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 system activation 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 [68]. In order to meet the energy sustainability and minimize the climate change, the European Commission agreed a set of binding legislation inside the EU 2020 package. One of the spotlighted sectors regarding 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 the 36% of CO2 emissions in the EU. Therefore, efficient management of building energy plays a vital role and is becoming the trend for 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 consequence, 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 KDD can be understood as a five steps process leading to the extraction of useful knowledge from raw data [26], 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.

In this paper the EEPSA (Energy Efficiency Prediction Semantic Assistant) process is presented. The EEPSA process takes leverage of the Semantic Web Technologies (SWT) to enhance the KDD process for achieving energy efficiency and comfort in tertiary buildings. For that purpose, expert knowledge in buildings, deployed devices and observations are used. The proposed process assists the data analyst during the different KDD phases to improve the robustness and performance of machine learning algorithms applied in the data mining phase and ease the interpretation of the obtained results.

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. 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 [32], 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. [34] 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 [69]. But not all external weather factors have the same impact in the energy consumption forecasting in buildings. In the use case presented in [50] for instance, effects of humidity and sun radiation had a less significant impact in energy consumption, compared with the external temperature.

Related work in [48], [65] and [78] 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. [23] proposes different HVAC (Heating, Ventilation and Air Conditioning) control strategies based on occupancy prediction of rooms. In a similar way [64] focuses on a better heating scheduling by predicting future occupancy. Wireless motion sensors and door sensors are used in [46] to infer occupants presence and activate or deactivate HVAC systems accordingly. [54] 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. This research proposes the use of SWT towards 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 [40] and [59]. 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 under-

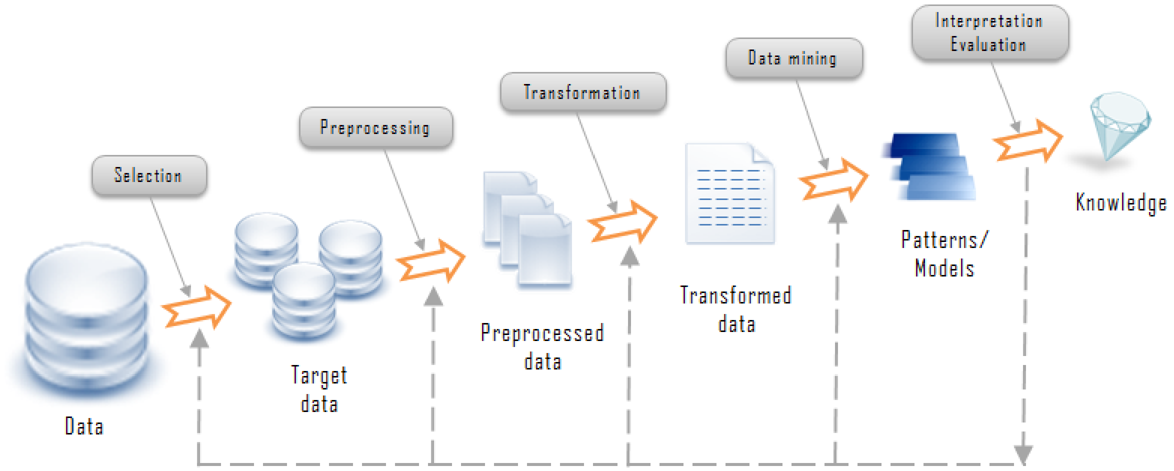

Fig. 1. An overview of the steps that compose the KDD Process proposed in [26].

stand which datasets can be meaningfully integrated, and how to communicate the results to humans and machines alike. This is why making sense of data and gaining new insights works best if inductive and deductive techniques go hand-in-hand.

According to [19], 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 [27], [28] and [29], 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 [45]. In [60] desiderata and challenges for developing a framework for unsupervised generation of data mining features from Linked Data are identified. [49], [55] and [63] are examples of systems for enriching data with features that are derived from LOD (Linked Open Data). In [75] a feature-selection method based on ontology is proposed. The data mining environment RapidMiner [38] includes a LOD extension which provides a set of operators for augmenting existing datasets with additional attributes from open data sources [61]. In [52] 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. [36] presents an ontology to support the meta-learning for algorithm selection in the data mining, while in [5] one of the first intelligent discovery assistants is proposed. An overview of existing intelligent assistants for data analysis is provided in [66]. In [6] 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 [62]. 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.

This research contributes at exploiting SWT to enable an improved KDD process for energy efficiency in tertiary buildings.

2.3. *Existing Ontologies in the Field*

BIM (Building Information Modelling) deals with the representation of functional and physical characteristics of a building [21]. That is, in a BIM model static information of a building element can be queried, such as a door: its material, when it was installed, or even the changes the door received until date. 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 applications of SWT in this field are surveyed in [57]. 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. Other ontologies such as Semanco [47] or the Aemet Network of Ontologies [4] have also been analysed, but are not covered in such a depth. Some of the consulted surveys to identify these ontologies have been [22] and [41]. An interesting comparison between different IoT ontologies is also covered in [67]. The catalogues Linked Open Vocabularies [74] and LOV4IoT [33] have been used to search vocabularies covering desired concepts.

2.3.1. ifcOWL Ontology

IfcOWL ontology[1] 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 [56]. 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 [58]). 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 [18], 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[2] allows to formalize all the aspects of IDEs (Intelligent Domotic Environment) and it is designed with a particular focus on interoperation between domotic systems [7]. 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. SSN Ontology

The Semantic Sensor Network (SSN) ontology[3] 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 [12]. 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[4] 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[5] (Sensor, Observation, Sample, and Actuator) for its elementary classes and properties. In line with the changes implemented for the new SSN ontology,

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

[2] http://elite.polito.it/ontologies/dogont.owl

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

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

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

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.

