this feature generation task as many times as needed, all data was used in the following Data Mining phase. In this case, the RapidMiner Studio 7.1 version was used alongside with the Linked Open Data extension. Within this extension, the operator SPARQL Data Importer was used to query the RDF Store and retrieve the information. The Series extension is also in order to work with time series.

4.2. Experiments

A baseline model was developed without the support of the EEPSA process in the traditional KDD process. Different predictive models were built using different combinations of available variables and fine-tuning the parameters for their window sizes. Best results were obtained with a model built with Rapidminer's Vector Linear Regression algorithm[35] and con-

---

taining a window of 553 features: a window of last 504 hours (21 days) indoor temperature observations, last 24 hours outdoor temperature, last 24 hours HVAC value, and another one for the date time.

For the EEPSA-enabled model, first of all the Semantic Annotation phase was applied. Then, EEPSA data selection suggestions were taken into account and the outlier detection task was applied in observations gathered by devices. Thanks to the generation of new attributes, the available data pool became larger. Variable selection and their window sizes were fine tuned to create a model that accurately predicts Open Space's upcoming 24 indoor temperatures. The most accurate model was built with Rapidminer's Vector Linear Regression containing last 168 hours (7 days) indoor temperatures, last 24 hours observations for outdoor temperature, outdoor humidity, outdoor wind speed and HVAC status, 2 features to describe current space occupancy, and 4 features describing the date (month, hour, day of the week and date time). Table 2 shows the input data used by some of the models created with and without the support of the EEPSA process[36].

4.3. Evaluation and Results Discussion

Performance of the forecasters is characterized by two statistical estimates: the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). Measures based on percentage errors (e.g. Mean Absolute Percentage Error, MAPE) were dismissed because of their disadvantage of being infinite or undefined if data is zero, and having extreme values when close to zero. Therefore, a percentage error makes no sense when measuring the accuracy of temperature forecasts on the Fahrenheit or Celsius scales [41]. Predicted indoor temperatures for the future 24 hours in the Open Space have: a MAE of 0.80°C and a RMSE of 0.99°C for the baseline model, and a MAE of 0.57°C and a RMSE of 0.70°C for the EEPSA-enabled model.

Without a process recommending which variables to use, how to preprocess them or in which sources collect them, improving the baseline model would be an arduous task. Although being an expert in data analysis, being a non-expert in the energy efficiency in tertiary buildings domain would make it even more complicated, resorting to the trial and error approach. Following this trial and error approach, the whole KDD process and model generation would be a costly task in

---

[36]Blank spaces mean that no variable has been used, and var(s) is a contraction for variable(s)

terms of time and effort. This cost is considerably reduced thanks to the assistance provided by the EEPSA process.

Moreover, obtained results show that the model obtained after applying the EEPSA process, reduces the MAE and RMSE by over 28% (0.23°C in MAE and 0.29°C in RMSE), which could yield a more energy-efficient control [78]. However, as stated along the article, the true impact of the EEPSA process should not be solely based on predictions accuracy improvement. Table 3 show the MAE and RMSE obtained after applying different models generated after applying the EEPSA process.

The Data Selection of the EEPSA process suggested the incorporation of some variables such as wind speed and outdoor humidity to build the predictive model. For example, incorporating the suggested wind speed variable in the predictive model (which may have been overlooked by a data analyst not expert in the domain), MAE was reduced by 5%. Therefore, thanks to the EEPSA process, the data analyst gets an assistant to define and create the predictive model. Anyway, it will be the decision of the data analyst whether to incorporate or not the suggested variables.

Thanks to the SemOD Framework applied in the data preprocessing phase, 1,253 anomalous temperature measurements were detected in the data registered by the Tibucon device located outdoors. Apart from labelling all these data objects as outliers, they have also been classified according to their potential provenance (*eepsa:OutlierCausedBySunlight*). This proves that the sensing device located outdoor gets hit by the sun in certain time spans, making its measurements unreliable. Thanks to the knowledge stored in the EEPSA ontology, the two possible solutions to this problem can be inferred: sheltering the device, or relocating it in a place with less direct sunlight exposure. Keeping this in mind, a new device was located in a more adequate place where it is protected from direct solar radiation. Furthermore, replacing the outdoor temperature data provided by the Tibucon sensor (considered to be low quality data) with a higher quality outdoor temperature source (a nearby weather station), MAE can be reduced by 6%, and even by nearly 13% in some specific days (namely in days with reduced working hours).

For the period of available data, a day not following expected work schedule was found. Specifically, the 23rd March 2016 (Wednesday) was a reduced hours workday, when typically it should have been a split shift schedule. This happened because in 2016, Easter

Table 2
Predictive models and the variables used to build them.

Model	Indoor Temperature	Outdoor Temperature	Outdoor Humidity	Wind Speed	HVAC	Occupancy	Date
Baseline	3 Tibucon	1 Tibucon			OpenSpace		1 var
EEPSA #1	3 Tibucon	1 Tibucon	1 Tibucon		OpenSpace	2 vars	4 vars
EEPSA #2	3 Tibucon	Euskalmet	1 Tibucon		OpenSpace	2 vars	4 vars
EEPSA #3	3 Tibucon	1 Tibucon	1 Tibucon	Euskalmet	OpenSpace	2 vars	4 vars
EEPSA #4	3 Tibucon	Euskalmet	1 Tibucon	Euskalmet	OpenSpace	2 vars	4 vars

Table 3
MAE and RMSE obtained with different predictive models enabled by the EEPSA process (best results were obtained with EEPSA #4).

Model	MAE (all days)	RMSE (all days)	MAE (reduced working hour)	RMSE (reduced working hour)
EEPSA #1	0.63°C	0.77°C	0.67°C	1.10°C
EEPSA #2	0.60°C	0.74°C	0.57°C	0.91°C
EEPSA #3	0.61°C	0.74°C	0.64°C	1.02°C
EEPSA #4 (*)	0.57°C	0.70°C	0.56°C	0.85°C

started the 24th March. Comparing the predictions obtained with the baseline model, the EEPSA enabled model reduced MAE by 44% (0.28°C) and RMSE by 45% (0.38°C). As long as more data is available, it will be analysed to which extent the EEPSA enabled model reduces prediction errors in days with atypical work schedule.

