Politecnico di Milano
Department of architecture, built environment and construction engineering (ABC)
ABC-PhD Doctoral program

KNOWLEDGE NETWORK FOR INNOVATION OF CONSTRUCTION SECTOR

Increasing efficiency through process digitisation of the entire chain

PhD candidate
Claudio Mirarchi

Supervisor:
Prof. Alberto Pavan – Politecnico di Milano
Tutor:
Prof. Angelo Luigi Camillo Ciribini – Università degli studi di Brescia
The Chair of the Doctoral Program
Prof. Marco Scaioni – Politecnico di Milano

XXXI Cycle
March 2019

Politecnico di Milano
Department of architecture, built environment and construction engineering (ABC)
Via Giuseppe Ponzio 31, 20133 Milano (IT)

To my parents that never limited my dreams
and with their support have made this result possible.
To Silvia that accompanied me in this long journey.

Acknowledgements

This work titled “knowledge network for innovation of construction sector” born from the collaboration between Assimpredil ANCE and Politecnico di Milano. It integrated both the industry and the academy of construction on a journey that allowed me to meet and work with countless people from both sides.

A first and warm acknowledgement goes to my supervisor, Prof. Alberto Pavan, and to the director of Assimpredil ANCE, Gloria Domenighini, to have made possible the beginning and development of this research. Thanks to Alberto Pavan for his continuous support during these years and for the tireless willingness to share ideas and perspectives that have enormously contributed to shaping this work. Thanks to Gloria Domenighini for helping me to gain a different and privileged perspective toward the association and the industry and for paving the way in the work with the construction companies.

Undoubtedly, the people who contributed to this work with suggestions and ideas are countless and it is impossible to thank them all individually. Nevertheless, some deserve a special acknowledgement.

A heartfelt acknowledgement goes to Prof. Arto Kiviniemi who hosted and advised me during the research period at the University of Liverpool. Every single opportunity of discussion has allowed me to grow as a researcher and to improve my work. Thank also for accepting to review my thesis helping me to improve it up to its current form.

An acknowledgement goes to all the research team of the University of Liverpool who made me feel at home during the months of research abroad. Special thanks go to Gulnaz Aksenova who tirelessly reviewed my work diving me useful suggestions to improve it. Thanks also to Sujesh Sujan for the continuous comparison that allowed me to see the work done from new angles.

Thanks to all the Assimpredil ANCE staff who supported me in the many associative activities since the early days.

Thanks to my tutors, Prof. Bruno Daniotti and Prof. Angelo Camillo Luigi Ciribini, for having reviewed and directed my work.

Thanks to Prof. Žiga Turk for accepting to review my work and for pushing me to improve it up to its current form with new ideas and suggestions.

Thanks also to all the construction companies that I met during these years of research for having listened and hosted me. Special thanks go to Giovanni Deleo who has always kept the doors of his company opened to the research and who has listened to me on countless topics.

Thanks to Yang Zou, Jie Sze Fam and Prof. Alice Chang-Richards of the University of Auckland who assisted me in the development of machine learning applications for the organisation of text documents.

A heartfelt thanks also to all the friends and colleagues who supported me during this research and who contributed to making these years special, sometimes helping me to raise my head from the screen and look elsewhere to see better the direction to take.

To all of you, thank you. This work, and what I am, would not be the same without all of you.

Ringraziamenti

Questo lavoro intitolato "knowledge network for the innovation of construction sector" nasce dalla collaborazione tra Assimpredil ANCE e Politecnico di Milano, integrando sia l'industria che l'accademia delle costruzioni in un percorso che mi ha permesso di conoscere e confrontarmi con innumerevoli persone di entrambe le estrazioni.