2.3.4. SAREF Ontology

The Smart Appliances REFerence (SAREF) ontology[6] is a shared model of consensus that facilitates the matching of existing assets in the smart appliances domain [15]. The ontology is based on the fundamental 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[7] data model covering building-related concepts but this knowledge is not enough to describe building elements and their features.

SAREF4BLDG ontology[8] 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.

2.3.5. FIESTA-IoT

FIESTA-IoT Ontology[9] 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[10] (a lite version of M3 ontology), Basic Geo WGS84 vocabulary[11], IoT-lite ontology[12], OWL-Time ontology[13], 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.

2.3.6. IoT-O Ontology

IoT-O ontology[14] is a core-domain modular IoT ontology proposing a vocabulary to describe connected devices and their relation with their environment [67]. 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 facilitate 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[15].
- A service module, based on MSM (Minimal Service Model)[16].
- A lifecycle module, based on a lifecycle vocabulary and an IoT-specific extension.
- An energy module, based on PowerOnt [8].

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

2.3.7. SmartHomeWeather Ontology

Smart Home Weather[17] 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.

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

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

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

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

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

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

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

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

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

[15] https://www.irit.fr/recherches/MELODI/ontologies/SAN.owl

[16] http://iserve.kmi.open.ac.uk/ns/msm/msm-2014-09-03.rdf

[17] https://www.auto.tuwien.ac.at/downloads/thinkhome/ontology/WeatherOntology.owl

2.3.8. Discussion

The ontologies presented in this section overlap to a greater or lesser extent in some of their parts. 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. For instance, reusing *s4bldg:Building* and *s4bldg:BuildingObject* from SAREF4BLDG, or *sosa:Sensor* and *sosa:Actuator* from SSN, prevents from using their equivalents *ifc:Building* and *ifc:Bu ildingElement* from ifcOWL, or *saref:Sensor* from SAREF and *san:Actuator* from SAN. 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 EEPSA ontology is presented in the next section, along with the EEPSA process that it supports.

3. EEPSA in KDD Support

Nowadays data analysts receive no guidance in KDD processes and consequently, novice analysts are typically overwhelmed. They have no idea which variables and tasks can be confidently used, and often resort to trial and error. Furthermore, being a non-expert in the domain further complicates the process to make accurate predictions. Therefore, there is an urgent need to support both expert and novice data analysts during the whole KDD process. The EEPSA process makes use of SWT, such as ontologies, ontology-driven rules and data access as a contribution to overcoming this hurdle in the domain of energy efficiency in tertiary buildings. Therefore, the EEPSA ontology supporting the EEPSA process aims to capture the necessary vocabulary and expert knowledge mainly related to buildings, sensing and actuating devices, and their corresponding observations and actuations.

The EEPSA process targets different KDD phases. First of all, data needs to be semantically annotated with the selected 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 may help in this task. The preprocessing phase intends to clean data from noisy, missing, or outlier values. Ontology-driven rules may help in detecting such 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 these phases contribute to improving the robustness and performance of machine learning algorithms applied in the data mining phase and it eases 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.

Following the EEPSA process the user utilizes some off-the-shelf tools and others which are specifically designed. For the semantic annotation phase the user 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 provides domain experts with facilities to design and upload parameterized queries and rules that will be properly stored and later offered to 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.

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 by those devices. Among others, the ontolo-

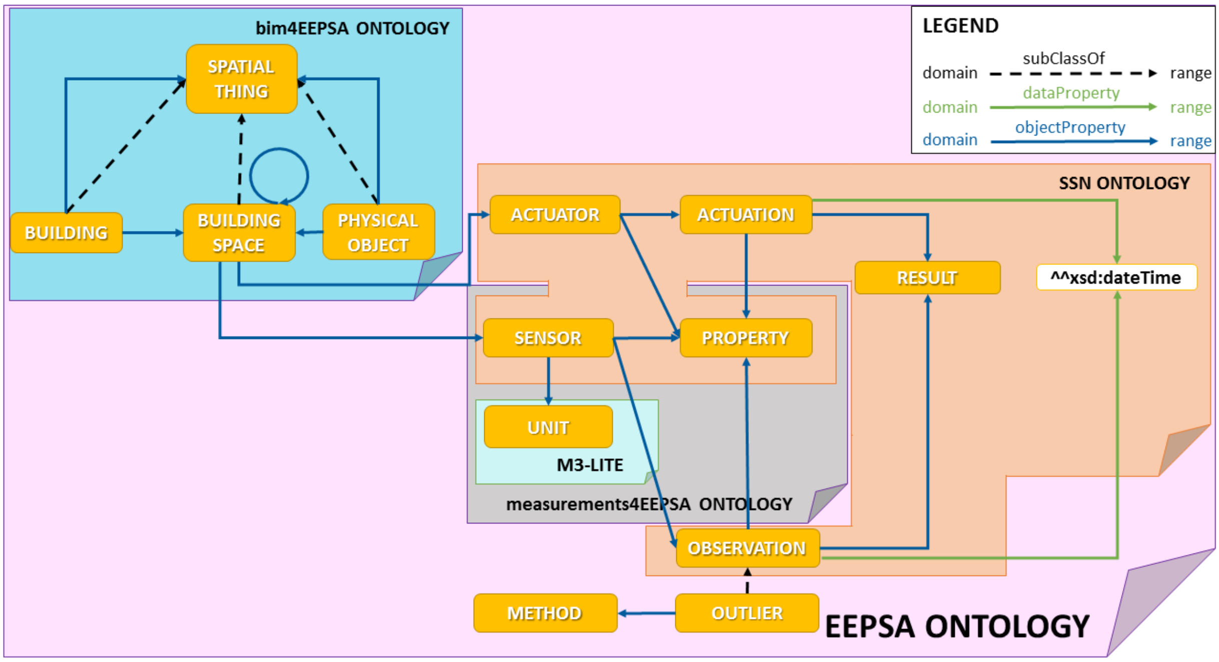

Fig. 2. An overview of the EEPSA ontology's main classes and properties.

gies presented in the previous section were assessed and, finally, parts of some of them were reused or re-engineered.