5. Conclusions

5.1. Benefits of the EEPSA Process

The EEPSA process leverages of SWT to enhance the KDD process towards the achievement of energy efficiency in tertiary buildings. The data analyst is guided through the different KDD phases in a semi-automatic manner. First of all, data is semantically annotated with terms contained in the EEPSA ontology, which aims to capture all the necessary expert knowledge for the EEPSA process mainly related to buildings, sensing and actuating devices, and their corresponding observations and actuations. This Semantic Annotation phase is fundamental for enriching data, integrating heterogeneous data and representing it in a more domain-oriented way, as well as for enabling the improvement of the upcoming KDD phases. In the data selection phase the data analyst is assisted by means of ontology-driven queries and inferences to decide which might be the most relevant variables for the matter at hand. The preprocessing phase leverages of a framework to detect outliers and propose possible methods to solve them to ensure data quality. The

transformation phase generates additional knowledge in the form of new attributes based on knowledge-driven rules and inferencing capabilities. All these tasks contribute to improve the robustness and performance of machine learning algorithms applied in the data mining phase and it eases the interpretation of the obtained results. Furthermore, the proposed process is expected to be reusable in similar use cases of the same domain due to its high abstraction level.

5.2. Future work

The EEPSA process proposed in this paper contributes to raise awareness of the possibilities of the SWT. However, SWT can be further exploited to improve the EEPSA process, implementing some of the tasks proposed in the article.

Data Selection phase: More expert knowledge elicitation should be performed, in order to define new space classes and variables affecting them, towards a more complete EEPSA process. Furthermore, more IFC PSETs should be re-engineered and captured in the EEPSA ontology.

Preprocessing phase: The EEPSA process mainly focuses on the outlier detection and classification by means of the SemOD Framework. However, current SemOD Framework only supports a SemOD Method, namely for the detection of outliers in temperature sensors caused by solar radiation. The SemOD Framework should be extended with further SemOD Methods (e.g. outliers caused by rain) for different sensor types (e.g. humidity or motion sensors), so that the data analyst could have a wide range of methods to detect

and classify outliers generated in different sensor types and by different causes. Regarding the Missing Values treatment, as explained in section 3.4.2, we believe that SWT could play a role assisting the data analyst by suggesting the most suitable imputation methods (depending on the missing values characteristics such as their length).

Transformation phase: The attribute generation task proposed by the EEPSA process takes leverage of meteorological measurements registered by Euskalmet weather stations. That is, the scope of the solution is limited to the Basque Country. Defining and implementing an ETL process for doing the same thing on AEMET weather stations would extend the applicability of this task to the whole Spanish territory. Furthermore, in section 3.5, another attribute generation method has been proposed, which consists in offering approximate attribute values depending on the context. This proposal should be further studied and implemented in further stages of the research.

Interpretation phase: Although not covered currently by the EEPSA process, the interpretation phase has a big potential for exploiting semantics of data. This is why research on this topic should be conducted.

The EEPSA Ontology: IFC contains a lot of information, which would be interesting for the EEPSA process. For instance, information to reflect the effect of features like materials or building envelope sealing. This information should be captured in the bim4EEPSA module that is imported by the EEPSA ontology. This is thought to enable a greater assistance during the KDD process.

Although not directly related with the SWT but towards the facilitation of the EEPSA process application, interaction with the system could be improved. The EEPSA process is intended to be used by nonexperts in the energy efficiency in buildings domain. If the semantic annotation of the target space has to be done manually, depending on the complexity of the space and the knowledge of the user, it can become a difficult and time-costing task. This task should be facilitated with a GUI where the user could add building elements and features to the space in an intuitive and easy manner.

Finally, in order to test the reusability of the EEPSA process, it is going to be applied in another tertiary building, namely in the Bilbao Exhibition Center (BEC). This building is located in Baracaldo (Basque Country, Spain) and covers an area of 251,055 square meters distributed in six pavilions intended for exhibitions.

References

[1] Radwan E. Abdel-Aal. Hourly temperature forecasting using abductive networks. *Engineering Applications of Artificial Intelligence*, 17(5):543–556, 2004. DOI https://doi.org/10.1016/j.engappai.2004.04.002.

[2] Alan C. Acock. Working with missing values. *Journal of Marriage and family*, 67(4):1012–1028, 2005. DOI https://doi.org/10.1111/j.1741-3737.2005.00191.x.

[3] Rachit Agarwal, David Gómez Fernandez, Tarek Elsaleh, Amelie Gyrard, Jorge Lanza, Luis Sánchez, Nikolaos Georgantas, and Valérie Issarny. Unified IoT ontology to enable interoperability and federation of testbeds. In *3rd IEEE World Forum on Internet of Things, WF-IoT 2016, Reston, VA, USA, December 12-14, 2016*, pages 70–75. IEEE Computer Society, 2016. DOI https://doi.org/10.1109/WF-IoT.2016.7845470.

[4] Songpu Ai, Mohan Lal Kolhe, Lei Jiao, Nils Ulltveit-Moe, and Qi Zhang. Domestic demand predictions considering influence of external environmental parameters. In *13th IEEE International Conference on Industrial Informatics, INDIN 2015, Cambridge, United Kingdom, July 22-24, 2015*, pages 640–644. IEEE, 2015. DOI https://doi.org/10.1109/INDIN.2015.7281810.

[5] Ghislain Atemezing, Oscar Corcho, Daniel Garijo, José Mora, María Poveda-Villalón, Pablo Rozas, Daniel Vila-Suero, and Boris Villazón-Terrazas. Transforming meteorological data into linked data. *Semantic Web*, 4(3):285–290, 2013. DOI https://doi.org/10.3233/SW-120089.

[6] Abraham Bernstein, Foster Provost, and Shawndra Hill. Toward intelligent assistance for a data mining process: An ontology-based approach for cost-sensitive classification. *IEEE Transactions on Knowledge and Data Engineering*, 17(4):503–518, 2005. DOI https://doi.org/10.1109/TKDE.2005.67.