Un primo e caloroso ringraziamento va al mio supervisor, Prof. Alberto Pavan, ed al direttore generale di Assimpredil ANCE, Dott.ssa Gloria Domenighini, per aver reso possibile l'inizio e lo sviluppo di questa ricerca. Grazie ad Alberto Pavan per il supporto continuo durante questi anni e l'instancabile voglia di condividere idee e prospettive che hanno enormemente contributo a plasmare questo lavoro. Grazie a Gloria Domenighini per avermi aiutato a guadagnare una prospettiva diversa e privilegiata verso l'associazione e l'industria e per avermi aperto la strada verso il rapporto con le imprese.

Indubbiamente le persone che hanno contribuito a questo lavoro con suggerimenti e idee sono innumerevoli ed è impossibile ringraziarle tutte singolarmente in queste pagine. Ciò nonostante, alcune meritano un ringraziamento speciale.

Un sentito ringraziamento va al Prof. Arto Kiviniemi per avermi ospitato e consigliato durante il periodo di ricerca presso l'università di Liverpool. Ogni singola occasione di confronto mi ha permesso di crescere come ricercatore e di migliorare il mio lavoro. Grazie anche per aver accettato di revisionare la mia tesi aiutandomi a migliorarla fino alla sua forma attuale.

Un ringraziamento a tutto il gruppo di ricerca dell'università di Liverpool che mi ha fatto sentire come a casa durante i mesi passati all'estero. Un ringraziamento speciale va a Gulnaz Aksenova che ha instancabilmente revisionato il mio lavoro regalandomi utili consigli per migliorarlo. Grazie anche a Sujesh Sujan per il confronto continuo che mi ha permesso di vedere il lavoro svolto da nuove angolazioni.

Grazie a tutto lo staff di Assimpredil ANCE che mi hanno supportato fin dai primi giorni nelle molteplici attività associative.

Grazie ai miei tutor, Prof. Bruno Daniotti and Prof. Angelo Camillo Luigi Ciribini, per aver revisionato e indirizzato il mio lavoro.

Grazie al Prof. Žiga Turk per aver accettato di revisionare questa tesi e per avermi spinto a migliorarla fino alla sua forma attuale con nuove idee e suggerimenti.

Un ringraziamento va anche tutte le imprese di costruzioni che ho incontrato durante questi anni di ricerca per avermi ascoltato e ospitato. Un ringraziamento speciale va a Giovanni Deleo che ha sempre tenuto aperte le porte della sua azienda alla ricerca e che mi ha ascoltato sui temi più vari.

Grazie a Yang Zou, Jie Sze Fam ed alla Prof. Alice Chang-Richards, dell'università di Auckland, che mi hanno affiancato nello sviluppo delle applicazioni di machine learning per l'organizzazione dei documenti testuali.

Un sentito ringraziamento anche a tutti gli amici e colleghi che mi hanno supportato durante questa ricerca e che hanno contribuito a rendere questi anni speciali, aiutandomi talvolta ad alzare la testa dallo schermo e guardare altrove per vedere meglio la direzione da prendere.

A tutti voi, grazie di cuore. Questo lavoro, e ciò che sono, non sarebbero gli stessi senza di voi.

Abstract

The construction industry has long been characterised as a project-based business that delivers one-of-a-kind products. No two buildings are alike, though they may appear similar from the outside. Hence, design and construction choices are made under uncertainty and the role of designers and constructors is consequently burdened with huge responsibilities about the success or failure of the project and its impact on the environment. Design and construction are processes that run in a cycle way through a trial-and-error procedure because it is not possible thought through the entire picture in advance. They can be considered as iterative procedures composed by a creative stage and a mechanical stage where the hypothesis of the former are tested with reference to project requirements. Moreover, the inherent variables of the construction site that can include machine/human failures, adverse weather conditions, and disadvantageous geological conditions brings the site management to face deviations between the as-planned and as-built construction performance, requiring real-time decisions during the construction development.