The EEPSA ontology[18] has been designed by dividing it in loosely coupled, self-contained components [16], 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 the main classes and properties from the EEPSA ontology.

With respect to building spaces representation, the Building Ontology Topology (BOT)[19] was taken into account, but finally the top levels of the hierarchy presented in the SAREF4BLDG ontology were selected. This decision was based on the clean and simple conceptualization of *s4bldg:Building*, *s4bldg:BuildingSpace*, and *s4bldg:PhysicalObject*, in addition to a proper integration with the SAREF ontology, links to ifcOWL ontology, and a well explained documentation. Those top level classes were extended with some other generic classes and properties. For instance, *bim4eepsa:WeatherStation* as subclass of *s4bldg:Building*, or *bim4eepsa:Door*, *bim4eepsa:Wall*, and *bim4eepsa:Window* as subclasses of *bim4eepsa:BuildingElement*, which is subclass of *s4bldg:PhysicalObject*. All those axioms were gathered in a module named bim4EEPSA[20] 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.

Hierarchy structure below the *s4bldg:BuildingSpace* class is not developed in SAREF4BLDG. 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 [7].

[18] http://w3id.org/eepsa
[19] http://www.student.dtu.dk/~mhoras/bot/
[20] http://w3id.org/bim4eepsa

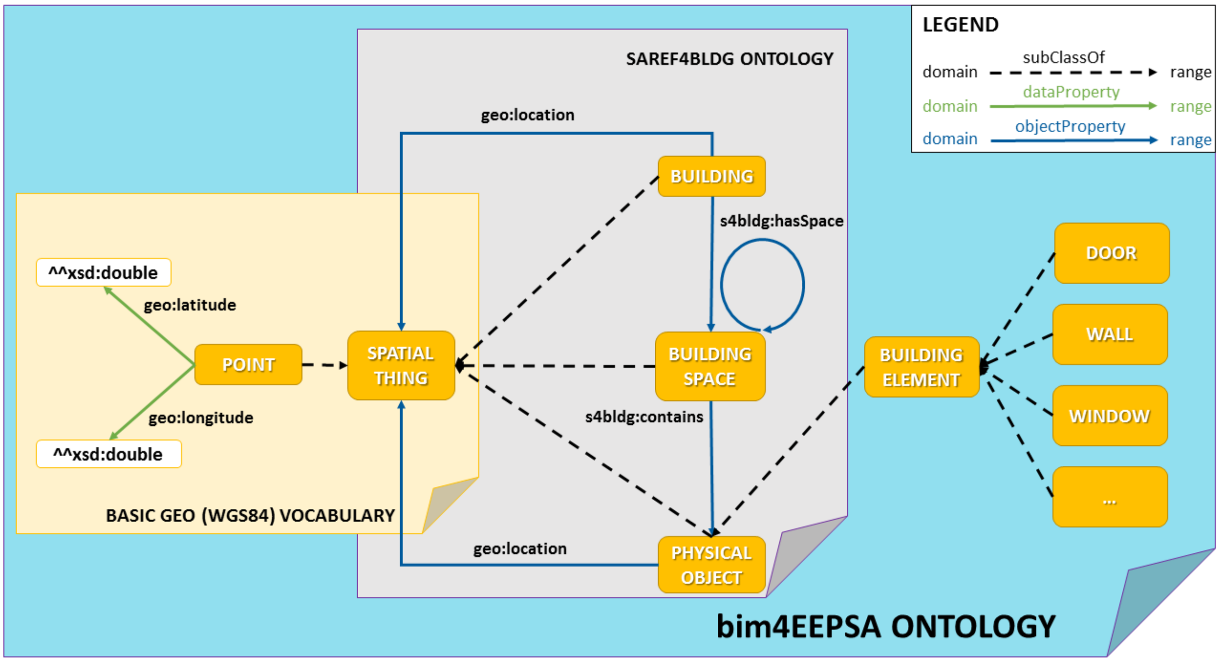

Fig. 3. An overview of the bim4EEPSA ontology's main classes and properties.

IfcOWL represents the IFC open standard for building and construction data. It was 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 [18], some of those properties (for instance, PSET Building Common) have been re-engineered and used to describe specific spaces such as the ones 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 ifcOWL models could be automatically translated to EEPSA models applying the simplification processes explained in [58].

Furthermore, the EEPSA ontology includes some properties that support the definition of queries and rules for the EEPSA process. For instance, the object property *eepsa:isAffectedBy* which relates spaces to variables (in the class *ssn:Property*) 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*.

As regards sensing devices, 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 classes of sensors, observable properties or units of measurements, the EEPSA ontology imports the module measurements4EEPSA[21]. This module is composed of a set of subclasses of *sosa:Sensor*, *ssn:Property* and *qudt:Unit* (and their corresponding properties) from the M3-Lite ontology. The Locality Module Extractor[22] tool [13] was used for automatically extract proper subclasses to be reused. Some of these classes were extended with some additional classes to improve coverage such as observable properties *m4eepsa:SpaceOccupancy* and *m4eepsa:WaterFlow*. Furthermore, the EEPSA ontology creates the property *eepsa:hasDataSource* to de-

[21] http://w3id.org/measurements4eepsa
[22] https://www.cs.ox.ac.uk/isg/tools/ModuleExtractor/

scribe the data sources where observable properties can be retrieved from when they are not measured by sensors.