[7] Eva Blomqvist. The use of semantic web technologies for decision support - a survey. *Semantic Web*, 5(3):177–201, 2014. DOI https://doi.org/10.3233/SW-2012-0084.

[8] Dario Bonino and Fulvio Corno. DogOnt - Ontology modeling for intelligent domotic environments. In Amit P. Sheth, Steffen Staab, Mike Dean, Massimo Paolucci, Diana Maynard, Timothy W. Finin, and Krishnaprasad Thirunarayan, editors, *The Semantic Web - ISWC 2008, 7th International Semantic Web Conference, ISWC 2008, Karlsruhe, Germany, October 26-30, 2008. Proceedings*, volume 5318 of *Lecture Notes in Computer Science*, pages 790–803. Springer, 2008. DOI https://doi.org/10.1007/978-3-540-88564-1_51.

[9] Dario Bonino, Fulvio Corno, and Luigi De Russis. PowerOnt: An ontology-based approach for power consumption estimation in smart homes. In Raffaele Giaffreda, Radu-Laurentiu Vieriu, Edna Pásher, Gabriel Bendersky, Antonio J. Jara, Joel J. P. C. Rodrigues, Eliezer Dekel, and Benny Mandler, editors, *Internet of Things. User-Centric IoT - First International Summit, IoT360 2014, Rome, Italy, October 27-28, 2014, Revised Selected Papers, Part I*, volume 150 of *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, pages 3–8. Springer, 2014. DOI https://doi.org/10.1007/978-3-319-19656-5_1.

[10] Diego Calvanese, Benjamin Cogrel, Sarah Komla-Ebri, Roman Kontchakov, Davide Lanti, Martin Rezk, Mariano Rodriguez-Muro, and Guohui Xiao. Ontop: Answering SPARQL queries over relational databases. *Semantic Web*, 8(3):471–487, 2016. DOI https://doi.org/10.3233/SW-160217.

[11] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Computing Surveys*, 41 (3):15:1–15:58, 2009. DOI https://doi.org/10.1145/1541880.1541882.

[12] Weiwei Cheng, Gjergji Kasneci, Thore Graepel, David H. Stern, and Ralf Herbrich. Automated feature generation from structured knowledge. In Craig Macdonald, Iadh Ounis, and Ian Ruthven, editors, *Proceedings of the 20th ACM Conference on Information and Knowledge Management, CIKM 2011, Glasgow, United Kingdom, October 24-28, 2011*, pages 1395–1404. ACM, 2011. DOI https://doi.org/10.1145/2063576.2063779.

[13] Michael Compton, Payam M. Barnaghi, Luis Bermudez, Raul Garcia-Castro, Óscar Corcho, Simon J. D. Cox, John Graybeal, Manfred Hauswirth, Cory A. Henson, Arthur Herzog, Vincent A. Huang, Krzysztof Janowicz, W. David Kelsey, Danh Le Phuoc, Laurent Lefort, Myriam Leggieri, Holger Neuhaus, Andriy Nikolov, Kevin R. Page, Alexandre Passant, Amit P. Sheth, and Kerry Taylor. The SSN ontology of the W3C semantic sensor network incubator group. *Journal of Web Semantics*, 17:25–32, 2012. DOI https://doi.org/10.1016/j.websem.2012.05.003. URL https://doi.org/10.1016/j.websem.2012.05.003.

[14] Bernardo Cuenca Grau, Ian Horrocks, Yevgeny Kazakov, and Ulrike Sattler. Modular reuse of ontologies: Theory and practice. *Journal of Artificial Intelligence Research*, 31:273–318, 2008. DOI https://doi.org/10.1613/jair.2375.

[15] Aba-Sah Dadzie and Matthew Rowe. Approaches to visualising linked data: A survey. *Semantic Web*, 2(2):89–124, 2011. DOI https://doi.org/10.3233/SW-2011-0037.

[16] Laura Daniele, Frank T. H. den Hartog, and Jasper Roes. Created in close interaction with the industry: The Smart Appliances REFerence (SAREF) ontology. In Roberta Cuel and Robert Young, editors, *Formal Ontologies Meet Industry - 7th International Workshop, FOMI 2015, Berlin, Germany, August 5, 2015, Proceedings*, volume 225 of *Lecture Notes in Business Information Processing*, pages 100–112. Springer, 2015. DOI https://doi.org/10.1007/978-3-319-21545-7_9.

[17] Mathieu d'Aquin. Modularizing ontologies. In Mari Carmen Suárez-Figueroa, Asunción Gómez-Pérez, Enrico Motta, and Aldo Gangemi, editors, *Ontology Engineering in a Networked World.*, pages 213–233. Springer, 2012. DOI https://doi.org/10.1007/978-3-642-24794-1_10.

[18] Mathieu d'Aquin and Nicolas Jay. Interpreting data mining results with linked data for learning analytics: motivation, case study and directions. In Dan Suthers and Katrien Verbert, editors, *Third Conference on Learning Analytics and Knowledge, LAK '13, Leuven, Belgium, April 8-12, 2013*, pages 155–164. ACM, 2013. DOI https://doi.org/10.1145/2460296.2460327.

[19] Tarcisio Mendes de Farias, Ana Roxin, and Christophe Nicolle. IfcWoD, semantically adapting IFC model relations into OWL properties. In *Proceedings of the 32nd CIB W78 Conference on Information Technology in Construction, October 27-29, 2015, Eindhoven, The Netherlands*, 2015.

[20] Wassim Derguech, Eanna Bruke, and Edward Curry. An autonomic approach to real-time predictive analytics using open data and internet of things. In *2014 IEEE 11th Intl Conf on Ubiquitous Intelligence and Computing and 2014 IEEE 11th Intl Conf on Autonomic and Trusted Computing and 2014 IEEE 14th Intl Conf on Scalable Computing and Communications and Its Associated Workshops, Bali, Indonesia, December 9-12, 2014*, pages 204–211. IEEE Computer Society, 2014. DOI https://doi.org/10.1109/UIC-ATC-ScalCom.2014.137.