In this context, decisions usually rely upon the knowledge of experts that act, however, under uncertainty and with access to a limited amount of information. Several studies in the field of psychology demonstrated the difficulties of human brain in evaluating samples, as people's intuition appears to satisfy the law of small numbers. Hence, a possible precaution to the above-mentioned issues is the use of computation intended as the introduction of techniques and technologies able to support the deciders through data driven suggestions. However, the introduction of digital means in the construction sector demonstrated to be all but easy and linear. Thus, the peculiarities of the construction sector impose the development of social analysis to evaluate possible frameworks able to support the introduction of digital technologies and techniques devoted to improve knowledge management and decision making processes. On the other hand, the need to provide computational analysis in the construction sector impose technological challenges requiring the integration of disparate data generated along the construction chain in a highly fragmented environment. In first instance, this can be interpreted in the need of integrate data, information, and knowledge provided by projects that are not based on a shared rule-set to create the conditions to push the introduction of computation in the construction sector.

Following this focus, the research that is here presented adopts a multidisciplinary approach including construction, knowledge and management, psychology and social science, and computer science areas. Starting from a theoretical background based on the hypertext organisation schema, the cultural historical activity theory, and the heuristic decision theory, this research investigate a novel framework for the development of a knowledge management platform for the construction sector. The theoretical analysis results in an integrated schema that forms the basis for the development of the desired framework prospecting a multi-level approach to develop collaborative environments and related platforms in the construction sector. This framework is supported by the development of dedicated case studies to consolidate the proposed processes. The experimentations are 1) a hybrid classification and clustering approach to organise digital documents from heterogeneous sources and 2) a machine learning image recognition process to verify and semantically enrich building information models. The evidences collected during the literature review and the results obtained in the development of the framework generated two implications. On the one hand, the presentation of an evolutionary view on collaborative environments in the construction sector. On the other hand, the analysis of the possible integration between the existing platforms developed at national level and the proposed framework.

Keywords

Collaborative Digital Platform; Knowledge Sharing; Knowledge generation; Data Management

Contents

List of figures

List of tables

1. Introduction

The construction industry has long been characterised as a project-based business that delivers one-of-a-kind products. No two buildings are alike, though they may appear similar from the outside (Kazi, 2005). Design and construction choices are made under uncertainty and the role of designers and constructors is consequently burdened with huge responsibilities about the success or failure of the project and its impact on the environment (Meadows *et al.*, 1972; Diamonds, 2011). Design and construction are processes that runs in a cycle way through a trial-and-error procedure because it is not possible thought through the entire picture in advance. They can be considered as iterative procedures composed by a creative stage and a mechanical stage where the hypothesis of the former are tested with reference to project requirements (Llach, 2015). Moreover, the inherent variables of the construction site that can include machine/human failures, adverse weather conditions, and disadvantageous geological conditions brings the site management to face deviations between the as-planned and as-built construction performance (Horenburg and Günthner, 2013), requiring real-time decisions during the construction development.

In this context, decisions usually rely upon the knowledge of experts that act under uncertainty and with access to a limited amount of information (Kargul *et al.*, 2015). Several studies in the field of psychology demonstrated the difficulties of human brain in evaluating samples as people's intuition appears to satisfy the law of small numbers[1] (Tversky and Kahneman, 1971). Consequently, subjects tend to overestimate power, i.e. they tend to test and use their hypothesis on small samples without be aware of the high level of risk. They overestimate significance, i.e. they tend to believe in the early results of a test and in its stability during time. They underestimate the breadth of confidence intervals, i.e. they tend to believe in the replicability of the results obtained on small samples (Tversky and Kahneman, 1971). According to Tversky and Kahneman (1971), a decider *"may be willing to regard his statistical intuitions with proper suspicion and replace impression formation by computation whenever possible"*. Proposing as a precaution to the above-mentioned issues the use of computation, i.e. the introduction of techniques and technologies able to support the deciders through data driven suggestions. However, the introduction of digital means in the construction sector demonstrated to be all but easy and linear. For example, several experimentations developed in the field of knowledge management (KM) reported poor results due to the difficulties in understanding the social context of application hindering the diffusion of KM practices (Pathirage, 2007). Looking at building information modelling (BIM), its introduction in the construction sector is still an open area of research even if it started more than 30 years ago.