Sensing device measurements (represented with *sosa:Observation*) and actuating devices actuations (represented with *sosa:Actuation*) are determined by the instant of time when the action is completed (data property *sosa:resultTime*) and its value, which can be represented with the *sosa:Result* class or the data property *sosa:hasSimpleResult* (in case of a simple value). 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 populated 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 assigned with 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:OutlierCausedBySunlight*) is assigned with two recommended solution methods: *eepsa:DeviceRelocation*, which recommends to relocate the device to an adequate place where it is not hit by sun and *eepsa:DeviceShelter*, recommending to shield the device with a Stevenson Screen or a similar one to cover it from direct heat radiation. Following any of these advices should avoid the device getting heated by direct sunlight and measuring erroneous observations.

This EEPSA Ontology is the base for all the KDD steps as detailed in the next sections.

3.2. Semantic Annotation

A preliminary phase to the KDD process that consists in annotating data with terms selected from appropriate ontologies and thus provide them with semantic. 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 [51], and additional background knowledge can be added to the dataset. Furthermore, the resulting dataset improves semantic interoperability [53], 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 design favours 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 [14]. However, no relevant work that supports the data analyst in 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 data analyst has the target building space 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 (*eepsa:NaturallyEnlightened Space*) 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 instantiating 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).

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 [35]. 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 [42]. 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 [10] and [37].

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 [25]. 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 [76] a domain ontology has been used to support the outlier detection based on a statistical method. In [30] 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 its 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 Detection) Framework [24], 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 resources, 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. Data analyst is then assisted to make use of these methods to generate a SPARQL query (SemOD Query) which retrieves sensor measurements that are presumably outliers because of being measured under a certain circumstance.

For example, a SemOD Method for detecting outliers caused by direct sun radiation, firstly offers a constraint pattern describing sun exposure times as presented in Listing 3 (in the Appendix A). Then, the method proposes the SPARQL query pattern shown in Listing 4 (in the Appendix A) for obtaining values asserted in the ontology to fill in the constraint pattern. 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_GENERATE D_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 need also to be specified with the corresponding variable's URI. Finally, the SPARQL query is generated and executed to obtain 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 [24].

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 [20] to social sciences [2]. Sensors are no exception and usually suffer from missing values caused by several reasons like a communication malfunction [44]. 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 elicitated, 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 [31]. This should be further studied in the future.

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 into a form in which data mining algorithms can accept as input. Amongst all the possible tasks in the Transformation phase (e.g. feature extraction), the EEPSA process focuses in the feature generation task.

The vast majority of existing feature generation solutions such as [11], [55] and [49] 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 obtained 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[23], and makes them publicly available[24] in a Virtuoso Open Source version 07.20.3217 Server[25]. 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

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

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

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

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. In addition, the EEPSA framework provides data analysts with parameterized SPARQL rules to infer values of variables based on existing data.

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 [43].

In [17] and [71] Linked Open Data has been proposed as a source of additional information to support the interpretation of the data mining method results. 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 EEPSA process addresses the question of energy efficiency while maintaining users' comfort in tertiary buildings. There are many complementary ways to save and optimize energy use in buildings, but since temperature is the most important weather parameter affecting electric load, forecasted temperatures constitute a basic ingredient in energy efficiency plans [1]. However, it is important to make clear that temperature forecasting is not the goal of the EEPSA process. These predictions are used to support prescriptive HVAC system activation strategies. That is, temperature of a space is predicted simulating the activation of HVAC systems at different times and intensities. For example, prediction of the temperature in a room when all HVAC systems are activated four hours in advance, when half of existing HVAC units are activated six hours in advance, etc. Estimating the temperature obtained with different strategies in advance, the one that uses energy in a more efficient way while maintaining the optimal comfort[26] can be chosen.

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 activation 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[27] that observe temperature, humidity and illuminance at five minutes intervals. There are three Tibucon devices located indoors and one located outdoors[28]. The Open Space is also equipped with eight HVAC units and collected information is simplified to whether any HVAC unit is activated or not.

A baseline experiment is developed without the support of the EEPSA process. This baseline's results

[26] 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.

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

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

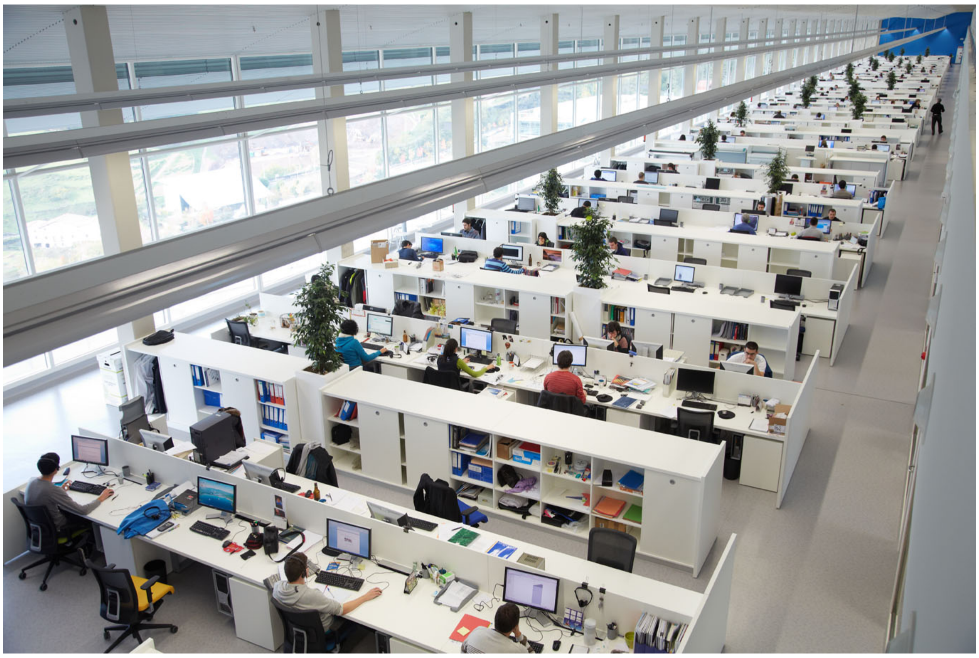

Fig. 4. IK4-TEKNIKER building's Open Space.