[21] A. Rogier T. Donders, Geert J.M.G. van der Heijden, Theo Stijnen, and Karel G.M. Moons. Review: A gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology*, 59(10):1087–1091, 2006. DOI https://doi.org/10.1016/j.jclinepi.2006.01.014.

[22] Chuck Eastman, Paul Teicholz, Rafael Sacks, and Kathleen Liston. *BIM Handbook: A Guide to Building Information Modeling for Owners, Managers, Designers, Engineers and Contractors,*. John Wiley & Sons, 2nd edition, 2011.

[23] Roger Eastman, Craig Schlenoff, Stephen Balakirsky, and Tsai Hong. A sensor ontology literature review. NISTIR 7908, National Institute of Standards and Technology (NIST), 2013. URL https://doi.org/10.6028/NIST.IR.7908.

[24] Varick L. Erickson and Alberto E. Cerpa. Occupancy based demand response HVAC control strategy. In Antonio G. Ruzzelli, editor, *BuildSys'10, Proceedings of the 2nd ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings, Zurich, Switzerland, November 3-5, 2010*, pages 7–12. ACM, 2010. DOI https://doi.org/10.1145/1878431.1878434.

[25] Iker Esnaola-Gonzalez, Jesús Bermúdez, Izaskun Fernández, Santiago Fernández, and Aitor Arnaiz. Towards a semantic outlier detection framework in wireless sensor networks. In Rinke Hoekstra, Catherine Faron-Zucker, Tassilo Pellegrini, and Victor de Boer, editors, *Proceedings of the 13th International Conference on Semantic Systems, SEMANTICS 2017, Amsterdam, The Netherlands, September 11-14, 2017*, pages 152–159. ACM, 2017. DOI https://doi.org/10.1145/3132218.3132226.

[26] Asmaa Fawzy, Hoda M.O. Mokhtar, and Osman Hegazy. Outliers detection and classification in wireless sensor networks. *Egyptian Informatics Journal*, 14(2):157–164, 2013. DOI https://doi.org/10.1016/j.eij.2013.06.001.

[27] Usama M. Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery in databases. *AI Magazine*, 17(3):37–54, 1996. DOI https://doi.org/10.1609/aimag.v17i3.1230.

[28] Christian Fürber. *Data Quality Management with Semantic Technologies*. Springer, 2015. DOI https://doi.org/10.1007/978-3-658-12225-6.

[29] Christian Fürber and Martin Hepp. Using semantic web resources for data quality management. In Philipp Cimiano and Helena Sofia Pinto, editors, *Knowledge Engineering and Management by the Masses - 17th International Conference, EKAW 2010, Lisbon, Portugal, October 11-15, 2010. Proceedings*, volume 6317 of *Lecture Notes in Computer Science*, pages 211–225. Springer, 2010. DOI https://doi.org/10.1007/978-3-642-16438-5_15.

[30] Christian Fürber and Martin Hepp. Using SPARQL and SPIN for data quality management on the Semantic Web. In Witold Abramowicz and Robert Tolksdorf, editors, *Business Information Systems, 13th International Conference, BIS 2010, Berlin, Germany, May 3-5, 2010. Proceedings*, volume 47 of *Lecture Notes in Business Information Processing*, pages 35–46. Springer, 2010. DOI https://doi.org/10.1007/978-3-642-12814-1_4.

[31] Lianli Gao, Michael Bruenig, and Jane Hunter. Semantic-based detection of segment outliers and unusual events for wireless sensor networks. In *Proceedings of the 18th International Conference on Information Quality, ICIQ 2013, Little Rock, AR, USA, November 7-9, 2013.*, pages 127–144. MIT Information Quality Program, 2013. URL http://arxiv.org/abs/1411.2188.

[32] Unai Garciarena. *An investigation of imputation methods for discrete databases and multi-variate time series*. Master's thesis, University of the Basque Country, 2016. URL http://hdl.handle.net/10810/19052.

[33] Pedro A. González and Jesus M. Zamarreno. Prediction of hourly energy consumption in buildings based on a feedback artificial neural network. *Energy and Buildings*, 37(6):595–601, 2005. DOI https://doi.org/10.1016/j.enbuild.2004.09.006.

[34] Amelie Gyrard, Christian Bonnet, Karima Boudaoud, and Martin Serrano. LOV4IoT: A second life for ontology-based domain knowledge to build semantic web of things applications. In Muhammad Younas, Irfan Awan, and Winston Seah, editors, *4th IEEE International Conference on Future Internet of Things and Cloud, FiCloud 2016, Vienna, Austria, August 22-24, 2016*, pages 254–261. IEEE Computer Society, 2016. DOI https://doi.org/10.1109/FiCloud.2016.44.

[35] Hani Hagras, Ian Packharn, Yann Vanderstockt, Nicholas McNulty, Abhay Vadher, and Faiyaz Doctor. An intelligent agent based approach for energy management in commercial buildings. In *FUZZ-IEEE 2008, IEEE International Conference on Fuzzy Systems, Hong Kong, China, 1-6 June, 2008, Proceedings*, pages 156–162. IEEE, 2008. DOI https://doi.org/10.1109/FUZZY.2008.4630359.

[36] Mihail Halatchev and Le Gruenwald. Estimating missing values in related sensor data streams. In Jayant R. Haritsa and T. M. Vijayaraman, editors, *Advances in Data Management 2005, Proceedings of the Eleventh International Conference on Management of Data, January 6, 7, and 8, 2005, Goa, India*, pages 83–94. Computer Society of India, 2005. URL http://comad2005.persistent.co.in/COMAD2005Proc/pages083-094.pdf.

[37] Jiawei Han, Micheline Kamber, and Jian Pei. *Data Mining: Concepts and Techniques, 3rd edition*. Morgan Kaufmann, 2011. URL http://hanj.cs.illinois.edu/bk3/.

[38] Melanie Hilario, Alexandros Kalousis, Phong Nguyen, and Adam Woznica. A data mining ontology for algorithm selection and meta-mining. In *Proceedings of SOKD-2009: Service-oriented knowledge discovery workshop at ECML/PKDD-2009, Bled, Slovenia, September 7-11, 2009*, pages 76–87, 2009.