The peculiarities of the construction sector impose the development of social analysis to evaluate possible frameworks able to support the introduction of digital technologies and techniques devoted to improve knowledge management and decision making processes. On the other hand, the need to provide computational analysis in the construction sector impose technological challenges requiring the integration of disparate data generated along the construction chain in a highly fragmented environment. In first instance, this can be interpreted in the need of integrate data, information, and knowledge provided by projects that are not based on a shared rule-set[2] to create the conditions to push the introduction of computation in the construction sector. Several efforts have been spent in this direction usually in the knowledge management field (chapter 2). However,

[1] *Judgmental bias which occurs when it is assumed that the characteristics of a sample population can be estimated from a small number of observations or data points (WebFinance Inc, 2018b)*

[2] *Data, information and knowledge created during a construction project are affected by several internal and external factors. Among these, one can list international, national and local regulations, project requirements, internal rules of the design teams, etc. Thus, different construction projects usually do not have a shared data structure and/or a shared ontology.*

the study of a domain-based application empowered by the explosion of available digital data seems lacking in the research field of the construction sector.

This situation may create an imbalance between the possibilities offered by the latest developments in artificial intelligence and machine learning applications and the processes applied in the sector. BIM can support the transition toward the use of AI but cannot represent on its own the solution to the above-mentioned issues. There is the need to overcome the boundaries of the project looking to company and domain dimensions where data can be related and aggregated to push the introduction of dedicated services along the idea of man-computer symbiosis[3] (Licklider, 1960).

Nevertheless, the construction process is characterised by high velocity of the information and a high number of possible changes. Thus, the understanding of how to guarantee the velocity and dynamicity required in day-to-day work integrating processes of man-computer design (or decision) support represents one of the focus of the proposed research. The approach that guided the development of the study is about an evolutionary view of this concept integrating the ability of the computer in assisting the designer in real time applications. This kind of paradigm, that is a collaborative man-machine design approach, has been already explored in the research field (Kvan, 2000; Achten and Beetz, 2009; Carrara, Fioravanti and Nanni, 2009; Fioravanti, Novembri and Rossini, 2017). However several issues have been highlighted including the inconsistency of data in BIM tools, the incoherence in the semantic representation, and a still missing relation between data and semantic[4] (Beetz, Van Leeuwen and De Vries, 2006).

The objective of this research is to propose a framework to pave the way for the introduction of the man-computer symbiosis paradigm in the construction sector through a knowledge network platform. The proposed approach starts from a theoretical development that consider and integrate social and psychological aspects related to the introduction of digital technologies and techniques in the construction sector. This main objective will be supported through the development of dedicated case studies to understand the capabilities offered by the developed framework.

1.1. Motivations

The digital transition of the construction sector is dominated by the progressive use of BIM as demonstrated by the increasing efforts in the research field (Yalcinkaya and Singh, 2015) and by its introduction in several national regulations (e.g. Cabinet Office, 2011; Stato Italiano, 2016). However, BIM represents only one of the main aspects of the digital transition and must be considered in coordination with the management of data, information and knowledge that will be generated in new digital ecosystems considering also the interaction between stakeholders. Due to the inherent characteristics of the construction sector that will be explored in the following of this thesis, there are several issues in the effective and efficient comprehension and application of digital processes and instruments. Collaboration and interoperability are critical aspects in this direction and they represent two of the main areas of research in the field. There is a vast literature on collaboration and interoperability and several initiatives that are working in the definition of further technical and technological advancements. However, focusing on BIM, the central point of both research and industry seems following the project centric nature of the sector and its temporal fragmentation. Nevertheless, the explosion in volume and variety of digital data and information and

[3] *The man-computer symbiosis paradigm changes the perception of computers role in the process. Usually, programs are developed to solve pre-defined problems and/or to complete pre-defined processes. Nevertheless, many problems (e.g. the design of a building) are too complex to be thought through in advance. It would be easier to solve these problems through a trial and error procedure with a cooperation between computer and human not only in the solution of the problem but also in its progressive definition. This approach is what can be considered as a man-computer symbiosis*

[4] *These concepts are broadly explored in the thesis (chapter 3.2 and 5.2)*

the increase in the awareness on the capability of advanced systems based on AI are imposing a change in the approach.