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 January 31st 2016 to August 1st 2016 is sampled hourly. Around 20% of data in this period is not measured due to external problems and in many circumstances, devices measure 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 *s4bldg:Building* is created to represent the IK4-TEKNIKER building (eepsa:ik4tekniker) in which it is contained. Then, the *eepsa:openSpace* is created as an individual of class *s4bldg:BuildingSpace*, related with *eepsa:ik4tekniker* by means of the property *s4eepsa:hasSpace*. Building elements of the Open Space are represented with individuals of classes such as *bim4eepsa:Door* or *bim4eepsa:Window* and are assigned with the property *s4bldg:contains*. 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[29] 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[30] is used. Ontop is an OBDA (Ontology-Based Data Access) tool which enables mappings between relational DB and an ontology [9]. 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. Nevertheless, 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.

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

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

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

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[31] 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 type classes, it is inferred that Open Space's indoor temperature might be affected by the following variables:

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

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) show that the outdoor device suffers from 1,253 outliers. This, together with the missing values the dataset has, is 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 is 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 data) need to be generated. In the Open Space, as previously stated, the outdoor temperature is considered as a low quality dataset due to its outliers and missing values, so it is decided to generate its values in this phase. Owing to the EEPSA ontology's OWL axioms, a reasoner infers that the outdoor temperature can be obtained from a weather station.

The first step is to check if there are any weather stations measuring outdoor temperature nearby the Open Space. To do so, a data analyst executes the GeoSPARQL query shown in Listing 8 (in the Appendix A) in the aforementioned Virtuoso SPARQL endpoint containing Euskalmet weather stations information[32]. The execution of this query returns 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 fac-

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

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

tors 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 decides which is the weather station chosen to retrieve the data, a parameterized SPARQL query has to be performed over the same endpoint. This time, the data analyst needs 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 is set with the variable outdoor temperature, the weather station Eitzaga and the time span between 1st January and 1st of August of 2016. The query will return the outdoor temperature values measured in the Eitzaga between the 1st January 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 is observed that another variable that is not being collected but affects the Open Space is the Wind Speed. This variable can also be retrieved from a weather station, so the same process as for outdoor temperature is followed.

After repeating this feature generation task as many times as needed, all data is used in the following Data Mining phase. In this case, the RapidMiner Studio 7.1 version is used alongside with the Linked Open Data extension. Within this extension, the operator SPARQL Data Importer is 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 experiment is developed without the support of the EEPSA process in the traditional KDD process. Different predictive models are built using different combinations of available variables and fine-tuning the parameters for their window sizes. Best results are obtained with a model built with Rapidminer's Vector Linear Regression algorithm[33] and containing 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[34].

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 [39]. Predicted indoor temperatures for the future 24 hours in the Open Space are: 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.

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 [77]. 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.

[33] https://docs.rapidminer.com/studio/operators/modeling/predictive/functions/vector_linear_regression.html

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

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

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 data analysts decision whether to incorporate or not the suggested variables.

Thanks to the knowledge-based outlier detection task, it was detected that the Tibucon device located outdoors had 1,253 anomalous temperature potentially caused by receiving direct sun radiation. 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. Thanks to this information a new device was located in a more adequate place where it is protected from direct sun 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 started the 24th March. Comparing the predictions obtained with the Baseline, 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, helping both novice and KDD experts. 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, currently only the SemOD Method for detect temperature outliers caused by sun radiation is implemented. The SemOD Framework should be extended with further SemOD Methods (e.g. outliers caused by rain), so that the data analyst could have a wide range of methods to identify and classify outliers generated 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 non-experts 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.

6. Acknowledgment

Part of the presented work is based on research contacted within the project BID3ABI (Big Data para RIS3 2016), which has received funding from the Basque Government (ELKARTEK 2016) under grant agreed reference KK-2016/00096. This work is also supported by FEDER/TIN2013-46238-C4-1-R and FEDER/TIN2016-78011-C4-2-R.

We thank Euskalmet (Basque Meteorology Agency) for assistance with weather stations and observations, as well as Zuzenean (Basque Citizen's Advice Service) for helping us with Open Data Euskadi (Basque Open Data portal).

References

[1] R. Abdel-Aal, *Hourly temperature forecasting using abductive networks*, Engineering Applications of Artificial Intelligence 17 (2004) 543-556.

[2] A.C. Acock, *Working with missing values*, Journal of Marriage and family 67 (2005) 1012-1028.

[3] R. Agarwal, D.G. Fernandez, T. Elsaleh, A. Gyrard, J. Lanza, L. Sanchez, N. Georgantas and V. Issarny, *Unified IoT Ontology to Enable Interoperability and Federation of Testbeds*, 3rd IEEE World Forum on Internet of Things (2016).

[4] G. Atemezing, O. Corcho, D. Garijo, J. Mora, M. Poveda-Villalón, P. Rozas, D. Vila-Suero and B. Villazón-Terrazas, *Transforming meteorological data into linked data*, Semantic Web 4 (2013) 285-290.

