Towards Data-Driven Sustainable Design

Decision Support based on Knowledge Discovery in Disparate Building Data

Petrova, Ekaterina Aleksandrova; Pauwels, Pieter; Svidt, Kjeld; Jensen, Rasmus Lund

Published in:
Architectural Engineering and Design Management

DOI (link to publication from Publisher):
10.1080/17452007.2018.1530092

Publication date:
2019

Document Version
Accepted author manuscript, peer reviewed version

Towards Data-Driven Sustainable Design: Decision Support based on Knowledge Discovery in Disparate Building Data

Ekaterina Petrova[a], Pieter Pauwels[b], Kjeld Svidt[a], and Rasmus Lund Jensen[a]

[a]*Department of Civil Engineering, Aalborg University, Aalborg, Denmark*
[b]*Department of Architecture and Urban Planning, Ghent University, Ghent, Belgium*

Sustainable building design requires an interplay between multidisciplinary input and fulfilment of diverse criteria to align into one high-performing whole. BIM has already brought a profound change in that direction, by allowing execution of efficient collaborative workflows. However, design decision-making still relies heavily on rules of thumb and previous experiences, and not on sound evidence. To improve the design process and effectively build towards a sustainable future, we need to rely on the multiplicity of data available from our existing building stock. The objective of this research is, therefore, to transform existing data, discover new knowledge and inform future design decision-making in an evidence-based manner. This article looks specifically into this task by (1) outlining and distinguishing between the diverse building data sources and types, (2) indicating how the data can be analysed, (3) demonstrating how the discovered knowledge can be implemented in a semantic integration layer and (4) how it can be brought back to design professionals through the design aids they use. We, therefore, propose a performance-oriented design decision support system, relying on BIM, data mining and semantic data modelling, thereby allowing customised information retrieval according to a defined goal.

Keywords: BIM, Sustainability, Building Design, Semantics, Data Mining, Pattern Recognition, Knowledge Discovery, Information Retrieval

Introduction

Sustainable building design requires an optimal interplay between diverse criteria, susceptible to both the fulfilment of strictly formulated requirements, as well as their interpretation, translation and implementation by the design team. Hence, a performance-oriented design process requires multidisciplinary input to align into one high-performing 'whole', simultaneously with that being done in the most efficient way. 'Whole' as a concept, and the derived term 'holism', was defined by Smuts (1926) as *'a unity of parts, which is so close and intense as to be more than the sum of its parts'*. That means that all parts should function towards the whole, determine each other and eventually merge their individual characters, which makes the holistic character

This is an Accepted Manuscript of an article published by Taylor & Francis in Architectural Engineering and Design Management, Special Issue on Intelligent Building Paradigms and Data-Driven Models of Innovation, available online: http://dx.doi.org/10.1080/17452007.2018.1530092.

discoverable in the functions of both the parts and the whole. This concept is translated into whole building design by the implementation of the integrated design approach. Therefore, sustainable design requires a holistic approach, in which there are no individual parts constituting a design, only synergetic multidisciplinary inputs that contribute to the targeted overall performance of the whole.

In that relation, Building Information Modelling (BIM) (Eastman et al., 2011; Sacks et al., 2018) has already brought a profound change to the Architecture, Engineering and Construction (AEC) industry by allowing much more efficient integrated workflows. Open data standards and protocols, including Information Delivery Manuals (IDMs), Model View Definitions (MVDs), Industry Foundation Classes (IFC), etc. (buildingSMART, 2016) have served as catalysts towards increased collaboration between stakeholders. This is crucial for obtaining efficiency gains and successful fulfilling of performance targets related to sustainability in the building design domain. By definition, BIM allows integration of multidisciplinary information within a single coordinated building model and empowers collaborative practices (Zanni et al., 2017).

Furthermore, BIM practice strongly advises the use of a Common Data Environment (CDE) to manage information from all stakeholders. The CDE is defined as '*a central repository where construction project information is housed. The contents of the CDE are not limited to assets created in a 'BIM environment' and it will therefore include documentation, graphical model and non-graphical assets.'* (British Standards Institute, 2013). In a CDE, distinct viewpoints on a building are brought together, thus providing the place where a holistic view is possible. That includes data that is often not captured directly in a BIM model (e.g. design briefs, point cloud data, etc.) (Fig. 1).

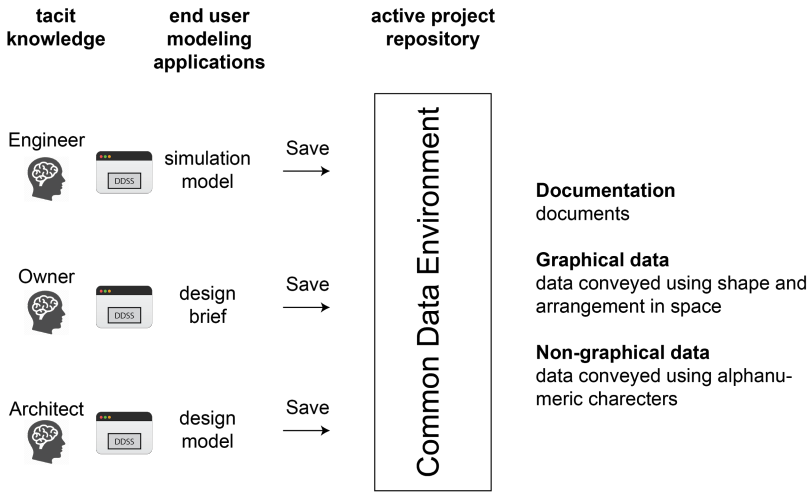

Figure 1. Use of a Common Data Environment in collaborative building design

As a result of the strong focus on BIM, BIM-based sustainable design has received major attention, and is a part of fundamental research within the construction

industry (Cemesova et al., 2015; Lu et al., 2017; Wong & Zhou, 2015). A considerable research effort, aiming for the seamless integration of BIM and building performance assessment in the (early) design process has also taken place in the last decade (El-Diraby et al., 2017; Ilhan & Yaman, 2016; Jalaei & Jrade, 2014; Liu et al., 2015; Schlueter & Thesseling, 2009; Shadram et al., 2016; Underwood & Isikdag, 2010; Yalcinkaya & Singh, 2015).

Even though BIM offers possibilities for synergy with sustainable design, many of the decisions taken during the design process are based on rules of thumb and previous experiences (Heylighen et al., 2007), which are not directly applicable or are not based on sound evidence. Polanyi (1958) defines such rules of thumb and experiences as tacit knowledge, and indicates that it is hard to capture, formalize and make explicit because of its context-specific nature. The increase in experience leads to more complex rules of thumb, which evolve into design patterns (Alexander, 1977). These patterns are crucial in one's understanding of what constitutes and satisfies the design context and heavily influence the design process.

Nevertheless, knowledge discovered in data from past projects and buildings in operation can be combined with the tacit knowledge for informing future design decision-making. As a result, huge potential would arise in achieving building design in a sustainable, efficient and evidence-based manner. One of the main research objectives in this regard is to leverage the multiplicity of data sources and types, and thus pave the way to knowledge discovery for evidence-based processes in design and engineering practice. To advance towards achievement of this objective, this study aims to employ the latest advances in three main areas:

(1) the full use of BIM as a means to reuse existing project data (e.g. through a CDE),

(2) the deployment of Knowledge Discovery in Databases (KDD) (Fayyad, 1996) to discover hidden knowledge in operational building data and inform future building design decision-making, and

(3) the reliance on semantic data modelling to represent the discovered knowledge in a semantically rich graph of data.

Despite not being the main focus, we hereby aim to also take into account the tacit knowledge and expertise used in design decision-making. The main principle is to identify meaningful and relevant patterns from previous projects and buildings in operation, transform information, discover new knowledge and better predict outcomes. The discovered knowledge will provide the basis for a design decision support system (DDSS), which is performance- and data-informed, rather than just data-dependent. Decision support systems are regarded here as computer-based tools adapted to support and aid complex decision-making and problem solving (Arnott & Pervan, 2008; Shim et al., 2002). Research in this area typically highlights the importance of information technology in improving the efficiency and effectiveness of decision-makers (Alter, 2004; Pearson & Shim, 1995). In the context of architectural design and engineering,

research limits more specifically to DDSS targeting the end user (Timmermans, 2016). Many commercial tools (CAD tools, BIM tools, simulation, visualization and coordination tools, etc.) have also been widely adopted in practice. However, they are most often stand-alone applications that do not implement the concept of knowledge reuse. We therefore aim to bring those features together in a DDSS that enables both knowledge sharing and reuse.

Methodological approach

This research relies on an extensive literature review aiming to identify both seminal works and state of the art developments within multiple research areas. Included here are design thinking and theory, BIM, sustainable building design and performance assessment, data analysis and artificial intelligence in performance-oriented architecture and civil engineering, as well as emerging technologies and computational approaches for improvement of design decision-making. We hereby also try to take into account design workflows in various settings. Based on this background research, we investigate the existing types of building data, their representations, formats, storage methods, and the way in which they can be handled by various algorithms, relative to variable goals of the knowledge discovery processes.

Next, we devise a system architecture that aims to bring the knowledge discovered in the available data to the end user and thereby support decision-making in future performance-oriented design processes. This system relies on three main approaches targeting knowledge discovery, namely data mining, geometric feature matching and direct semantic queries. We investigate to what extent the results of geometric similarity matching and data mining can be represented in semantic graphs, thereby relying on earlier work (Petrova et al., 2018a, 2018b). The resulting framework would therefore be able to successfully combine these approaches in support of AEC domain specialists working towards improving the built environment.

In this article, we first document key efforts for information exchange and data analysis in sustainable building design (Section 2). Section 3 proposes a system outline for holistic sustainable design relying on operational building data and project data repositories. Sections 4 and 5 summarize the proposed system, thereby indicating the main implementation methods, i.e. data mining, geometric feature matching and direct semantic queries. Finally, Section 6 presents a conclusion and outlines future work.