The study and implementation of specific solutions focused on specific areas of the process is fundamental due to the variety of competencies and the high specialisation required for the development of construction projects. However, it is required a higher vision on data and information management towards its possible uses to produce value on a domain base. Hence, the need to understand through an holistic approach how to support the construction sector in the integration of knowledge management processes based on digital technologies with the scope of introduce man-machine supportive applications.

1.2. Methodology and methods

1.2.1. Methodology

This research focuses on the development of an information system framework for knowledge management in the construction industry. According to this focus, the Design Science Research (DSR) methodology represent an optimal solution. DSR differs from natural science in the scope of the research. The former focuses on enriching a body of knowledge about the design of an artefact that needs to be useful in the context of application. While the latter, focuses on the explanation of the behaviours and interactions between object of phenomenon, looking at the "discovering of the truth" (Vaishnavi, Kuechler and Petter, 2004).

There is not a commonly accepted reference process model for design science research (Winter, 2008). Nevertheless, Peffers *et al.* (2006) analysed the main existing process proposals defining a design science research process based on six activities, namely 1) problem identification and motivation, 2) objectives of a solution, 3) design and development, 4) demonstration, 5) evaluation, and 6) communication.

Problem identification and motivation – also known as awareness of problem – represents a fundamental step in DSR. In DSR the identified problem should refer to a real-world problem with interest in the practice. Moreover, the identification of a problem must deal with the behavioural-science, i.e. it must consider the implications of external factors (context) on the designed artefact (Hevner *et al.*, 2004). According to these requirements, this work analyses the societal context of the construction sector from the knowledge management point of view (chapter 2) integrating different disciplines in the problem identification stage (Figure 1). DSR differs from routine design according to the innovative value of the problem. This last can be evaluated looking at the maturity of the technologies in the domain of application and at the maturity about the problem domain. Thus, it is required a specific analysis to clarify the context of application from both the technological context and the problem awareness (chapter 3).

The second stage of the methodology focuses on the identification of a possible solution to the problem. This is a creative step where a tentative design is identified paving the way to further developing the design solution (chapter 4.1).

Design and development represent the core activities of the methodology where the solution is defined. The third stage requires a more formal representation of the solution compared to stage two and a description of the components of the solution as well as of the internal and external factors that may affect the solution itself. This step is inherently combined with stages four and five where the performance (effectiveness and utility) of the solution is demonstrated and evaluated. According to the type of artefact the evaluation means can change including case studies, qualitative evaluation in the context, etc. (chapters 4.2, 4.3, 5 and 6).

The last stage of the methodology is the communication of the solution. This thesis represent the main means to communicate the results of the work. Nevertheless, the author developed several publications related to the work (chapter 7.3) and integrated part of the results in a proposal presented at European level (chapter 6.3).

1.2.2. Methods

To approach the data, information and knowledge collection, management, maintenance and use were identified several means explored in this thesis. The broad area of knowledge and knowledge management was explored through an extensive literature review focused on the construction sector. This analysis was enriched with the presentation of techniques and technologies in the digital area devoted to manage and generate knowledge such as support decision making processes and knowledge discovery in databases techniques. Moreover, to understand the dynamics of knowledge the literature review includes a study of the main related theories in the field of psychology and social science. Figure 1 represents the main arguments that compose the research source of knowledge, where has been identified four main fields, i.e. construction, knowledge and management, psychology and social science, and computer science. In construction field, the theoretical analysis of the existing studies has been joined with the analysis of real world data provided by several stakeholders, i.e. construction companies, engineering companies, public administrations, clients, and manufacturers. Moreover, the development of the study was supported by direct experiences matured in the collaboration with the association of construction companies of Milan, Monza e Brianza and Lodi (Assimpredil ANCE).