[5] A. Bernstein, F. Provost and S. Hill, *Toward intelligent assistance for a data mining process: An ontology-based approach for cost-sensitive classification, Knowledge and Data Engineering*, IEEE Transactions on 17 (2005) 503-518.

[6] E. Blomqvist, *The use of Semantic Web technologies for decision support - a survey*, Semantic Web 5 (2014) 177-201.

[7] D. Bonino and F. Corno, *Dogont-ontology modeling for intelligent domotic environments*, International Semantic Web Conference (2008) 790-803.

[8] D. Bonino, F. Corno and L. De Russis, *Poweront: An ontology-based approach for power consumption estimation in smart homes* , Internet of Things. User-Centric IoT, Springer, 2015, pp. 3-8.

[9] D. Calvanese, B. Cogrel, S. Komla-Ebri, R. Kontchakov, D. Lanti, M. Rezk, M. Rodriguez-Muro and G. Xiao, *Ontop: answering SPARQL queries over relational databases*, Semantic Web (2016) 1-17.

[10] V. Chandola, A. Banerjee and V. Kumar, *Anomaly detection: A survey*, ACM computing surveys (CSUR) 41 (2009) 15.

[11] W. Cheng, G. Kasneci, T. Graepel, D. Stern and R. Herbrich, *Automated feature generation from structured knowledge*, Proceedings of the 20th ACM international conference on Information and knowledge management (2011) 1395-1404.

[12] M. Compton, P. Barnaghi, L. Bermudez, R. García-Castro, O. Corcho, S. Cox, J. Graybeal, M. Hauswirth, C. Henson and A. Herzog, *The SSN ontology of the W3C semantic sensor network incubator group*, Web Semantics: Science, Services and Agents on the World Wide Web 17 (2012) 25-32.

[13] B. Cuenca Grau, I. Horrocks, Y. Kazakov and U. Sattler, *Modular reuse of ontologies: Theory and practice*, Journal of Artificial Intelligence Research 31 (2008) 273-318.

[14] A. Dadzie and M. Rowe, *Approaches to visualising linked data: A survey*, Semantic Web 2 (2011) 89-124.

[15] L. Daniele, F. den Hartog and J. Roes, *Created in close interaction with the industry: the smart appliances reference (SAREF) ontology*, International Workshop Formal Ontologies Meet Industries (2015) 100-112.

[16] M. d'Aquin, *Modularizing ontologies*, Ontology Engineering in a Networked World, Springer, 2012, pp. 213-233.

[17] M. d'Aquin and N. Jay, *Interpreting data mining results with linked data for learning analytics: motivation, case study and directions*, Proceedings of the Third International Conference on Learning Analytics and Knowledge (2013) 155-164.

[18] T.M. de Farias, A. Roxin and C. Nicolle, *IfcWoD, semantically adapting IFC model relations into OWL properties*, arXiv preprint arXiv:1511.03897 (2015).

[19] W. Derguech, E. Bruke and E. Curry, *An Autonomic Approach to Real-Time Predictive Analytics Using Open Data and Internet of Things*, Ubiquitous Intelligence and Computing, 2014 IEEE 11th Intl Conf on and IEEE 11th Intl Conf on and Autonomic and Trusted Computing, and IEEE 14th Intl Conf on Scalable Computing and Communications and Its Associated Workshops (UTC-ATC-ScalCom) (2014) 204-211.

[20] A.R.T. Donders, van der Heijden, Geert JMG, T. Stijnen and K.G. Moons, *A gentle introduction to imputation of missing values*, Journal of clinical epidemiology 59 (2006) 1087-1091.

[21] C.M. Eastman, C. Eastman, P. Teicholz, R. Sacks and K. Liston, *BIM handbook: A guide to building information modeling for owners, managers, designers, engineers and contractors*, John Wiley & Sons, 2011.

[22] R. Eastman, C. Schlenoff, S. Balakirsky and T. Hong, *A sensor ontology literature review*, 2013.

[23] V.L. Erickson and A.E. Cerpa, *Occupancy based demand response HVAC control strategy*, Proceedings of the 2nd ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Building (2010) 7-12.

[24] I. Esnaola-Gonzalez, J. Bermúdez, I. Fernández, S. Fernández and A. Arnaiz, *Towards a Semantic Outlier Detection Framework in Wireless Sensor Networks*, Proceedings of the 13th International Conference on Semantic Systems (2017)

[25] A. Fawzy, H.M. Mokhtar and O. Hegazy, *Outliers detection and classification in wireless sensor networks*, Egyptian Informatics Journal 14 (2013) 157-164.

[26] U. Fayyad, G. Piatetsky-Shapiro and P. Smyth, .From data mining to knowledge discovery in databases, AI magazine 17 (1996) 37.

[27] C. Fürber, D*ata quality management with semantic technologies*, Springer, 2015.

[28] C. Fürber and M. Hepp, *Using semantic web resources for data quality management*, International Conference on Knowledge Engineering and Knowledge Management (2010) 211-225.

[29] C. Fürber and M. Hepp, *Using SPARQL and SPIN for data quality management on the semantic web*, International Conference on Business Information Systems (2010) 35-46.

[30] L. Gao, M. Bruenig and J. Hunter, Semantic-based detection of segment outliers and unusual events for wireless sensor networks, arXiv preprint arXiv:1411.2188 (2014).

[31] U. Garciarena Hualde, *An investigation of imputation methods for discrete databases and multi-variate time series*, Master's Thesis, (2016).

[32] P.A. González and J.M. Zamarreno, *Prediction of hourly energy consumption in buildings based on a feedback artificial neural network*, Energy and Buildings 37 (2005) 595-601.