[39] Victoria J. Hodge and Jim Austin. A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22(2):85–126, 2004. DOI https://doi.org/10.1023/B:AIRE.0000045502.10941.a9.

[40] Markus Hofmann and Ralf Klinkenberg. *RapidMiner: Data Mining Use Cases and Business Analytics Applications*. Chapman & Hall/CRC, 2013.

[41] Rob J. Hyndman and George Athanasopoulos. *Forecasting: Principles and Practice*. OTexts, 2014.

[42] Krzysztof Janowicz, Frank van Harmelen, James A. Hendler, and Pascal Hitzler. Why the data train needs semantic rails. *AI Magazine*, 36(1):5–14, 2015. URL http://www.aaai.org/ojs/index.php/aimagazine/article/view/2560.

[43] Venkata Narasimha Pavan Kappara, Ryutaro Ichise, and O. P. Vyas. Liddm: A data mining system for linked data. In Christian Bizer, Tom Heath, Tim Berners-Lee, and Michael Hausenblas, editors, *WWW2011 Workshop on Linked Data on the Web, Hyderabad, India, March 29, 2011*, volume 813 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2011. URL http://ceur-ws.org/Vol-813/ldow2011-paper07.pdf.

[44] Maxim Kolchin, Nikolay Klimov, Alexey Andreev, Ivan Shilin, Daniil Garayzuev, Dmitry Mouromtsev, and Danil Zakoldaev. Ontologies for web of things: A pragmatic review. In Pavel Klinov and Dmitry Mouromtsev, editors, *Knowledge Engineering and Semantic Web - 6th International Conference, KESW 2015, Moscow, Russia, September 30 - October 2, 2015, Proceedings*, volume 518 of *Communications in Computer and Information Science*, pages 102–116. Springer, 2015. DOI https://doi.org/10.1007/978-3-319-24543-0_8.

[45] Vijay Kotu and Bala Deshpande. *Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner*. Morgan Kaufmann, 2014.

[46] Freddy Lécué, Robert Tucker, Veli Bicer, Pierpaolo Tommasi, Simone Tallevi-Diotallevi, and Marco Luca Sbodio. Predicting severity of road traffic congestion using semantic web technologies. In Valentina Presutti, Claudia d'Amato, Fabien Gandon, Mathieu d'Aquin, Steffen Staab, and Anna Tordai, editors, *The Semantic Web: Trends and Challenges - 11th International Conference, ESWC 2014, Anissaras, Crete, Greece, May 25-29, 2014. Proceedings*, volume 8465 of *Lecture Notes in Computer Science*, pages 611–627. Springer, 2014. DOI https://doi.org/10.1007/978-3-319-07443-6_41.

[47] Siaw-Teng Liaw, Alireza Rahimi, Pradeep Ray, Jane Taggart, S. Dennis, Simon de Lusignan, B. Jalaludin, A. E. T. Yeo, and Amir Talaei-Khoei. Towards an ontology for data quality in integrated chronic disease management: A realist review of the literature. *International Journal of Medical Informatics*, 82(1):10–24, 2013. DOI https://doi.org/10.1016/j.ijmedinf.2012.10.001.

[48] Jiakang Lu, Tamim I. Sookoor, Vijay Srinivasan, Ge Gao, Brian Holben, John A. Stankovic, Eric Field, and Kamin Whitehouse. The smart thermostat: Using occupancy sensors

to save energy in homes. In Jan Beutel, Deepak Ganesan, and Jack A. Stankovic, editors, *Proceedings of the 8th International Conference on Embedded Networked Sensor Systems, SenSys 2010, Zurich, Switzerland, November 3-5, 2010*, pages 211–224. ACM, 2010. DOI https://doi.org/10.1145/1869983.1870005.

[49] Leandro Madrazo, German Nemirovski, and Álvaro Sicilia. Shared vocabularies to support the creation of energy urban systems models. In Rogelio Segovia and Régis Decorme, editors, *EEBuilding Data Models Energy Efficiency Vocabularies and Ontologies. Proceedings of the 4th Workshop organised by the EEB Data Models Community ICT for Sustainable Places, Nice, France, 9th-11th September, 2013*, pages 130–150. European Commission, DG CONNECT, H5 Smart Cities & Sustainability, 2014. URL http://www.sustainableplaces.eu/wp-content/uploads/2017/01/Proceedings-of-4th-EEB-Data-Models-Community-workshop_ICT-for-Sustainable-Places_2013.pdf.

[50] Claudio Martani, David Lee, Prudence Robinson, Rex Britter, and Carlo Ratti. ENERNET: Studying the dynamic relationship between building occupancy and energy consumption. *Energy and Buildings*, 47:584–591, 2012. DOI https://doi.org/10.1016/j.enbuild.2011.12.037.

[51] Alberto Hernandez Neto and Flávio Augusto Sanzovo Fiorelli. Comparison between detailed model simulation and artificial neural network for forecasting building energy consumption. *Energy and Buildings*, 40(12):2169–2176, 2008. DOI https://doi.org/10.1016/j.enbuild.2008.06.013.

[52] Natalya F. Noy. Semantic integration: A survey of ontology-based approaches. *ACM Sigmod Record*, 33(4):65–70, 2004. DOI https://doi.org/10.1145/1041410.1041421.

[53] Mustafa V. Nural, Michael E. Cotterell, and John A. Miller. Using semantics in predictive big data analytics. In Barbara Carminati and Latifur Khan, editors, *2015 IEEE International Congress on Big Data, New York City, NY, USA, June 27 - July 2, 2015*, pages 254–261. IEEE Computer Society, 2015. DOI https://doi.org/10.1109/BigDataCongress.2015.43.

[54] Leo Obrst. Ontologies for semantically interoperable systems. In *Proceedings of the 2003 ACM CIKM International Conference on Information and Knowledge Management, New Orleans, Louisiana, USA, November 2-8, 2003*, pages 366–369. ACM, 2003. DOI https://doi.org/10.1145/956863.956932.