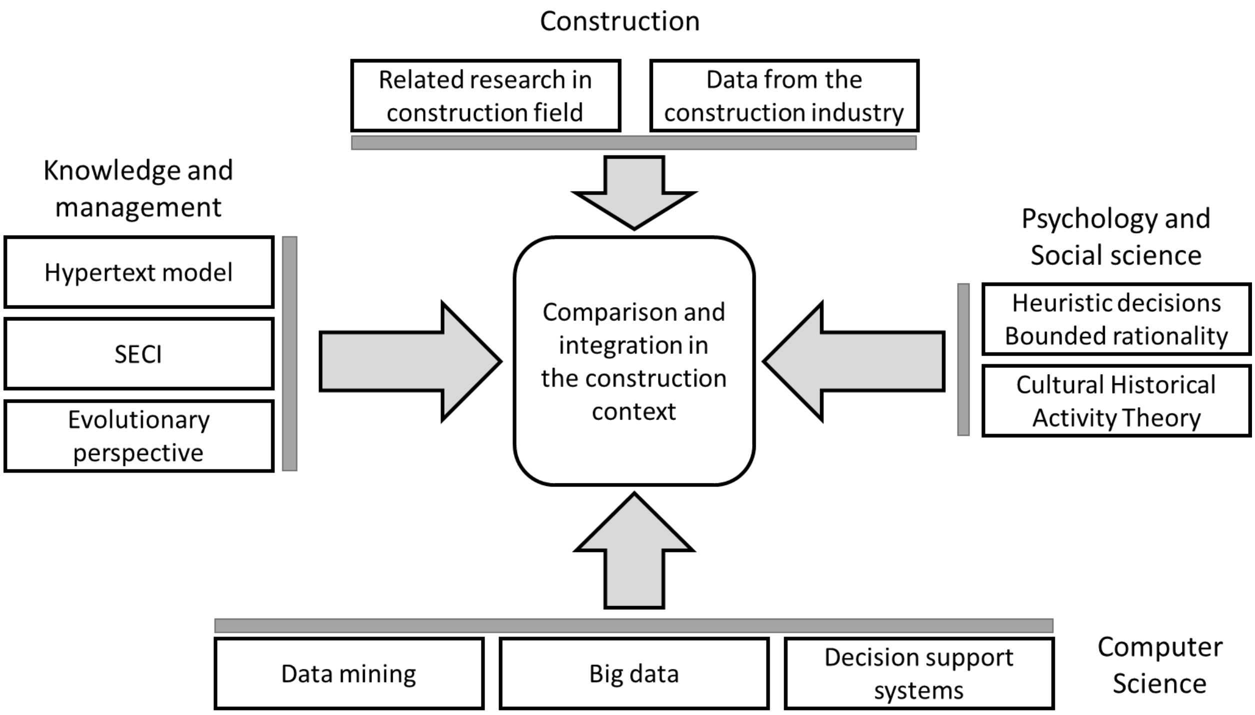


Figure 1. Epistemological map of the research

As a source of creation of digital data and information and a spreading mean in data and information management, BIM is analysed in this research and included as one of the fundamental components of the proposed results. In this context, the study of collaboration, interoperability and data management is a natural consequence to explore actual means and possibilities offered by the technology in these areas.