Data Exchange and Analysis in Collaborative Sustainable Building Design

Data-Driven and Experience-Based Design

Sustainability is a multi-dimensional matter, aiming for equal balance between economic and social development, and environmental protection (United Nations, 2010). From a collaborative perspective, Senciuc et al. (2015) define sustainable design as a complex system of elements linked by interdependencies and a process of

managing numerous perspectives. Furthermore, Kocaturk (2017) underlines the important role that technology plays in transforming the understanding of sustainability as a concept in the built environment, by enabling design innovation at product, process and operational levels. Sonetti et al. (2018) further highlight the potential of artificial intelligence and ICT tools for human-centric regenerative design. Building performance, on the other hand, besides being a criterion itself, is an outcome of a multidisciplinary set of multiple-criteria design decisions (Jalaei, et al., 2015). In that relation, the availability of data and the efficiency of its exchange are highly influential to both the design decision-making and its results. However, building design is characterized by fragmentation of processes and heterogeneity of actors, competencies and information sources. As a result, data is not readily available and not necessarily easily exchanged. As stated by Akin (2014), the information created and associated with the design must be available and applicable at all stages, without any losses, duplication of trivial processes or backtracking.

According to Aksamija (2012), high-performance design requires *"building performance predictions, use of simulations and modelling, research-based and data-driven processes."* BIM can facilitate knowledge transfer and experience between ongoing projects, but it is also important to use the experience from previous projects to adopt a holistic standpoint (Goldman & Zarzycki, 2014). Thus, for the design intent and performance targets to be achieved, the building operation needs to inform the design, and both phases should not be considered separate or independent, but parts of a cause and effect relationship. Furthermore, Goldman & Zarzycki (2014) claim that much of the data initially required for modelling could be based on predictions relying on data from previous projects. That would require pairing substantial data collection with captured professional expertise. Yet, the result would be a refined outcome, where quantified knowledge and professional experience are used in decision-making in a dedicated and structured way. According to Isikdag (2015), such a future transformation needs a *"focus on enabling an (i) integrated environment of (ii) distributed information which is always (iii) up to date and open for (iv) derivation of new information."* Goldman & Zarzycki (2014) further stipulate that a future data exchange network also has to be based on reuse of experience across designers, and requires knowledge to be modular and shareable.

Basics of Data Analytics and Application of KDD in the AEC Industry

Data analysis is becoming increasingly important for the built environment. Through the emergence of BIM, information as a concept has paved the way to changing the way professionals in the industry work. However, many questions still need to be answered with regards to what should be measured, how the information should be reported and stored, and most importantly, how it should be translated to knowledge and applied in practice. In that relation, Starkey & Garvin (2013) take a step back and highlight the variable, sometimes intertwining definitions of the terms data, information and knowledge from philosophical, semiotic and cybernetic points of view. From a

knowledge management perspective, Thierauf (1999) defines data as "unstructured facts and figures that have the least impact". Davenport & Prusak (2000) claim that, for data to become information, it needs to be contextualised, categorised, calculated and condensed, whereas knowledge implies know-how, meaning and understanding.

This article adopts the term data in a foundational way, as the building blocks for information, which in turn allows purposeful pattern discovery in various datasets, by the use of dedicated analytical approaches. The obtained analytical results would further allow combinations in support of cognitive processes in design. More specifically, the term 'data' in the current context refers to various types and representations of digital data, generated and available throughout the entire building life cycle. That includes generated design documentation (design brief databases) graphical design data (BIM models, simulation models, numeric geometric data), and non-graphical data (semantic design data, numeric simulation output, monitored operational performance data from sensor networks), etc. In other words, we refer to digital building data types in representations useful for further computational analyses. We explicitly focus on digital data and its representations to reflect and comply with the BIM and CDE-based workflows. The article further highlights the potential impact that discovered applicable knowledge in digital data can have on the future built environment.

From an analytical perspective, large volumes of data prove to be overwhelming when using traditional methods, which generate informative reports, but fail when it comes to analysis of their content (Soibelman & Kim, 2002). On the other hand, data mining, KDD and pattern recognition excel at the analysis of data and extraction of knowledge, and can facilitate an effective design space exploration.

Hand et al. (2001) define data mining as *"the analysis of large observational datasets to find unsuspected relationships and summarize the data in novel ways so that data owners can fully understand and make use of the data."* Additionally, Bishop (2006) states that *'pattern recognition is concerned with the automatic discovery of regularities in data through the use of computer algorithms and with the use of these regularities to take actions such as classifying the data into different categories'*. In that context, Piatetsky-Shapiro (1991) formulates knowledge as the end product of a data-driven discovery, whereas KDD represents the overall process of the extraction of useful knowledge. Data mining is the step in that process which employs specific algorithms to discover useful and previously unknown patterns in the data. Fayyad et al. (1996) state that the essential purpose is to discover high-level knowledge in low-level data. Furthermore, they define five essential steps, which transform the available raw data into actionable knowledge and insights of immediate value to the end user (Fig. 2).

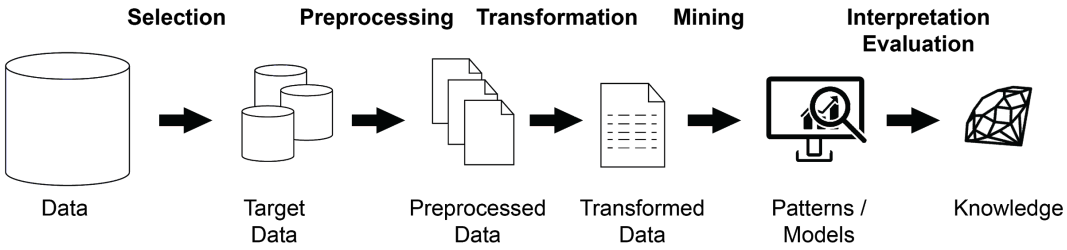

Figure 2. Knowledge Discovery in Databases (KDD) process, Fayyad et al. (1996)

(1) Selection

Data selection deals with the necessity to develop and understand the application domain, capture the relevant prior knowledge and identify the goal of the KDD process from an end-user perspective. Thereafter, a suitable target dataset or subset of variables should be chosen.

(2) Pre-processing

Pre-processing includes cleansing of the data in terms of handling of missing data fields, removal of duplicates, as well as fusion and resolution of conflicts due to the data originating from heterogeneous sources. Soibelman & Kim (2002) argue the significant importance of data preparation to the generation of high-quality knowledge through KDD. In addition, Cabena et al. (1998) point out that 60% of the time goes into data preparation, whereas the mining itself accounts for only 10% of the overall effort.

(3) Transformation

Transformation is concerned with reduction and projection of data with the purpose of finding useful features and representing the data according to the needs of the stated goal and the chosen algorithms. That includes finding invariant data representations and using dimensionality reduction methods to reduce the effective number of considered variables.

(4) Data mining

Data mining deals with matching the defined KDD goals with a particular method, e.g. classification, regression, or clustering. That includes the selection of algorithms and pattern extraction methods, as well as considerations concerning the end user's capabilities for interpretation of the chosen model vs. the model's predictive capabilities and accuracy. The actual data mining can then take place, i.e. searching for patterns in a particular representational form or set of representations, such as rule sets, trees, clusters, etc.

(5) Interpretation / Evaluation

The last step involves interpretation of the mined patterns and examination of their validity. That may include visualization of the discovered patterns and

assessment of their usefulness. Of particular importance is acting on the discovered knowledge, e.g. documenting it, using it directly, or implementing it into another system for further use.

Related Works

Fayyad et al. (1996) define six widely accepted data mining categories, namely classification, clustering, association rule mining, regression, summarization and anomaly detection. Han et al. (2012) further detail each of these techniques and highlight their belonging to two main categories: predictive (supervised) and descriptive (unsupervised). Supervised techniques are powerful for predictive modelling and knowledge representations (regression or classification models). They describe the qualitative or quantitative relationships between the input and output variables, and rely on domain expertise and training data (a set of observations, for which both the input and output variables are given). Thus, discovery of novel knowledge with predictive techniques is therefore unlikely, because inputs and outputs are predefined.

Unsupervised techniques (e.g. clustering, association rule mining, etc.), on the other hand, hold a significant potential in discovering the intrinsic structure, correlations and associations in data. Training data has no relation to the success of unsupervised analytics, as inputs and outputs are not predefined. In that relation, Han et al. (2012) state that the fundamental advantage of unsupervised methods lies within the ability to discover previously unknown and hidden knowledge in the given data. Unlike supervised approaches that adopt a backward approach by having a predefined target, unsupervised analytics are forward oriented, which gives the possibility of discovering interesting relationships and bringing out the value in the data (Fan et al., 2018).

As a result of their potential, KDD and data mining approaches have received major attention in the AEC industry. We performed a literature review that identifies main areas of application in the context of sustainability and energy efficiency, both from predictive and descriptive perspectives. Predictive applications include building energy use and demand prediction (Ahmed et al., 2011; Wang & Srinivasan, 2017; Zhao & Magoulès, 2012), prediction of building occupancy and occupant behaviour (D'Oca & Hong, 2014; Zhao et al., 2014), and fault detection diagnostics for building systems (Cheng et al., 2016; Pena et al., 2016). Descriptive tasks, on the other hand, are concerned with framework development (D'Oca & Hong, 2015; Fan et al., 2015a, 2015b; Park et al., 2016; Yu et al., 2013; Zhou et al., 2015), patterns in occupant behaviour (Capozzoli et al., 2017), building modelling and optimal control (Xiao & Fan, 2014), as well as discovering and understanding energy use patterns (Gaitani et al., 2010; Miller et al., 2015; Wu and Clements-Croome, 2007). Other efforts include the use of data mining for high-performance building design based on classification models for sustainability certification evaluation (Jun & Cheng, 2017), use of BIM-based data mining approaches for improvement of facility management (Peng et al., 2017), use of semantic modelling, neural networks and data mining algorithms for building energy management (McGlinn et al., 2017), etc.

However, the use of KDD and pattern recognition has been dedicated mostly to improvement of the building operation. Using discovered knowledge to improve future building design processes is an area that is rarely explored in detail. Efforts include pattern recognition in simulation data and extraction of information from BIM design log files (Yarmohammadi et al., 2016), use of data-driven approaches to design energy-efficient buildings by mining of BIM data (Liu et al., 2015) and data mining for extracting and recommending architectural design concepts (Mirakhorli et al., 2015).