[55] Frauke Oldewurtel, Alessandra Parisio, Colin N. Jones, Dimitrios Gyalistras, Markus Gwerder, Vanessa Stauch, Beat Lehmann, and Manfred Morari. Use of model predictive control and weather forecasts for energy efficient building climate control. *Energy and Buildings*, 45:15–27, 2012. DOI https://doi.org/10.1016/j.enbuild.2011.09.022.

[56] Heiko Paulheim and Johannes Fürnkranz. Unsupervised generation of data mining features from linked open data. In Dumitru Dan Burdescu, Rajendra Akerkar, and Costin Badica, editors, *2nd International Conference on Web Intelligence, Mining and Semantics, WIMS '12, Craiova, Romania, June 6-8, 2012*, pages 31:1–31:12. ACM, 2012. DOI https://doi.org/10.1145/2254129.2254168.

[57] Pieter Pauwels and Ana Roxin. SimpleBIM: From full IfcOWL graphs to simplified building graphs. In Symeon Christodoulou and Raimar Scherer, editors, *eWork and eBusiness in Architecture, Engineering and Construction: ECPPM 2016: Proceedings of the 11th European Conference on Product and Process Modelling (ECPPM 2016), Limassol, Cyprus, 7-9 September 2016*, pages 11–18. CRC Press, 2017.

[58] Pieter Pauwels and Walter Terkaj. Express to owl for construction industry: Towards a recommendable and usable ifcowl ontology. *Automation in Construction*, 63:100–133, 2016. DOI https://doi.org/10.1016/j.autcon.2015.12.003.

[59] Pieter Pauwels, Sijie Zhang, and Yong-Cheol Lee. Semantic web technologies in AEC industry: A literature overview. *Automation in Construction*, 2016. DOI https://doi.org/10.1016/j.autcon.2016.10.003.

[60] Qudamah K. Quboa and Mohamad Saraee. A state-of-the-art survey on semantic web mining. *Intelligent Information Management*, 5:10–17, 2013. DOI https://doi.org/10.4236/iim.2013.51002.

[61] Mads Holten Rasmussen, Pieter Pauwels, Christian Anker Hviid, and Jan Karlshøj. Proposing a central AEC ontology that allows for domain specific extensions. In *Lean and Computing in Construction Congress (LC3): Volume I Proceedings of the Joint Conference on Computing in Construction (JC3), July 4-7, 2017, Heraklion, Greece*, pages 237–244, 2017. DOI https://doi.org/10.24928/JC3-2017/0153.

[62] Petar Ristoski. Towards linked open data enabled data mining - strategies for feature generation, propositionalization, selection, and consolidation. In Fabien Gandon, Marta Sabou, Harald Sack, Claudia d'Amato, Philippe Cudré-Mauroux, and Antoine Zimmermann, editors, *The Semantic Web. Latest Advances and New Domains - 12th European Semantic Web Conference, ESWC 2015, Portoroz, Slovenia, May 31 - June 4, 2015. Proceedings*, volume 9088 of *Lecture Notes in Computer Science*, pages 772–782. Springer, 2015. DOI https://doi.org/10.1007/978-3-319-18818-8_50.

[63] Petar Ristoski and Heiko Paulheim. Feature selection in hierarchical feature spaces. In Saso Dzeroski, Pance Panov, Dragi Kocev, and Ljupco Todorovski, editors, *Discovery Science - 17th International Conference, DS 2014, Bled, Slovenia, October 8-10, 2014. Proceedings*, volume 8777 of *Lecture Notes in Computer Science*, pages 288–300. Springer, 2014. DOI https://doi.org/10.1007/978-3-319-11812-3_25.

[64] Petar Ristoski and Heiko Paulheim. Semantic web in data mining and knowledge discovery: A comprehensive survey. *Journal of Web Semantics*, 36:1–22, 2016. DOI https://doi.org/10.1016/j.websem.2016.01.001.

[65] Petar Ristoski, Christian Bizer, and Heiko Paulheim. Mining the web of linked data with rapidminer. *Journal of Web Semantics*, 35:142–151, 2015. DOI https://doi.org/10.1016/j.websem.2015.06.004.

[66] James Scott, A. J. Bernheim Brush, John Krumm, Brian Meyers, Mike Hazas, Steve Hodges, and Nicolas Villar. PreHeat: Controlling home heating using occupancy prediction. In James A. Landay, Yuanchun Shi, Donald J. Patterson, Yvonne Rogers, and Xing Xie, editors, *UbiComp 2011: Ubiquitous Computing, 13th International Conference, UbiComp 2011, Beijing, China, September 17-21, 2011, Proceedings*, pages 281–290. ACM, 2011. DOI https://doi.org/10.1145/2030112.2030151.

[67] Tiina Sekki, Miimu Airaksinen, and Arto Saari. Impact of building usage and occupancy on energy consumption in finnish daycare and school buildings. *Energy and Buildings*, 105:247–257, 2015. DOI https://doi.org/10.1016/j.enbuild.2015.07.036.

[68] Floarea Serban, Joaquin Vanschoren, Jörg-Uwe Kietz, and Abraham Bernstein. A survey of intelligent assistants for data

analysis. *ACM Computing Surveys*, 45(3):31:1–31:35, 2013. DOI https://doi.org/10.1145/2480741.2480748.

[69] Nicolas Seydoux, Khalil Drira, Nathalie Hernandez, and Thierry Monteil. IoT-O, a core-domain IoT ontology to represent connected devices networks. In Eva Blomqvist, Paolo Ciancarini, Francesco Poggi, and Fabio Vitali, editors, *Knowledge Engineering and Knowledge Management - 20th International Conference, EKAW 2016, Bologna, Italy, November 19-23, 2016, Proceedings*, volume 10024 of *Lecture Notes in Computer Science*, pages 561–576, 2016. DOI https://doi.org/10.1007/978-3-319-49004-5_36.

[70] Paul Staroch. *A weather ontology for predictive control in smart homes*. Master's thesis, Technische Universität Wien, 2013.

[71] Ilaria Tiddi. *Explaining Data Patterns using Knowledge from the Web of Data*. PhD dissertation, The Open University, 2016.

[72] U.S. Department of Energy. Energy efficiency trends in residential and commercial buildings. Technical report, 2010.