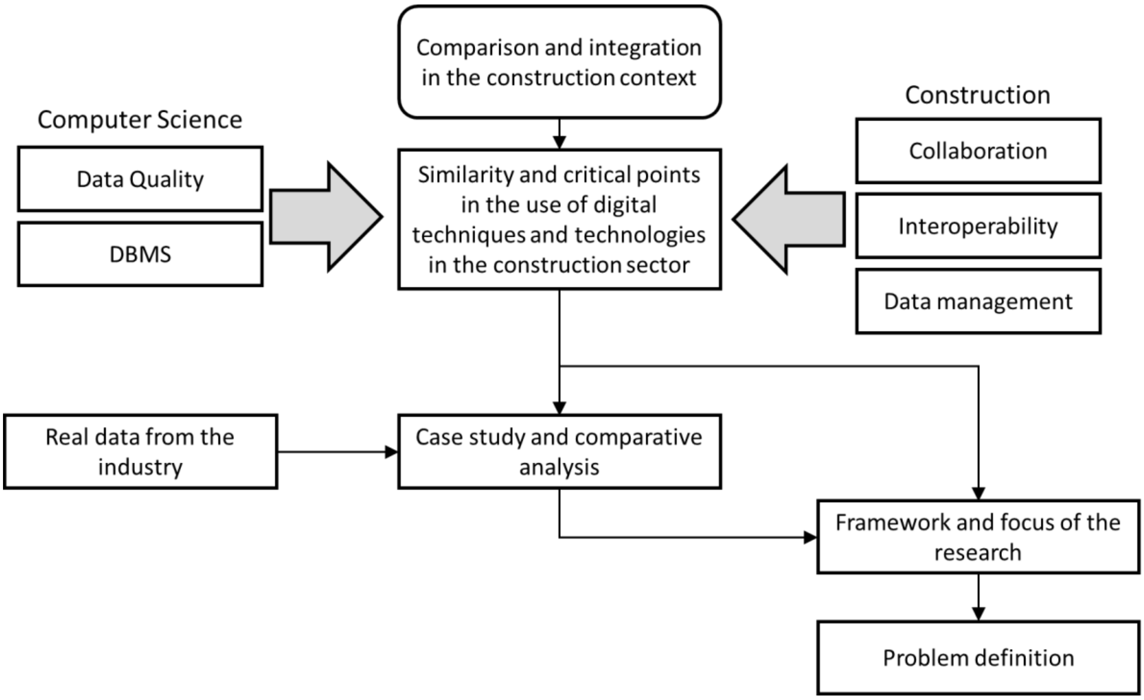


Figure 2. Methods and research path for problem definition

Starting from the definition of the state of the art in accordance to the above-described methodology, it was possible proceed with the development of the main objectives of the work, i.e. the study of the framework for a knowledge network for the construction sector interpreted as an inclusive platform as extensively described in this thesis. Figure 2 shows the methods and research path applied for the problem definition. In particular, defined the state of the art, the existing technologies, techniques and processes used in the construction field in the areas of collaboration, interoperability and data management has been studied and compared to the main related themes in data science, i.e. data quality and database manage systems (DBMS). To evaluate the impact of the highlighted similarities and issues is proposed a case study developed using real world data from the industry. The background analysis provided the basis to develop the study of the framework starting from its high-level definition (objectives of a solution). Figure 3 proposes the path of the analysis. In particular, from the high level framework it was possible defining the requirements of each part of the framework combining it with the literature review and with the observations of the existing practices in the industry (this last limited to the Italian context). Combining the high level framework and the requirements analysis it was possible to define the system framework.