Reuse of similarities for design decision support has also been recognised in design practice. This is prominent in case-based reasoning (CBR), which provides decision makers with a problem solving framework involving recalling and reusing previous knowledge and experience (Aamodt & Plaza, 1994). CBR approaches in design differ based on the method of their implementation (Elouti, 2009; Heylighen & Neuckermans, 2000; Richter et al., 2007). Example implementations in the context of sustainable architectural design can be found in (Sabri et al., 2017; Shen et al., 2017; Xiao et al., 2017).

In addition, research targeting the creation of a "repository of knowledge" for decision support based on patterns in thermal simulation output has been significantly extended in de Souza & Tucker (2015), de Souza & Tucker (2016) and Tucker & de Souza (2016). All similarity retrieval efforts mentioned above occupy the same conceptual space and are of high relevance to this research. Yet, despite coming a step closer to realizing the targeted future process, they rely on patterns only in design and simulation data. Thus, we aim to contribute further by adopting the latest semantic technologies, adding operational data mining and geometry matching capacities, and taking into account BIM and CDE-based workflows in early design.

The data analysis results coming from existing buildings and designs can rarely be linked to an early stage design using computational tools, mainly because the data representations do not match. This is not the case for tacit knowledge, which facilitates intuitive associations to any visual representation in an early design stage. A design professional would therefore tend to rely primarily on that knowledge instead of tangible performance data. In terms of data analysis, traditional approaches typically start from the available data and focus on retrieving the inherent insights. Decision-makers then determine how these insights may help them. As a result, despite the importance of the KDD goal definition, the knowledge discovery is driven only to a limited extent by the needs of the decision-maker.

Advanced analytical approaches start from the decision-makers and the identification of the most critical decisions, including the variability of their potential outcomes. As a result, the necessary insights to clarify those decisions can be identified, the type of information they may stem from, the data sources that could provide this information, and the knowledge to extract. Thus, a more user-oriented analysis is targeted, resulting in useful and practically applicable design decision support.

Towards Holistic Sustainable Design Relying on Operational Building Data and BIM Data Repositories

The ultimate objective of this research effort is to propose a DDSS that can bring forward a much more efficient sustainable building design process. More specifically, we aim to achieve informed decision-making by reusing existing BIM data repositories and operational building data. BIM data can include BIM models, simulation data, design briefs, etc.; operational data includes monitored data from existing buildings, i.e. sensor data, building use data, and so forth. The purpose is to integrate the DDSS in both the CDE as well as the individual end-user applications. That is found necessary, as the CDE hosts the information related to the building design process, and the end user applications host the individual decisions.

Data and Knowledge with Potential Impact on Design Decision Support

When implementing an advanced data analytics approach, there are several considerations, pertaining not only to the goals and criticality of the decisions, but also to the ability to generalize over the discovered patterns. Meaningful patterns are those that can be statistically justified, hence they should be based on the exploration of significant volumes of heterogeneous data. Furthermore, such an approach has highest impact when it can affect both the design process and the final product. In summary, the suggested approach works best in an environment that hosts simultaneously:

- decisions with high impact and criticality, namely early-stage design decisions with high level of variability of outcome
- specific performance criteria, concerning the practical implications of the decisions with regards to targeted building performance
- data from a high number of reference buildings
- data in big amounts and diversity

Many of the critical early decisions and the related requirements and constraints are interdependent. These dependencies can be captured in diagrams, which give a full overview of the relevant decision-making criteria and relations. Predictive models can hereby contribute further, by quantifying the weights of the dependencies, the criticality of the decisions, the variability of outcomes and the potential impacts. Figure 3 shows the developed dependency diagram capturing the relevant decision-making criteria in high-performance design. The grey nodes with most dependencies highlight not only the criticality of the related decisions, but also the data that would be most relevant for goal-oriented analytics. AEC projects generate various kinds of data in different formats, however, not all data are equally useful to all pattern recognition techniques. The following sections categorize the diverse data types based on their origin.

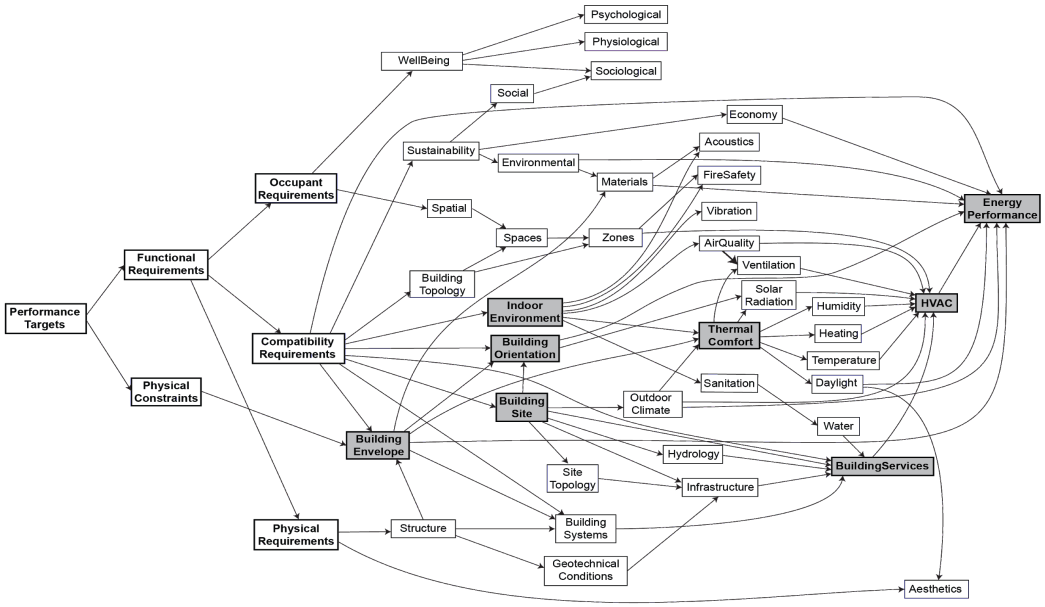

Figure 3. Criteria dependency in a typical sustainable design process

Data Types and Hidden Knowledge at Building Operation Stage

Operational building data is usually represented in a two-dimensional structured tabular way, with columns representing variables and rows storing the measurements at given time steps. Collected data usually includes time and date of measurement, energy consumption data (e.g. power consumption, cooling and heating loads, etc.), HVAC system operating conditions (temperature, flow rates, etc.), and environmental data (e.g. indoor and outdoor climate, humidity, solar radiation, etc.). These data types consist of parameters that are directly influencing building performance and are dynamically changing. Such data are a valuable input for data-driven simulations, HVAC system optimization and improvement of the building operation. Figure 4 represents the dynamic parameters and therefore operational data types typically collected from Building Management Systems (BMS).

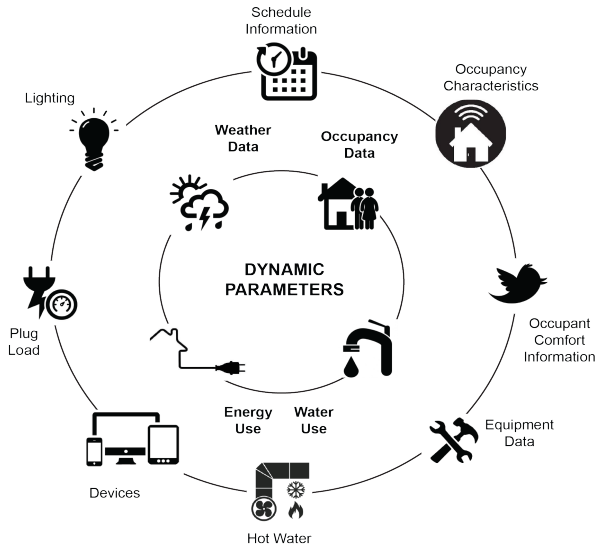

Figure 4. Dynamic parameters, based on taxonomy by Mantha et al. (2015)

According to Han et al. (2015), the typical formats and the tabular representation of operational building data gives an opportunity for discovery of two main types of knowledge: cross-sectional (static) and temporal (dynamic). Cross-sectional knowledge can be discovered when treating each row as an independent observation. The discovered knowledge is static, as the temporal dependencies between the rows are ignored (the knowledge discovered mainly includes the concurrent relationships among the different variables). Static knowledge discovery is useful for the identification of interaction between system components, atypicality in operation, etc. Han et al. (2015) further state that, in contrast, temporal knowledge can be discovered by mining data along both axes of the two-dimensional table and is very useful for characterizing dynamics in building operations. The insights obtained can be used for developing dynamic solutions for optimal building control, fault detection and diagnosis. Capturing the temporal dependencies in the data are much more challenging, but give a possibility for discovering unsuspected patterns and their relationships.

Data Types and Hidden Knowledge at Building Design Stage

The knowledge discovered in design data is much more static, even when taking into account versioning possibilities. Data at the building design stage typically starts with a design brief and a design model. Crucial choices on building orientation, zoning, spatial arrangement, and building materials are made in the earliest design stages. This data typically responds to the requirements and constraints listed earlier in the dependency diagram in Fig. 3 and represents important static parameters defining the character of the building (Fig. 5).

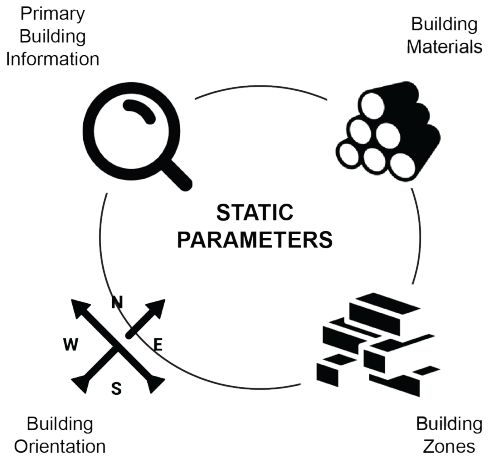

Figure 5. Static parameters, based on taxonomy by Mantha et al. (2015)

A lot of hidden knowledge is also available in the simulation data. This data can inform the design according to the paths defined in the dependency diagram by giving an insight into the building performance. Yet, they are typically a lot more optimistic compared to the actual performance. Building geometry is also valuable, as it provides many of the inputs required for simulation and compliance checking.