[33] A. Gyrard, C. Bonnet, K. Boudaoud and M. Serrano, *LOV4IoT: A second life for ontology-based domain knowledge to build Semantic Web of Things applications*, IEEE 4th International Conference on Future Internet of Things and Cloud (FiCloud) (2016) 254-261.

[34] H. Hagras, I. Packharn, Y. Vanderstockt, N. McNulty, A. Vadher and F. Doctor, *An intelligent agent based approach for energy management in commercial buildings*, IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2008) (2008) 156-162.

[35] J. Han, J. Pei and M. Kamber, *Data mining: concepts and techniques*, Elsevier, 2011.

[36] M. Hilario, A. Kalousis, P. Nguyen and A. Woznica, *A data mining ontology for algorithm selection and meta-mining*, Pro-

ceedings of the ECML/PKDD09 Workshop on 3rd generation Data Mining (SoKD-09) (2009) 76-87.

[37] V.J. Hodge and J. Austin, A survey of outlier detection methodologies, Artificial Intelligence Review 22 (2004) 85-126.

[38] M. Hofmann and R. Klinkenberg, *RapidMiner: Data mining use cases and business analytics applications*, CRC Press, 2013.

[39] R.J. Hyndman and G. Athanasopoulos, *Forecasting: principles and practice*, OTexts, 2014.

[40] K. Janowicz, F. Van Harmelen, J.A. Hendler and P. Hitzler, *Why the data train needs semantic rails*, AI Magazine (2014).

[41] M. Kolchin, N. Klimov, A. Andreev, I. Shilin, D. Garayzuev, D. Mouromtsev and D. Zakoldaev, *Ontologies for Web of Things: A Pragmatic Review*, in:Anonymous , Knowledge Engineering and Semantic Web, Springer, 2015, pp. 102-116.

[42] V. Kotu and B. Deshpande, *Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner*, Morgan Kaufmann, 2014.

[43] F. Lécué, R. Tucker, V. Bicer, P. Tommasi, S. Tallevi-Diotallevi and M. Sbodio, *Predicting severity of road traffic congestion using semantic web technologies, in:Anonymous , The Semantic Web: Trends and Challenges*, Springer, 2014, pp. 611-627.

[44] M.H. Le Gruenwald, *Estimating missing values in related sensor data streams*, COMAD (2005).

[45] S. Liaw, A. Rahimi, P. Ray, J. Taggart, S. Dennis, S. de Lusignan, B. Jalaludin, A. Yeo and A. 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 (2013) 10-24.

[46] J. Lu, T. Sookoor, V. Srinivasan, G. Gao, B. Holben, J. Stankovic, E. Field and K. Whitehouse, *The smart thermostat: using occupancy sensors to save energy in homes*, Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems (2010) 211-224.

[47] L. Madrazo, G. Nemirovski and A. Sicilia, *Shared vocabularies to support the creation of energy urban systems models* (2013).

[48] C. Martani, D. Lee, P. Robinson, R. Britter and C. Ratti, *ENERNET: Studying the dynamic relationship between building occupancy and energy consumption*, Energy and Buildings 47 (2012) 584-591.

[49] V. Narasimha, P. Kappara, R. Ichise and O. Vyas, *LiDDM: A Data Mining System for Linked Data*, Workshop on Linked Data on the Web. CEUR Workshop Proceedings 813 (2011).

[50] A.H. Neto and F.A.S. Fiorelli, *Comparison between detailed model simulation and artificial neural network for forecasting building energy consumption*, Energy and Buildings 40 (2008) 2169-2176.

[51] N.F. Noy, *Semantic integration: a survey of ontology-based approaches*, ACM Sigmod Record 33 (2004) 65-70.

[52] M.V. Nural, M.E. Cotterell and J.A. Miller, *Using Semantics in Predictive Big Data Analytics*, Big Data (BigData Congress), 2015 IEEE International Congress on (2015) 254-261.

[53] L. Obrst, *Ontologies for semantically interoperable systems*, Proceedings of the twelfth international conference on Information and knowledge management (2003) 366-369.

[54] F. Oldewurtel, A. Parisio, C.N. Jones, D. Gyalistras, M. Gwerder, V. Stauch, B. Lehmann and M. Morari, *Use of model predictive control and weather forecasts for energy efficient building climate control*, Energy and Buildings 45 (2012) 15-27.

[55] H. Paulheim and J. Fümkranz, *Unsupervised generation of data mining features from linked open data*, Proceedings of the 2nd international conference on web intelligence, mining and semantics (2012) 31.

[56] P. Pauwels and W. Terkaj, *EXPRESS to OWL for construction industry: Towards a recommendable and usable ifcOWL ontology*, Automation in Construction 63 (2016) 100-133.

[57] P. Pauwels, S. Zhang and Y. Lee, *Semantic web technologies in AEC industry: A literature overview*, Automation in Construction (2016).

[58] P. Pauwels and A. Roxin, *SimpleBIM: From full ifcOWL graphs to simplified building graphs*, Proceedings of the 11th European Conference on Product and Process Modelling (ECPPM 2016) (2017).

[59] Q.K. Quboa and M. Saraee, *A state-of-the-art survey on semantic web mining*, Intelligent Information Management 5 (2013) 10-17.

[60] P. Ristoski, *Towards Linked Open Data Enabled Data Mining*, in:Anonymous , The Semantic Web. Latest Advances and New Domains, Springer, 2015, pp. 772-782.

[61] P. Ristoski, C. Bizer and H. Paulheim, *Mining the web of linked data with rapidminer*, Web Semantics: Science, Services and Agents on the World Wide Web 35 (2015) 142-151.

[62] P. Ristoski and H. Paulheim, *Semantic Web in data mining and knowledge discovery: A comprehensive survey*, Web Semantics: Science, Services and Agents on the World Wide Web (2016).