[73] U.S. Energy Information Administration. International energy outlook 2016. Technical Report DOE/EIA-0484(2016), 2016.

[74] Pierre-Yves Vandenbussche, Ghislain Atemezing, María Poveda-Villalón, and Bernard Vatant. Linked open vocabularies (LOV): A gateway to reusable semantic vocabularies on the web. *Semantic Web*, 8(3):437–452, 2017. DOI https://doi.org/10.3233/SW-160213.

[75] Paul Waide and Daniele Gerundino. International standards to develop and promote energy efficiency and renewable energy sources. IEA Information Paper. in support of the G8 plan of action, International Energy Agency, 2007. URL https://webstore.iea.org/international-standards-to-develop-and-promote-energy-efficiency-and-renewable-energy-sources.

[76] Bill B. Wang, Robert I. McKay, Hussein A. Abbass, and Michael Barlow. A comparative study for domain ontology guided feature extraction. In Michael J. Oudshoorn, editor, *Computer Science 2003, Twenty-Sixth Australasian Computer Science Conference (ACSC2003), Adelaide, South Australia, February 2003*, volume 16 of *CRPIT*, pages 69–78. Australian Computer Society, 2003. URL http://crpit.com/confpapers/CRPITV16BWang.pdf.

[77] Yongheng Wang and Shenghong Yang. Outlier detection from massive short documents using domain ontology. In *2010 IEEE International Conference on Intelligent Computing and Intelligent Systems (ICIS)*, volume 3, pages 558–562. IEEE, 2010. DOI https://doi.org/10.1109/ICICISYS.2010.5658426.

[78] Francisco Zamora-Martínez, Pablo Romeu, Paloma Botella-Rocamora, and Juan Pardo. Towards energy efficiency: Forecasting indoor temperature via multivariate analysis. *Energies*, 6(9):4639–4659, 2013. DOI https://doi.org/10.3390/en6094639.

[79] Hai-xiang Zhao and Frédéric Magoulés. A review on the prediction of building energy consumption. *Renewable and Sustainable Energy Reviews*, 16(6):3586–3592, 2012. DOI https://doi.org/10.1016/j.rser.2012.02.049.

Appendix

A. Semantic Resources

```
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix eepsa: <http://w3id.org/eepsa#> .
@prefix bim4eepsa: <http://w3id.org/bim4eepsa#> .
@prefix m3-lite: <http://purl.org/iot/vocab/m3-lite#> .
@prefix bot: <https://w3id.org/bot#> .

###  http://w3id.org/eepsa#NaturallyEnlightenedSpace
eepsa:NaturallyEnlightenedSpace rdf:type owl:Class ;
  owl:equivalentClass
    [ owl:intersectionOf ( bot:Space
        [ rdf:type owl:Class ;
          owl:unionOf
            ( [ rdf:type owl:Restriction ;
                owl:onProperty bot:containsElement ;
                owl:minQualifiedCardinality "1"^^xsd:nonNegativeInteger ;
                owl:onClass bim4eepsa:ExternalWindow ]
             [ rdf:type owl:Restriction ;
                owl:onProperty bot:containsElement ;
                owl:minQualifiedCardinality "1"^^xsd:nonNegativeInteger ;
                owl:onClass bim4eepsa:Skylight ]
            ) ] ) ;
      rdf:type owl:Class ] ;
  rdfs:subClassOf bot:Space ,
     [ owl:intersectionOf
         ( [ rdf:type owl:Restriction ;
             owl:onProperty eepsa:isHighlyAffectedBy ;
             owl:hasValue m3-lite:CloudCover ]
           [ rdf:type owl:Restriction ;
             owl:onProperty eepsa:isHighlyAffectedBy ;
             owl:hasValue m3-lite:SolarRadiation ]
           [ rdf:type owl:Restriction ;
             owl:onProperty eepsa:isHighlyAffectedBy ;
             owl:hasValue m3-lite:SunPositionDirection ]
           [ rdf:type owl:Restriction ;
             owl:onProperty eepsa:isHighlyAffectedBy ;
            owl:hasValue m3-lite:SunPositionElevation ] ) ;
         rdf:type owl:Class ] ;
  rdfs:comment "A space enlightened with a source of light from the exterior
  (and not enlightened by a source of light inside the space)."@en ;
  rdfs:label "Naturally Enlightened Space"@en .
```

Listing 1: eepsa:NaturallyEnlightenedSpace class axiom.

```
PREFIX sosa: <http://www.w3.org/ns/sosa/>
PREFIX eepsa: <http://w3id.org/eepsa#>
PREFIX bim4eepsa: <http://w3id.org/bim4eepsa#>

SELECT DISTINCT ?affectingProperty
WHERE {
  { eepsa:mySpace eepsa:isAffectedBy ?affectingProperty. }
  MINUS  { eepsa:mySpace bim4eepsa:contains ?sensor.
           ?sensor sosa:observes ?affectingProperty. }
}
```

Listing 2: SPARQL query for retrieving properties that affect but are not observed within a space "eepsa:mySpace".

```
(month(?date)) = MONTH_VALUE &&
?time >= (STARTING_TIME) && ?time <= (ENDING_TIME)
```

Listing 3: SemOD Method's constraint pattern describing an object's sun exposure times.

```
SELECT *
WHERE {
        $OBJECT eepsa:hasSunExposurePeriod ?period.
        ?period eepsa:startingTime ?startingTime;
                eepsa:endingTime ?endingTime;
                eepsa:hasMonth ?monthValue.
}
```

Listing 4: SemOD Method's constraint pattern describing $OBJECT's sun exposure times.

```
CONSTRUCT {?obs1 rdf:type eepsa:OutlierCausedBySunRadiation}
FROM <$RDF_GRAPH>
WHERE {
?sensor1 sosa:observedProperty m3-lite:Temperature;
        m3-lite:hasDirection ?orientation.
?sensor2 sosa:observedProperty $PROPERTY;
        eepsa:hasUnitOfMeasure $UNIT_OF_MEASUREMENT;
        m3-lite:hasDirection ?orientation.
?obs1 sosa:isObservedBy ?sensor1;
        eepsa:obsTime ?time;
        eepsa:obsDate ?date;
        sosa:hasSimpleResult ?value1.
?obs2 sosa:isObservedBy ?sensor2;
        eepsa:obsTime ?time;
        eepsa:obsDate ?date;
        sosa: hasSimpleResult ?illuminanceValue.
FILTER( $PREVIOUSLY_GENERATED_CONSTRAINTS  )
```