Data Type Definition from Analytical Perspective

To achieve high success rate in terms of analytical evaluation, it is important to match the types of data with the most suitable analytical techniques. Different data types can be recognized, informing the choice of analytical techniques and the structure of the data to enable effective knowledge discovery and performance-oriented decision support. The list below presents a data type definition from an analytical perspective.

- Semantic design data: semantic data describing design features, which include building elements, materials, object types, design brief data, etc.
- Numeric geometric data: geometric data in a format optimized for geometric analysis.
- Numeric sensor data: tabular sensor data with real-time data from supervisory control and data acquisition systems.
- Numeric simulation data: data models containing simulation results.

A Holistic Approach to a Data-Driven Sustainable Design System

This section proposes a system architecture that combines the available data with data analytics in a sensible way for decision support. This analysis is put forward through Fig. 6, which shows the main approach and the overall flow of proposed activities.

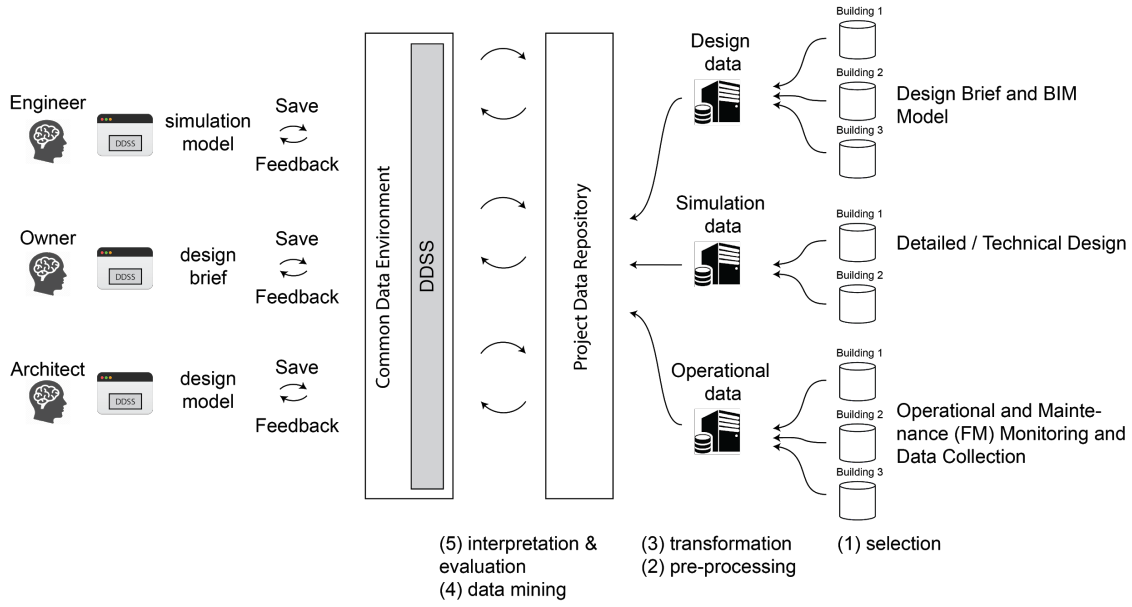

Figure 6. Proposed flow of data from existing buildings and project data repositories towards the diverse end-users

The active design environment (left in Fig. 6) may include BIM authoring tools, parametric design tools, simulation tools, etc. Design professionals iterate through a number of proposals within their individual tools and with the rest of the team. While designing, project data is stored in the CDE as files being uploaded to a central server.

In this study, DDSS systems are proposed both in the CDE and in the individual applications, where the DDSS in the CDE communicates to a project repository (Fig. 6). This repository collects the data available from previous projects and existing buildings, which comes from various heterogeneous sources. For example, BIM data captures the design, but typically comes in different representations, including a native 3D model, a neutral IFC data model, schedules, etc. Sensor data comes in different representations, depending on the system from which it originates. Storing local copies facilitates the execution of the data selection part of the KDD process defined by Fayyad et al. (1996) together with the maintaining of the original data. The selected data can then be cleansed and transformed, thus following steps 2 and 3 by Fayyad et al. (1996). After cleansing and transformation of the selected datasets, the results are stored in a project data repository, which hosts disparate data. While this allows diverse analysis techniques, integration across the data types will be needed.

The following sections indicate how the different components of the proposed system can be set up. We focus specifically on how different approaches may be effectively combined to achieve useful design decision support. Section 4 deals with the part of the system architecture related to the active design environment, including the semantic integration of data, while Section 5 introduces the use of KDD for creating a project data repository.

The Active Design Environment

End-users approach decision-making in an iterative problem-solution oriented manner, in which they put forward solutions based on tacit knowledge. When it comes to the DDSS, an insight into the cognitive processes within design decision-making provides an invaluable input for system design. We therefore first consider the overall design thinking processes, after which we outline how this takes form in a BIM-based process that relies on a CDE with heterogeneous data.

Design Thinking and Problem Solving as a part of Data-Driven Design

The background knowledge of the decision-maker determines the course of the design process. With each design iteration, designers explore a problem/solution space, thereby going through a continuous co-evolution of problem and solution (Dorst & Cross, 2001). As already indicated, the digital part of this process typically happens in a CDE, which stores the multidisciplinary design solutions as they come in sequentially. All actors go through a co-evolution process using their own tacit background knowledge and technology stack. The design requirements, typically captured in the design brief, drive the design decisions and follow the co-evolution of problem and solution. In the context of sustainable design, both the tacit definition for sustainable design and the solution responding to the particular requirements evolve throughout the design process. Ideally, the design team converges over time, under the influence of the design brief and the performance targets, both in the problem and solution spaces (Fig. 7). Convergence brings the team closer to a solution that fulfils the targets. The purpose is to avoid regress, e.g. widening of cycles at any given point in the evolution of the time dimension.

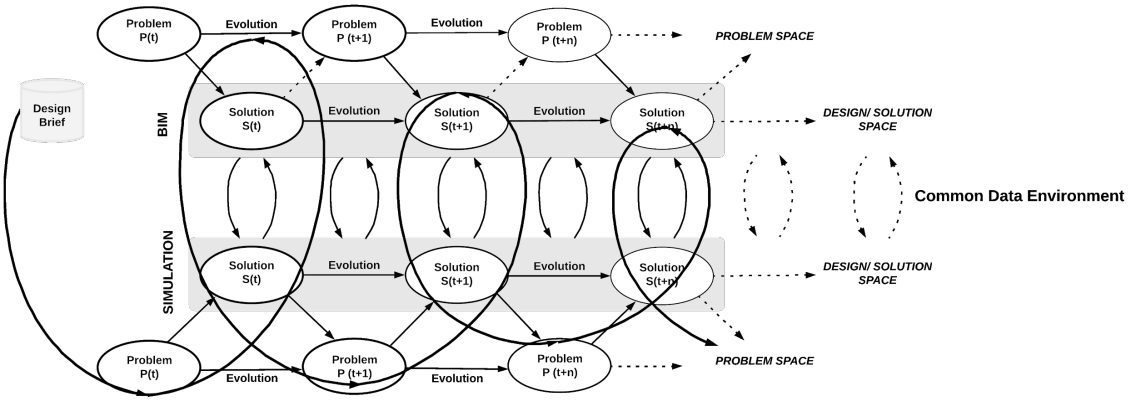

Figure 7. Problem-Solution cycle in collaborative design

In order to give tangible performance data a better role in the above process, the way in which decision-makers connect to their own background knowledge needs to be influenced. This can only be done by presenting the decision-maker useful alternatives

(problem-solution space), which match the goal and build on the tacit experience in a structured way.

Tools and Data Flows in the Active Design Environment

Even if a CDE is used, data is typically kept in separate files. This makes an integrated view over the available information very difficult to achieve. More recent initiatives aim at making the data available in an integrated manner using web technologies. As the web is evolving into a web of data instead of a web of documents (Berners-Lee et al., 2001), technology can be used to make the CDE web-compliant and data-oriented, as opposed to its current document-based nature. Such a system is much more attractive as (1) it makes project data available for semantic information retrieval and management, (2) it allows a larger diversity of data mining approaches, as data can be processed multiple times for different purposes while maintaining the same semantic identifiers, and (3) advanced semantic data mining techniques are within scope. Building a web-based semantic CDE results in the design environment outlined in Fig. 8.

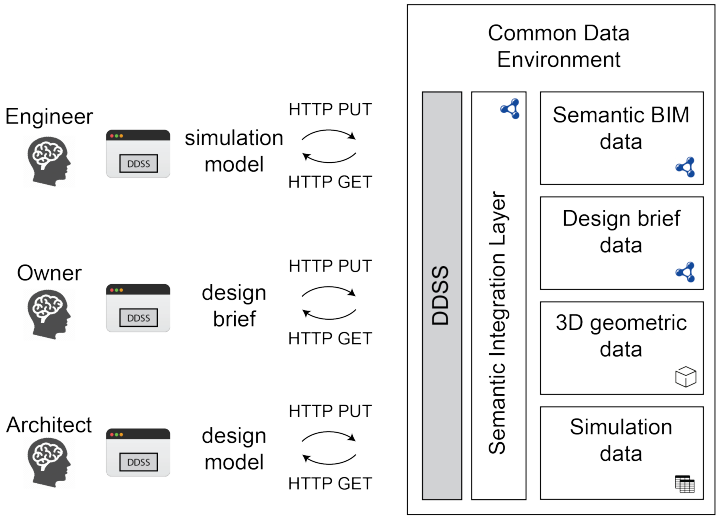

Figure 8. Integration of datasets in a web-based CDE

As the CDE has a web-based structure, applications and users are less occupied with manually storing files in an online server. Instead, the CDE is automatically filled with data using the HTTP protocol. By doing so, a lot more versioning and data logging can be achieved. Considering that data is gathered from multiple heterogeneous sources, the CDE would function optimally with a decentralized structure, which is most commonly realized using graph database approaches. Promising solutions in this regard for the AEC domain relate to deployment of linked data and semantic web technologies (Pauwels et al., 2017a). These technologies allow to build a decentralized web of semantic information, which serves perfectly for maintaining the backbone of a web-based CDE, thereby allowing to link the diverse datasets together, while respecting their original data structures.