[63] P. Ristoski and H. Paulheim, *Feature selection in hierarchical feature spaces*, International Conference on Discovery Science (2014) 288-300.

[64] J. Scott, A. Bernheim Brush, J. Krumm, B. Meyers, M. Hazas, S. Hodges and N. Villar, *PreHeat: controlling home heating using occupancy prediction*, Proceedings of the 13th international conference on Ubiquitous computing (2011) 281-290.

[65] T. Sekki, M. Airaksinen and A. Saari, *Impact of building usage and occupancy on energy consumption in Finnish daycare and school buildings*, Energy and Buildings 105 (2015) 247-257.

[66] F. Serban, J. Vanschoren, J. Kietz and A. Bernstein, *A survey of intelligent assistants for data analysis*, ACM Computing Surveys (CSUR) 45 (2013) 31.

[67] N. Seydoux, K. Drira, N. Hernandez and T. Monteil, *IoT-O, a Core-Domain IoT Ontology to Represent Connected Devices Networks*, Knowledge Engineering and Knowledge Management: 20th International Conference, EKAW 2016 (2016) 561-576.

[68] P.H. Shaikh, N.B.M. Nor, P. Nallagownden, I. Elamvazuthi and T. Ibrahim, *A review on optimized control systems for building energy and comfort management of smart sustainable buildings*, Renewable and Sustainable Energy Reviews 34 (2014) 409-429.

[69] A. Songpu, M.L. Kolhe, L. Jiao, N. Ulltveit-Moe and Q. Zhang, *Domestic demand predictions considering influence of external environmental parameters*, IEEE 13th International Conference on Industrial Informatics (INDIN) (2015) 640-644.

[70] P. Staroch, *A weather ontology for predictive control in smart homes*, Master's Thesis, 2013.

[71] I. Tiddi, *Explaining Data Patterns using Knowledge from the Web of Data*, The Open University (2016).

[72] U.S. Department of Energy , *Energy Efficiency Trends in Residential and Commercial Buildings*, (2008).

[73] U.S. Energy Information Administration, *International Energy Outlook 2016*, (2016).

[74] P. Vandenbussche, G.A. Atemezing, M. Poveda-Villalón and B. Vatant, *Linked Open Vocabularies (LOV): a gateway to reusable semantic vocabularies on the Web*, Semantic Web 8 (2017) 437-452.

[75] B.B. Wang, R.I. Mckay, H.A. Abbass and M. Barlow, *A comparative study for domain ontology guided feature extraction*, Proceedings of the 26th Australasian computer science conference-Volume 16 (2003) 69-78.

[76] Y. Wang and S. Yang, *Outlier detection from massive short documents using domain ontology*, IEEE International Conference on Intelligent Computing and Intelligent Systems (ICIS) (2010) 558-562.

[77] F. Zamora-Martínez, P. Romeu, P. Botella-Rocamora and J. Pardo, *Towards energy efficiency: Forecasting indoor temperature via multivariate analysis*, Energies 6 (2013) 4639-4659.

[78] H. Zhao and F. Magoulès, *A review on the prediction of building energy consumption*, Renewable and Sustainable Energy Reviews 16 (2012) 3586-3592.

Appendices

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 s4bldg: <https://w3id.org/def/saref4bldg#>

###  http://w3id.org/eepsa#NaturallyEnlightenedSpace
eepsa:NaturallyEnlightenedSpace rdf:type owl:Class ;
  owl:equivalentClass [ owl:intersectionOf ( s4bldg:BuildingSpace
       [ rdf:type owl:Class ;
         owl:unionOf ( [ rdf:type owl:Restriction ;
                 owl:onProperty eepsa:containsBuildingElement ;
                 owl:minQualifiedCardinality "1"^^xsd:nonNegativeInteger ;
                 owl:onClass bim4eepsa:ExternalWindow
               ]
               [ rdf:type owl:Restriction ;
                 owl:onProperty eepsa:containsBuildingElement ;
                 owl:minQualifiedCardinality "1"^^xsd:nonNegativeInteger ;
                 owl:onClass bim4eepsa:Skylight
               ]
             )
       ]
     ) ;
        rdf:type owl:Class
              ] ;
  rdfs:subClassOf s4bldg:BuildingSpace ,
          [ 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 (mainly sunlight
  but it can come from some artificial source of light) from the exterior.
  "@en .
```

Listing 1: eepsa:NaturallyEnlightenedSpace class axiom.

```
PREFIX s4bldg: <https://w3id.org/def/saref4bldg#>
PREFIX sosa: <http://www.w3.org/ns/sosa/>
PREFIX eepsa: <http://w3id.org/eepsa#>

SELECT DISTINCT ?affectingProperty
WHERE {
{
eepsa:mySpace eepsa:isAffectedBy ?affectingProperty.
}
MINUS
{
eepsa:mySpace s4bldg:contains ?sensor.
?sensor sosa:observes ?observedProperty
}
}
```

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 an 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>
?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 s4bldg: <https://w3id.org/def/saref4bldg#> .
@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 ,
           s4bldg:Building ;
           s4bldg:hasSpace eepsa:openSpace ;
           rdfs:comment "The IK4-TEKNIKER building" .

eepsa:openSpace rdf:type owl:NamedIndividual ,
           s4bldg:BuildingSpace ;
           eepsa:containsBuildingElement eepsa:door1 ,
                      eepsa:door2 ,
                      eepsa:door3 ,
                      eepsa:wall1 ,
                      eepsa:wall2 ,
                      eepsa:wall3 ,
                      eepsa:window1 ;
           s4bldg: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 s4bldg: <https://w3id.org/def/saref4bldg#>
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 s4bldg: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.