Listing 5: SemOD Query pattern for detecting outliers caused by sun radiation.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX sosa: <http://www.w3.org/ns/sosa/>
PREFIX eepsa: <http://w3id.org/eepsa#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX m3-lite: <http://purl.org/iot/vocab/m3-lite#>

CONSTRUCT {?obs1 rdf:type eepsa:OutlierCausedBySunRadiation}
FROM <myGraph>
WHERE{
?sensor1 sosa:observedProperty m3-lite:Temperature;
        m3-lite:hasDirection ?orientation.
?sensor2 sosa:observedProperty m3-lite:Illuminance;
        eepsa:hasUnitOfMeasure m3-lite:Lux;
        m3-lite:hasDirection ?orientation.
?obs1 sosa:isObservedBy ?sensor1;
        eepsa:obsTime ?time;
        eepsa:obsDate ?date;
        sosa:hasSimpleResult ?value1.
?obs2 sosa:isObservedBy ?sensor2;
        eepsa:obsTime ?time;
        eepsa:obsDate ?date;
        sosa: hasSimpleResult ?illuminanceValue.
FILTER(
month(?date) = 02 &&?time > xsd:time(18:00:00) && ?time < xsd:time(19:00:00)
            && xsd:integer(?illuminanceValue) > (15000) )
        || (...)
)
}
```

Listing 6: SemOD Query excerpt for detecting temperature outliers caused by sun radiation.

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix eepsa: <http://w3id.org/eepsa#> .
@prefix bim4eepsa: <http://w3id.org/bim4eepsa#> .
@prefix bot: <https://w3id.org/bot#> .
@prefix sosa: <http://www.w3.org/ns/sosa/> .
@prefix m3-lite: <http://purl.org/iot/vocab/m3-lite#> .

eepsa:ik4-tekniker rdf:type owl:NamedIndividual ,
        bot:Building ;
        bot:hasStorey eepsa:floor2 ;
        rdfs:comment "The IK4-TEKNIKER building" .

eepsa:floor2 rdf:type owl:NamedIndividual ,
        bot:Storey ;
        bot:hasSpace eepsa:openSpace ;
        rdfs:comment "The second storey of the IK4-TEKNIKER
```

```
                building" .

eepsa:openSpace rdf:type owl:NamedIndividual ,
            bot:BuildingSpace ;
            bot:containsElement eepsa:door1 ,
                        eepsa:door2 ,
                        eepsa:door3 ,
                        eepsa:wall1 ,
                        eepsa:wall2 ,
                        eepsa:wall3 ,
                        eepsa:window1 ;
            bim4eepsa:contains eepsa:OpenSpaceHVAC ,
                        eepsa:TibuconIndoor1 ,
                        eepsa:TibuconIndoor2 ,
                        eepsa:TibuconIndoor3 ,
                        eepsa:TibuconOutdoor1 ;
            rdfs:comment "Building space located at IK4-TEKNIKER building's
            second floor" .

eepsa:door1 rdf:type owl:NamedIndividual ,
            bim4eepsa:Door .

eepsa:door2 rdf:type owl:NamedIndividual ,
            bim4eepsa:Door .

eepsa:door3 rdf:type owl:NamedIndividual ,
            bim4eepsa:Door .

eepsa:wall1 rdf:type owl:NamedIndividual ,
            bim4eepsa:ExternalBuildingElement ,
            bim4eepsa:Wall ;
            m3-lite:hasDirection m4eepsa:northWestOrientation .

eepsa:wall2 rdf:type owl:NamedIndividual ,
            bim4eepsa:Wall .

eepsa:wall3 rdf:type owl:NamedIndividual ,
            bim4eepsa:Wall .

eepsa:window1 rdf:type owl:NamedIndividual ,
            bim4eepsa:ExternalBuildingElement ,
            bim4eepsa:Window ;
            m3-lite:hasDirection m4eepsa:southWestOrientation .

eepsa:TibuconIndoor1 rdf:type owl:NamedIndividual ,
            sosa:Sensor .

eepsa:TibuconIndoor2 rdf:type owl:NamedIndividual ,
            sosa:Sensor .

eepsa:TibuconIndoor3 rdf:type owl:NamedIndividual ,
```

```
        sosa:Sensor .

eepsa:TibuconOutdoor1 rdf:type owl:NamedIndividual ,
        sosa:Sensor .

eepsa:OpenSpaceHVAC rdf:type owl:NamedIndividual ,
        sosa:Actuator .
```

Listing 7: Excerpt of RDF representation of the Open Space.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX geo: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX ssn: <http://www.w3.org/ns/ssn/>
PREFIX sosa: <http://www.w3.org/ns/sosa/>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX bim4eepsa: <http://w3id.org/bim4eepsa#>
PREFIX m4eepsa: <http://w3id.org/measurements4eepsa#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

SELECT DISTINCT ?stationID ?stationName
(bif:st_distance((bif:st_point(xsd:float(?lat), xsd:float(?lon))),
(bif:st_point(xsd:float(43.19), xsd:float(-2.45)))))) AS ?distanceToBuilding
FROM <http://tekniker.es/euskalmetStations>
WHERE {
?weatherStation rdf:type bim4eepsa:WeatherStation.
?weatherStation foaf:name ?stationName.
?weatherStation geo:latitude ?lat.
?weatherStation geo:longitude ?lon.
?weatherStation dc:identifier ?stationID.
?weatherStation bim4eepsa:contains ?sensor.
?sensor ssn:hasSubSystem ?sensorComponent.
?sensorComponent sosa:observes ?property.

FILTER (
?property = m4eepsa:OutdoorTemperature )
}
ORDER BY ?distanceToBuilding
LIMIT 5
```

Listing 8: GeoSPARQL query for retrieving IK4-TEKNIKER building nearby weather stations measuring temperature.