Research has also shown that not all data can be efficiently maintained in a graph database or triple store (Pauwels et al., 2017b). We suggest that vast amounts of numeric data, such as geometric, simulation, and sensor data are therefore explicitly kept out of the semantic graph. Geometric data, such as 3D meshes, 2D drawings, point cloud data, etc., are ideally maintained in formats that can efficiently be parsed by geometric analysis algorithms. Sensor and simulation data are typically stored in tabular formats. Therefore, we propose a semantic integration layer (Fig. 8), which maintains the links between the individual datasets. The semantic integration layer is a thin and modular structure, capturing the key semantics of the different data sources in a decentralized manner, while referring to the original data sources that are kept in their optimized structures. The CDE can then be used to query the project data repository.

Reusing BIM Project Data and Operational Building Data

Matching queries from the CDE with the project data repository can occur in a number of ways, depending on how the data is stored. In this section, we look into the structure of the project data repository, and how pattern recognition and matching techniques can be applied to the data (direct queries, geometric feature matching, data mining). An overview diagram of the project data repository is given in Fig. 9.

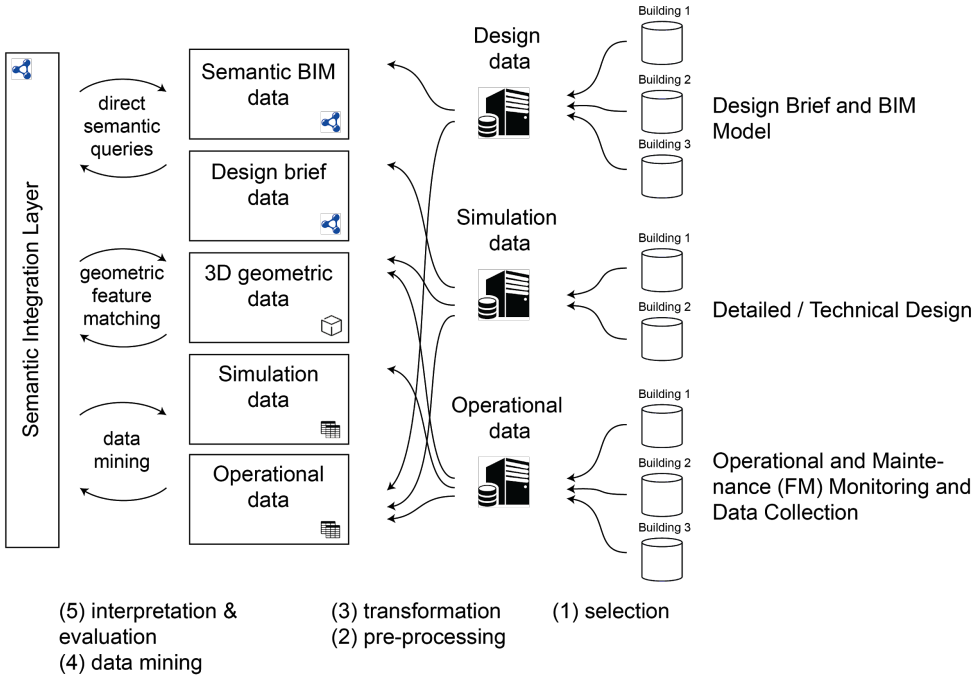

Figure 9. Overview of the project data repository

Structure of the Project Data Repository

Although a project data repository does not necessarily need to have the exact same structure as the CDE, it should be similarly well-structured. By maintaining this data

structure, and not converting all data into linked data, for example, we aim to allow as many as possible feature matching and data mining algorithms. Indeed, it is possible to transform all data to a semantic format, and then to query this data directly (Ristoski & Paulheim, 2016). Yet, this would disallow many of the efficient data mining and geometry matching algorithms that can be used for retrieving knowledge. Instead, we propose to store the semantic, geometric, and operational data separately. These datasets are then interlinked through the semantic data integration layer, which aims to link the semantic data model of a building with its numeric forms.

Clearly, the sole reliance on direct semantic information retrieval queries will be insufficient to give full feedback to an end user targeting a holistic performance-oriented design. The semantic queries do not capture the diversity of conclusions and matches that can be gathered from data mining techniques. Furthermore, relying solely on data mining techniques will not provide the integrated view over the diverse datasets. The same applies to geometric data; one cannot rely only on geometric data to retrieve valuable knowledge from a project repository to inform a designer aiming at holistic sustainable design solutions. Therefore, the diverse data sources need to be available and dynamically linked to allow information retrieval and design decision support.

To build a project data repository as proposed, a number of crucial steps need to be made. Data needs to be selected, cleansed and transformed so that it fits the project data repository. Furthermore, it is advisable to prepare separate local copies of the data in order not to intrude or violate data integrity at the source. In the selection process, it is possible to select only the data of relevance and place them on a local server (see step 1 in Fig. 9). For the static data, such as a design model, design brief, and simulation data, a direct copy can be used. For the dynamic data, such as the operational data and sensor data, data streams need to be accessed continuously. By implementing this data selection process, not only is the data in scope, but the original data is also maintained secure. In a next stage, data can be cleansed and transformed (steps 2 and 3 in Fig. 9). These are highly necessary steps to allow data mining with accurate results. The main purpose of the data transformation step is to end up in the structured project data repository as outlined above.

Recognizing Patterns from the Hive

Data Mining for Temporal Knowledge Discovery in Operational Building Data

Operational building data updates continuously with additional data points. The result is a data stream that gives an indication of the building operation (the heartbeat of the building). The dynamics in operation are usually very complex, due to changes in outdoor climate, indoor occupancy, systems utilization, etc., which rarely occur simultaneously. Discovering related temporal knowledge is of valuable importance to decision-making concerning building components, building automation and control systems, etc. Fan et al. (2015a) state that operational data is in essence multivariate time series data, where each observation is a vector of multiple measurements and control

signals, and time intervals between subsequent observations are usually fixed. That means that using temporal knowledge discovery can help capture relationships between variables over a particular time period.

Various approaches have been developed for temporal knowledge discovery of patterns, e.g. events, clusters, motifs (frequent sequential patterns), discords (infrequent sequential patterns) and temporal association rules, but rarely in the context of operational building data. A framework developed by Fan et al. (2015a) demonstrates encouraging potential in temporal knowledge discovery for improvement of building operations and performance management.

To inform design decision-making, it is important that the discovered patterns hold the potential to increase the confidence of the decisions, while still allowing creativity and variability of design space exploration. Considering the target data in this case and the goal for discovery of unsuspected patterns and relationships, unsupervised temporal knowledge mining should target motifs (and/or discords), as well as association rules (Fu, 2011). Motifs are by themselves valuable to temporal association rule mining and discord detection. We propose to use multivariate motif discovery as a first step (Vahdatpour et al., 2009), as it gives the possibility to discover both synchronous and asynchronous multivariate motifs consisting of univariate motifs or subsets of motifs. That is important, as in this context, motifs in operational building data do not necessarily start at the same time or have the same length. For example, turning the air conditioner on does not lead to an immediate change in indoor temperature due to the thermal mass (Fan et al., 2015a). Employing this method makes it possible to first discover univariate motifs and then use graph clustering approaches to identify multivariate motifs.

In addition, association rule mining (ARM) can help discover associations between variables (Agrawal et al., 1993). ARM usually targets cross-sectional knowledge and temporal dependencies are neglected. Due to the complexity and dynamics of operational building data, the use of temporal association rule mining (TARM) would be more useful, because it provides not only an insight into the associations between the variables, but also their temporal dependencies (Fournier-Viger et al., 2012). As a result, applying the above-mentioned techniques will allow decision-making support by identifying complex patterns over time, as well as the dependencies in their occurrence.

Feature Matching in Geometric Data

Geometric data can also be used for matching data in the CDE with data in the project data repository. Direct geometric pattern matching techniques can be implemented and used to return the most resembling results to a user. A number of geometry types and representations can be considered. One of the most commonly used is IFC, which is a neutral data model aiming to capture building semantics and object properties along with the full 3D geometry. IFC provides one of the most expressive neutral data models to describe building geometry in full semantic detail. A number of

alternative open data models are available as well. One example is the geometry ontology defined by Perzylo et al. (2015). Furthermore, Well-Known Text (WKT)[1] is a markup language that also allows specifying geometry with simple strings based on common agreement. Most WKT content refers to 2D geometry and is used for geospatial data, but it could also be used for representing 3D building geometry (Pauwels et al., 2017b).

Most of the above geometric data models can be captured in the form of labelled graphs. Yet, geometric topology graphs are slightly different, as they typically focus on the nodes and edges representing lines, boundaries, and points. An example of such a geometric topology graph is given for a room with four walls in Fig. 10.

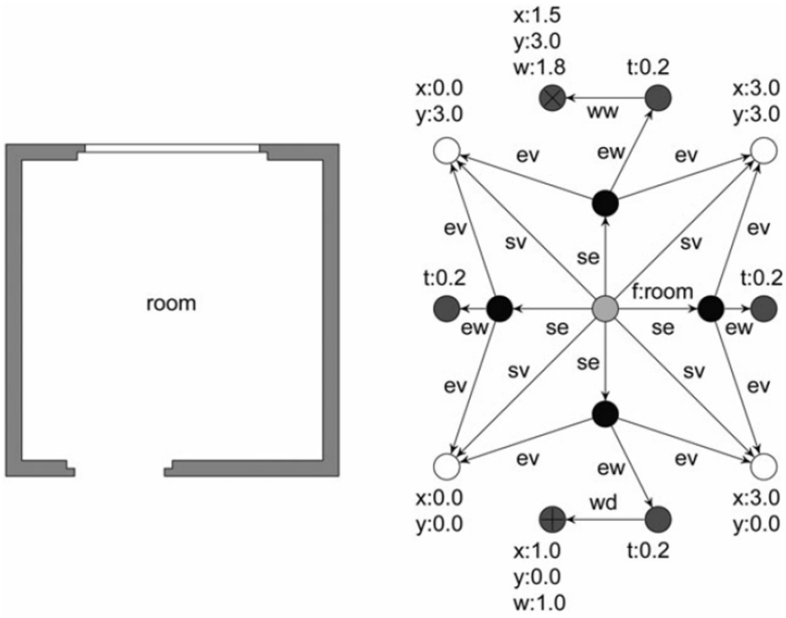

Figure 10. Geometric topology graph, Strobbe et al. (2016)

3D building data can also be represented using 3D mesh models. Yet, such data is semantically less defined and direct geometric feature matching techniques are less applicable. Point cloud data are also used to represent geometry, but, similarly to 3D mesh data, this data structure presents limited semantics.

For semantically rich geometric models, graph matching techniques can be used. Several direct graph matching techniques are available, in particular in data-oriented or web-oriented contexts. SPARQL, CYPHER, and GraphQL are graph query languages used for graph matching in a CDE. This technique assumes the target data to be available in graphs, which can be the case for IFC, WKT, and geometric topology models, but not for the rest.

Advanced geometric analysis algorithms can work with semantically unspecific data, such as point cloud data or 3D meshes, in order to make sense of the unstructured data and match them with the current geometric data in the CDE. Geometric analysis algorithms aim at parsing input geometry, including the unstructured mesh and point

[1] http://www.opengeospatial.org/standards/wkt-crs

cloud data. These are typically hardcoded algorithms, able to evaluate geometry and distil specific characteristics. The extracted characteristics are typically semantic and can thus be captured in a semantic data structure. Examples here are the GeoSPARQL[2] and BimSPARQL (Zhang et al., 2018) query languages, the first aiming at geospatial data and the second aiming at building data. The query languages contain statements such as "within" and "above", thus allowing to formulate geometric semantic queries.

Direct Semantic Queries

Another way to match data from a CDE to a project data repository is through direct semantic queries. Such queries can target the semantic integration layer, the semantic design model data and/or the semantic attributes that may be inferred from data mining or geometric feature recognition techniques.

The modular ontology structure proposed by the W3C Linked Building Data (LBD) Community Group[3] can serve to capture the considered semantics in an efficient way. This includes a number of ontologies, such as a Building Topology Ontology (BOT) (Rasmussen et al., 2017), a PRODUCT ontology, a PROPS ontology (properties), and an Ontology for Property Management (OPM). These ontologies allow to represent the building topology, product data, element properties and management of those properties. The OPM ontology is specifically useful, as it captures desired property values and whether they are achieved or not. Recent industry implementations further target the representation of design brief requirements in commercial graph databases, such as Neo4J, which is highly similar to the linked data approach. Hence, a semantically rich graph is possible based on OPM, BOT, PRODUCT, and PROPS ontologies.

Using linked data technologies, links can be maintained with the operational and geometric data. Device data can be captured using SAREF[4], home automation data can be represented using DogOnt (Bonino & Corno, 2008), and aggregate sensor data can be represented using SSN[5] and/or SOSA[6]. However, these ontologies do not serve well in case all operational data are targeted. In such case, a tabular format is still a lot more effective. The mentioned ontologies can be used to capture static characteristics, such as averages, min-max values, features of interest, devices, etc. The results of the geometric analysis algorithms can be captured in semantic graphs. These are static semantic annotations added to the semantic graph. Full geometric matching is however best done using the original data in a non-semantic format.

The semantic integration layer makes the connection with the non-semantic data possible, namely the reference source for operational data (web server address of

[2] http://www.opengeospatial.org/standards/geosparql
[3] https://www.w3.org/community/lbd/
[4] https://w3id.org/saref
[5] https://www.w3.org/TR/vocab-ssn/
[6] https://www.w3.org/ns/sosa/

specific sensor node data) and geometric data (web server address of specific geometric data file). The integration layer connects the semantic, geometric and operational data, so that any system accessing the data can recognize the associations.

Proposed System Architecture

The proposed system architecture utilizes measured operational building data and project data, which then serve as an input for the discovery of useful knowledge by the use of selected goal-oriented pattern recognition algorithms. The top in Fig. 11 represents the active design environment, which communicates with the project data repository (bottom in Fig. 11). This repository collects all reference data, linked together using the semantic integration layer, but also kept in their native formats. It is enriched using direct semantic queries, geometric feature matching, and data mining techniques, thereby allowing data-driven decision support for holistic performance-oriented design.

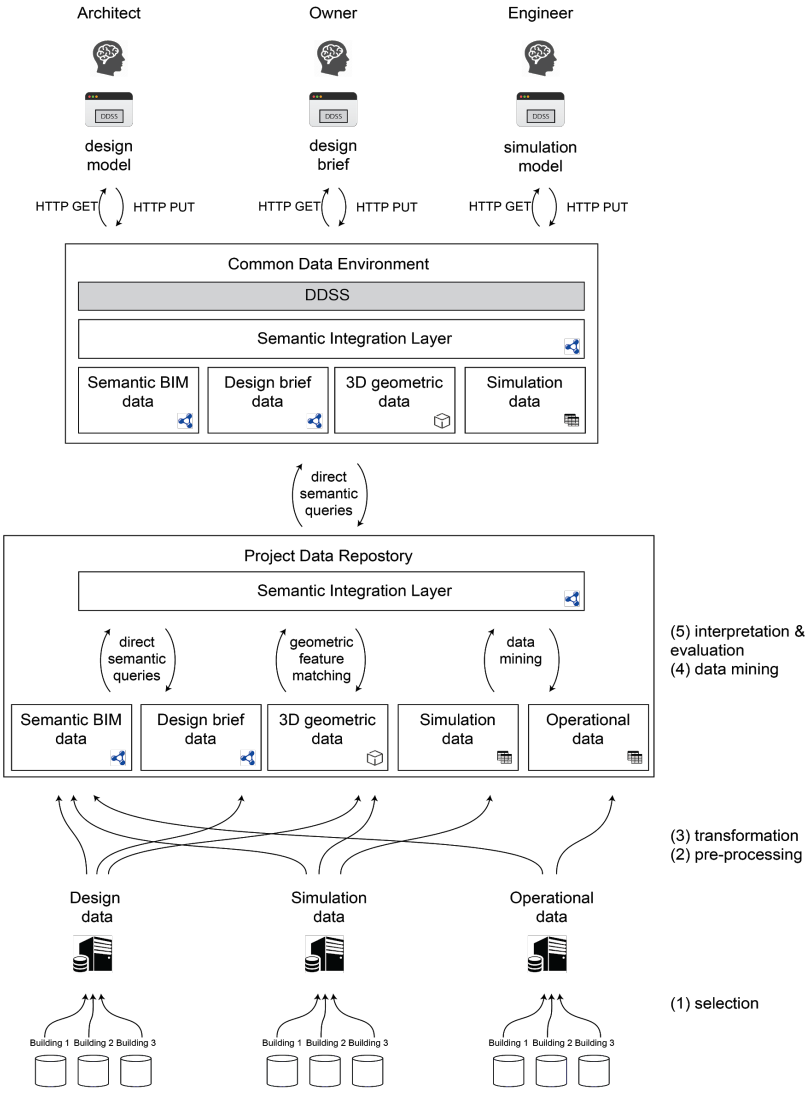

Figure 11. Proposed system architecture

Conclusion

This paper presents a framework for data-driven performance-oriented building design, relying on decision support from knowledge discovered in operational building data and project data repositories. The work identifies the relevant data types and combines three main approaches targeting knowledge discovery accordingly, namely data mining, geometric feature matching and direct semantic queries. The research identifies that the outcome of both the geometric similarity matching and the data mining can be represented in semantic graphs, which allows building a decision support system employing direct semantic queries. The combined approach allows semantic integration of heterogeneous datasets, their attributes and instances. The user-defined semantic queries allow customised information retrieval according to a defined goal.

One of the key challenges identified in this work is the implementation of a semantic integration layer, which combines data from various sources in a semantic graph, yet still allows to deploy data mining and geometric feature matching techniques. Although it is possible to include explicit results from these approaches in a graph (Petrova et al., 2018b), this might compromise the flexibility and modularity of the DDSS. By deploying the proposed web-based system architecture, we hope to overcome this challenge and make the data analysis and information retrieval user-driven. Such approach aims to integrate, yet also preserve the multiplicity of data and algorithms, allowing to deploy them to the maximum of their capabilities, in support of holistic sustainable design.

Future work needs to be done with regards to the testing and implementation of the proposed system in environments that can respond to the necessary requirements: design decisions with high impact and criticality, specific performance criteria, high number of reference buildings, and access to data in big amounts and diversity. Considering the diverse data analysis algorithms and web-based information retrieval approaches, the practical implementation needs to happen in an incremental and modular fashion, ideally involving a community knowledgeable in the architectural design, engineering and construction domains. This implementation process will indicate necessary changes in terms of performance, practical applicability, etc.

More importantly, however, this implementation process needs to reflect and capture the direct value that can be obtained in each concrete stakeholder environment. Of critical importance in future research are the methods that are used to ‘match past and present’ (CDE and project data repository). This match has not been discussed here at length. Choosing which matching mechanism (data mining, direct semantic queries, geometric feature matching) is used when, is of critical importance for the functioning of the system and needs to be investigated in further detail.

The proposed framework can be of significant importance for collaborative design teams aiming to improve the quality of the built environment in terms of sustainability, energy performance, indoor environmental quality, HVAC system design, etc. That includes a number of scenarios and contexts. This research effort targets the early design phase, where the decisions have the biggest impact on the future

performance. Thus, matching needs to be done as early as possible in the design process. The early design phase is, however, also one of the most difficult phases to provide decision support, because of the very limited amount of specific information that is available at this stage. Data is usually limited to an overall site definition, a design brief, and a preliminary layout of spaces. Most designers initially work in a 3D modelling environment, performing mass studies and spatial design exploration. Little semantic information can be obtained in such tools in contrast to the detailed data that can be accessed in the repository. Most useful data in this regard would likely be the building type, design brief, and overall structural system. Such information can inform and trigger queries to the repository, returning similarity-based matches in terms of structure, topology, and/or design requirements. Yet, specific features of retrieved cases, such as system components, material properties, operational performance parameters, etc. would potentially be retrieved in a second phase, which will naturally stimulate the use of BIM and CDE environments. This would in turn enhance further interpretation and learning by the design professionals, simultaneous with the implementation of their domain expertise in the decision-making. The proposed framework will also need to support that initial phase and infer design semantics and characteristics from very limited data. Further investigations are therefore needed to identify the efficiency of the proposed system in the very early design stages.

The devised framework can also be of direct relevance in the technical design phases, where many core decisions are already made, yet specific ones still need to be taken. Such environments rely heavily on digital models and tools, which once again reflects the positioning of the suggested framework in a BIM and CDE context. The above mentioned issues pertaining to availability of data in the early stages are generally not present here. This phase of the design process is strongly characterised by an abundance of data, both in terms of types and representations. As the proposed system aims to leverage exactly this multiplicity of data, it should fit in this part of the design and engineering process. As a result, the workflows characteristic to design practice at this stage would be preserved, apart from the additional presence of precise user-centred recommendations coming in through the BIM and CDE tools.

Using tangible performance data to impact decision-making and prevent errors early in the design phase is increasingly important. Leveraging computational approaches to enhance sustainability-oriented practices, and following an evidence-based path will empower knowledge sharing and reuse, and reduce knowledge vaporization and uncertainty in design decision-making.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Aamodt, A., & Plaza, E. (1994). Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches. *AI Communications*, *7*(1), 39-59.

Agrawal, R., Imielinski, T., & Swami, A. (1993, May). Mining association rules between sets of items in large databases. In *SIGMOD 1993*. *Proceedings of the 1993 ACM SIGMOD international conference on Management of data* (pp. 207–216). Washington, DC: ACM.

Ahmed, A., Korres, N.E., Ploennigs, J., Elhadi, H. & Menzel, K. (2011). Mining building performance data for energy-efficient operation. *Advanced Engineering Informatics, 25*, 341–354.

Akin, Ö. (2014). Necessity of Cognitive Modeling in BIM's Future. In: *Building Information Modeling BIM in Current and Future Practice.* New Jersey: Wiley, pp. 17-27.

Aksamija, A. (2012). *BIM-Based Building Performance Analysis: Evaluation and Simulation of Design Decisions.* Washington, DC: Omnipress.

Alexander, C. (1977). *A Pattern Language*. New York, NY: Oxford University Press.

Alter, S. (2004). A work system view of DSS in its fourth decade. *Decision Support Systems*, *38*, 319-327.

Arnott, D. & Pervan, G. (2008). Eight key issues for the decision support system discipline, *Decision Support Systems*, *44*, 657-672.

Berners-Lee, T., Hendler, J. & Lassila, O. (2001). The Semantic Web, *Scientific American*, 29-37.

Bishop, C.M. (2006). *Pattern Recognition and Machine Learning*. Cambridge, UK: Springer.

Bonino, D. & Corno, F. (2008). DogOnt - Ontology Modeling for Intelligent Domotic Environments. In *Lecture Notes in Computer Science: Vol: 5318*. *Proceedings of the International Semantic Web Conference* (pp. 790-803).

British Standards Institute (2013). PAS 1192-2:2013 Specification for information management for the capital/delivery phase of construction projects using building information modelling.

BuildingSMART (2016). BuildingSMART specifications. Retrieved from http://www.buildingsmart-tech.org/specifications.

Cabena, P., Hadjinian, P., Stadler, R., Verhees, J. & Zanasi, A. (1998). *Discovering data mining: From concept to implementation*, Upper Saddle River, NJ: Prentice Hall.

Capozzoli, A., Piscitelli, M.S., Gorrino, A., Ballarini, I. & Corrado, V. (2017). Data analytics for occupancy pattern learning to reduce the energy consumption of HVAC systems in office buildings. *Sustainable Cities and Society*, *35*, 191–208.

Cemesova, A., Hopfe, C. J. & Mcleod, R. S. (2015). PassivBIM: Enhancing interoperability between BIM and low energy design software. *Automation in Construction, 57*, 17-32.

Cheng, Z., Zhao, Q., Wang, F., Chen, Z., Jiang, Y. & Li, Y. (2016). Case studies of fault diagnosis and energy saving in buildings using data mining techniques. In *Proceedings of IEEE International Conference on Automation Science and Engineering* (pp. 646-651). Fort Worth, TX: IEEE.

Davenport, T. H. & Prusak, L. (2000). *Working Knowledge: How organizations manage what they know.* Boston, MA: Harvard Business School Press.

de Souza, C. B. & Tucker, S. (2015). Thermal simulation software outputs: a framework to produce meaningful information for design decision-making, *Journal of Building Performance Simulation*, *8*(2), 57-78.

de Souza, C. B. & Tucker, S. (2016). Thermal simulation software outputs: a conceptual data model of information presentation for building design decision-making. *Journal of Building Performance Simulation*, *9*(3), 227-254.

D'Oca, S. & Hong, T. (2014). A data-mining approach to discover patterns of window opening and closing behavior in offices. *Building and Environment*, *82*, 726-739.

D'Oca, S., & Hong, T. (2015). Occupancy schedules learning process through a data mining framework. *Energy and Buildings*, *88*, 395–408.

Dorst, K. & Cross, N. (2001). Creativity in the design process: co-evolution of problem-solution. *Design Studies*, *22*(5), 425-437.

Eastman, C., Teicholz, P., Sacks, R. & Liston, K. (2011). *BIM Handbook - A guide to Building Information Modeling for Owners, Managers, Designers, Engineers, and Contractors* (2nd ed.), Wiley.

Elouti, B.H. (2009). Design knowledge recycling using precedent-based analysis and synthesis models. *Design Studies*, *30*, 340-368.

El-Diraby, T., Krijnen, T. & Papagelis, M. (2017). BIM-based collaborative design and socio-technical analytics of green buildings. *Automation in Construction*, *82*, 59–74.

Fan, C., Xiao, F., Madsen, H. & Wang, D. (2015a). Temporal knowledge discovery in big BAS data for building energy management. *Energy and Buildings*, *109*, 75-89.

Fan, C., Xiao, F. & Yan, C. (2015b). A framework for knowledge discovery in massive building automation data and its application in building diagnostics. *Automation in Construction*, *50*, 81-90.

Fan, C., Xiao, F., Li, Z. & Wang, J. (2018). Unsupervised data analytics in mining big building operational data for energy efficiency enhancement: A review. *Energy and Buildings*, *159*, 296–308.

Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, *17*, 37-54.

Fournier-Viger P., Wu C.W., Tseng V.S. & Nkambou R. (2012). Mining Sequential Rules Common to Several Sequences with the Window Size Constraint. In *Lecture Notes in Computer Science: Vol. 7310. Advances in Artificial Intelligence* (pp. 299–304). Berlin, Heidelberg: Springer.

Fu, T.C. (2011). A review on time series data mining. *Engineering Applications of Artificial Intelligence*, *17*, 164–181.

Gaitani N., Lehmann C., Santamouris M., Mihalakakou G. & Patargias P. (2010). Using principal component and cluster analysis in the heating evaluation of the school building sector. *Applied Energy*, *87*, 2079 – 2086.

Goldman, G. & Zarzycki, A. (2014). Smart Buildings/ Smart(er) Designers: BIM and the Creative Design Process. In: *Building Information Modeling BIM in Current and Future Practices.* New Jersey: Wiley, pp. 3-16.

Han, J.W., Kamber, M. & Pei, J. (2012). *Data mining concepts and techniques* (3rd ed.) Waltham, US: Morgan Kaufmann.

Hand D., Mannila H. & Smyth P. (2001). *Principles of Data Mining*. Cambridge, USA: MIT Press.

Heylighen, A., Martin, M. & Cavallin, H. (2007). Building Stories Revisited: Unlocking the Knowledge Capital of Architectural Practice. *Architectural Engineering and Design Management*, *3*(1), 65-74.

Heylighen, A. & Neuckermans, H. (2000). DYNAMO: A Dynamic Architectural Memory On-line. *Educational Technology & Society, 3*(2), 86-95.

Ilhan, B. & Yaman, H. (2016). Green building assessment tool (GBAT) for integrated BIM-based design decisions. *Automation in Construction*, *70*, 26-37.

Isikdag, U. (2015). *Enhanced Building Information Models: Using IoT Services and Integration Patterns* (1st ed.). Istanbul: Springer.

Jalaei,. F. & Jrade, A. (2014). Integrating Building Information Modeling (BIM) and Energy Analysis Tools with Green Building Certification System to Conceptually Design Sustainable Buildings. *Journal of Information Technology in Construction (ITcon)*, *19*, 494-519.

Jalaei, F., Jrade, A. & Nassiri, M. (2015). Integrating decision support system (DSS) and building information modeling (BIM) to optimize the selection of sustainable building components. *Journal of Information Technology in Construction (ITcon), 20*, 399-420.

Jun, M.A. & Cheng, J.C.P. (2017). Selection of target LEED credits based on project information and climatic factors using data mining techniques. *Advanced Engineering Informatics*, *32*, 224–236.

Kocaturk, T. (2017). Towards An Intelligent Digital Ecosystem - Sustainable Data-Driven Design Futures. In: *Future Challenges for Sustainable Development within the Built Environment*. UK: Wiley-Blackwells, pp. 164-178.

Liu, Y., Huang, Y.C. & Stouffs, R. (2015a). Using a data-driven approach to support the design of energy-efficient buildings. *Journal of Information Technology in Construction (ITcon)*, *20*, 80-96.

Liu, S., Meng, X. & Tam, C. (2015b). Building information modeling based building design optimization for sustainability. *Energy and Buildings*, *105*, 139-153.

Lu, Y., Wu, Z., Chang, R. & Li, Y. (2017). Building Information Modeling (BIM) for green buildings: A critical review and future directions. *Automation in Construction*, *83*, 134-148.

Mantha, B.R.K., Menassa, C.C. & Kamat, V.R. (2015). A taxonomy of data types and data collection methods for building energy monitoring and performance simulation. *Advances in Building Energy Research*, *10*(2), 263-293.

McGlinn, K., Yuce, B., Wicaksono, H., Howell, S., & Rezgui, Y. (2017). Usability evaluation of a web-based tool for supporting holistic building energy management, *Automation in Construction*, *84*, 154–165.

Miller, C., Nagy, Z. & Schlueter, A. (2015). Automated daily pattern filtering of measured building performance data. *Automation in Construction*, *49*, 1-17.

Mirakhorli, M., Chen, H. & Kazman, R. (2015). Mining Big Data for Detecting, Extracting and Recommending Architectural Design Concepts. In *Proceedings of the 1st IEEE/ACM International Workshop on Big Data Software Engineering* (pp. 15-18). Florence: IEEE.

Park, H.S., Lee, M., Kang, H., Hong, T. & Jeong, J. (2016). Development of a new energy benchmark for improving the operational rating system of office buildings using various data-mining techniques. *Applied Energy*, *173*, 225-237.

Pauwels, P., Zhang, S. & Lee, Y.C. (2017a). Semantic web technologies in AEC industry: A literature overview. *Automation in Construction*, *73*, 145-165.

Pauwels, P., Krijnen, T., Terkaj, W., & Beetz, J. (2017b). Enhancing the ifcOWL ontology with an alternative representation for geometric data. *Automation in Construction*, *80*, 77-94.

Pearson, J.M., & Shim, J.P. (1995). An empirical investigation into DSS structure and environments. *Decision Support Systems*, *13*, 141-158.

Pena, M., Biscarri, F., Guerrero, J.I., Monedero, I. & León, C. (2016). Rule-based system to detect energy efficiency anomalies in smart buildings, a data mining approach. *Expert Systems With Applications*, *56*, 242–255.

Peng, Y., Lina, J.R., Zhang, J.P. & Hu, Z.Z. (2017). A hybrid data mining approach on BIM-based building operation and maintenance. *Building and Environment*, *126*, 483–495.

Perzylo, A., Somani, N., Rickert, M., & Knoll, A. (2015). An ontology for CAD data and geometric constraints as a link between product models and semantic robot task descriptions. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 4197-4203). Hamburg: IEEE.

Petrova, E., Pauwels, P., Svidt, K., & Jensen, R.L. (2018a). In Search of Sustainable Design Patterns: Combining Data Mining and Semantic Data Modelling on Disparate Building Data. In *Proceedings of the 35th CIB W78 Conference* (in press).

Petrova, E., Pauwels, P., Svidt, K., & Jensen, R.L. (2018b). From patterns to evidence: Enhancing sustainable building design with pattern recognition and information retrieval approaches. In Karlshøj & Scherer (Eds.)*, ECPPM 2018. Proceedings of the 12th European Conference on Product & Process Modelling. eWork and eBusiness in Architecture, Engineering and Construction* (pp. 391-399). Copenhagen: CRC Press.

Piatetsky-Shapiro, G. (1991). Knowledge Discovery in Real Databases: A Report on the IJCAI-89 Workshop. *AI Magazine*, *11*(5), 68–70.

Polanyi, M.(1958). *Personal Knowledge: Towards a Post-Critical Philosophy*. Chicago, IL: The University of Chicago Press.

Rasmussen, M.H., Pauwels, P., Hviid, C.A. & Karlshøj, J. (2017). Proposing a central AEC ontology that allows for domain specific extensions. In *Proceedings of the Joint Conference on Computing in Construction (JC3)* (pp. 237-244).

Richter, K., Heylighen, A., & Donath, D. (2007). Looking back to the future-an updated case base of case-based design tools for architecture. In *Proceedings of the 5th eCAADe Conference* (pp. 285-292).

Ristoski, P. & Paulheim, H. (2016). Semantic Web in data mining and knowledge discovery: A comprehensive survey. *Web Semantics: Science, Services and Agents on the World Wide Web*, *36*, 1–22.

Sabri, Q.U., Bayer, J., Ayzenshtadt, V., Bukhari, S.S., Althoff, K.D. & Dengel, A. (2017). Semantic Pattern-based Retrieval of Architectural Floor Plans with Case-based and Graph-based Searching Techniques and their Evaluation and Visualization. In *Proceedings of the 6th International Conference on Pattern Recognition Applications and Methods* (pp. 50-60).

Sacks, R., Eastman, C., Lee, G. & Teicholz, P. (2018). *BIM Handbook: A Guide to Building Information Modeling for Owners, Designers, Engineers, Contractors, and Facility Managers* (3rd ed.). Hoboken, NJ: Wiley.

Schlueter, A. & Thesseling, F. (2009). Building information model based energy/exergy performance assessment in early design stages. *Automation in Construction, 182*, 153-163.

Senciuc, A., Pluchinotta, I. & Rajeb, S. B. (2015). *Collective Intelligence Support Protocol: A Systemic Approach for Collaborative Architectural Design.* Mallorca: Springer.

Shadram, F., Johansson, T.D., Lu, W., Schade, J. & Olofsson, T. (2016). An integrated BIM-based framework for minimizing embodied energy during building design. *Energy and Buildings*, *128*, 592-604.

Shen, L., Yan, H., Fan, H., Wu, Y. & Zhang, Y. (2017). An integrated system of text mining technique and case-based reasoning (TM-CBR) for supporting green building design. *Building and Environment*, *124*, 388-401.

Shim, J.P., Warkentin, M., Courtney, J.F., Power, D.J., Sharda, R., & Carlsson, C. (2002). Past, present and future of decision support technology. *Decision Support Systems*, *33*, 111-126.

Smuts, J.C. (1926). *Holism and evolution*. London, UK: Macmillan.

Soibelman, L. & Kim, H. (2002). Data preparation process for construction knowledge generation through knowledge discovery in databases. *Journal of Computing in Civil Engineering*, *16*(1), 39–48.

Sonetti, G., Naboni, E. & Brown, M. (2018). Exploring the Potentials of ICT Tools for Human-Centric Regenerative Design. *Sustainability*, *10*.

Starkey, C. & Garvin, C. (2013). Knowledge from data in the built environment. *New York Academy of Sciences 2013 Annals, 1295*(1)*,* 1-9. *The implications of a data-driven built environment.* New York, NY: New York Academy of Sciences.

Strobbe, T., Eloy, S., Pauwels, P., Verstraeten, R., De Meyer, R., & Van Campenhout, J. (2016). A graph-theoretic implementation of the Rabo-de-Bacalhau

transformation grammar. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, *30*, 138 - 158.

Thierauf, R.J. (1999). *Knowledge management systems for business* (1st ed.). Westport, CT: Greenwood Publishing Group.

Timmermans, H. (2016). Design & Decision Support Systems in Architecture and Urban Planning. *Proceedings of the 13th International Conference on Design & Decision Support Systems in Architecture and Urban Planning*.

Tucker, S. & de Souza, C.B. (2016). Thermal simulation outputs: exploring the concept of patterns in design decision-making, *Journal of Building Performance Simulation*, *9*(1), 30-49.

Underwood, J. & Isikdag, U. (2010). *Handbook of Research on Building Information Modeling and Construction Informatics: Concepts and Technologies* (1st ed.). Hershey, PA: Information Science Reference- IGI Publishing.

United Nations (2010). Sustainable Development. [Online] Available at: http://www.un.org/en/ga/president/65/issues/sustdev.shtml

Vahdatpour, A., Amini, N. & Sarrafzadeh, M. (2009). Towards unsupervised activity discovery using multi-dimensional motif detection in time series. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI)* (pp. 1261-1266).

Wang, Z. & Srinivasan, R.S. (2017). A review of artificial intelligence based building energy use prediction: Contrasting the capabilities of single and ensemble prediction models. *Renewable and Sustainable Energy Reviews*, *75*, 796–808.

Wong, J.K.W. & Zhou, J. (2015). Enhancing environmental sustainability over building life cycles through green BIM: a review. *Automation in Construction*, *57*, 156–165.

Wu S. & Clements-Croome D. (2007). Understanding the indoor environment through mining sensory data—A case study. *Energy and Buildings*, *39*, 1183 – 1191.

Xiao, F., & Fan, C. (2014). Data mining in building automation system for improving building operational performance. *Energy and Buildings*, *75*, 109–118.

Xiao, X., Skitmore, M. & Hu, X. (2017). Case-based reasoning and text mining for green building decision making. *Energy Procedia*, *111*, 417 – 425.

Yalcinkaya, M. & Singh, V. (2015). Patterns and trends in Building Information Modeling (BIM) research: A Latent Semantic Analysis. *Automation in Construction, 59*, 68-80.

Yarmohammadi, S., Pourabolghasem, R., Shirazi, A. & Ashuri, B. (2016). A sequential pattern mining approach to extract information from BIM design log files. In *Proceedings of the 33rd International Symposium on Automation and Robotics in Construction* (pp. 174-181).

Yu, Z., Fung, B. & Haghighat, F. (2013). Extracting knowledge from building-related data - A data mining framework. *Building Simulation*, *6*(2), 207-222.

Zanni, M.A., Soetanto, R. & Ruikar, K. (2017). Towards a BIM-enabled sustainable building design process: roles, responsibilities, and requirements. *Architectural Engineering and Design Management*, *13*(2), 101-129.

Zhang, C., Beetz, J., & de Vries, B. (2018). BimSPARQL: Domain-specific functional SPARQL extensions for querying RDF building data. *Semantic Web*, *9*(6), 829-855.

Zhao, J., Lasternas, B., Lam, K.P., Yun, R. & Loftness, V. (2014). Occupant behavior and schedule modeling for building energy simulation through office appliance power consumption data mining. *Energy and Buildings*, *82*, 341-355.

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

Zhou, Q., Zhou, H., Zhu, Y. & Li, T. (2015). Data-driven solutions for building environmental impact assessment. In *Proceedings of the 9th International Conference on Semantic Computing* (pp. 316-319)