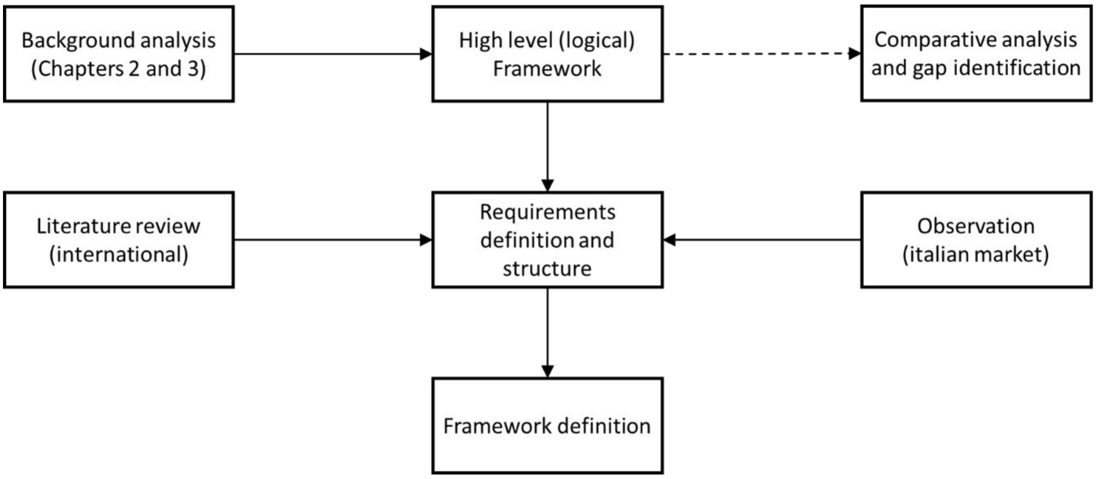


Figure 3. Analysis path for the definition of the system framework (from the tentative design to the design and development stage)

Developed the system framework, it was possible defining specific applications and case studies to validate the perspective implicitly included in it. Figure 4 represents the general path followed in the definition and development of each case study. In particular, starting from the proposed framework and from the study of the existing issues revealed in the literature review, can be identified some possible solutions. These last can be translated in applicative frameworks related to the specific solutions and compared to the existing studies in the literature to individuate possible improvements. Starting from the framework, a prototype of the application can be proposed to test and validate the individuated framework concluding the case study development. In some cases, the proposed frameworks bring to the identification of critical areas that need further developments where the inclusion of specific algorithms for the automation of exchange processes is foreseen and/or where the state of the art in the development of specific algorithms is not sufficient to guarantee the required performance.

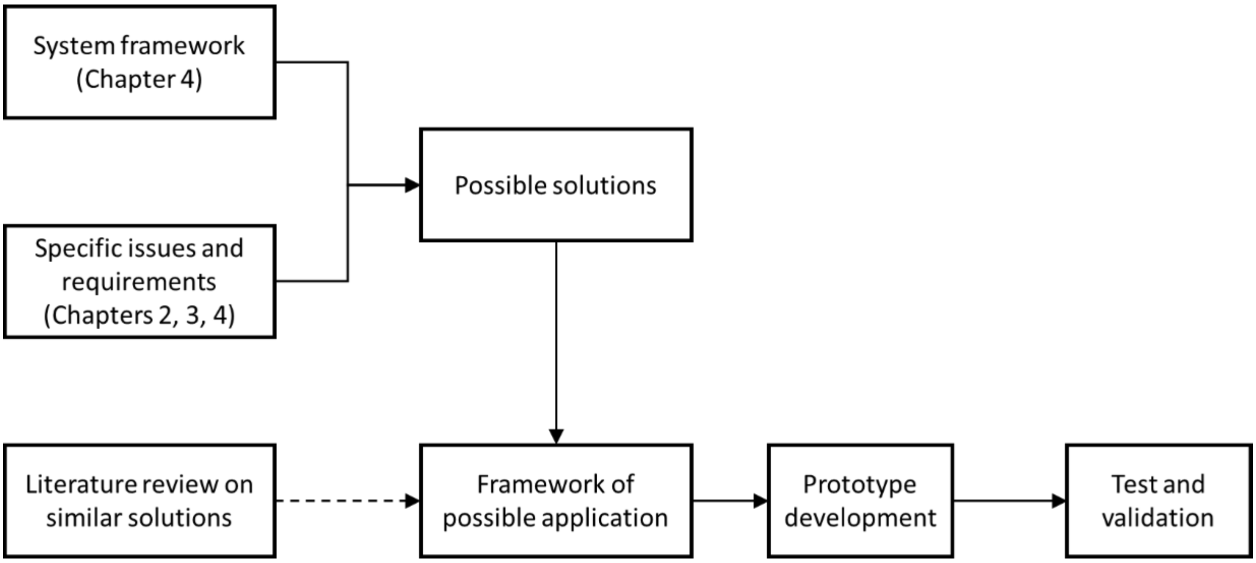


Figure 4. Case studies identification and development

In synthesis, one can list the main activities of the research:

- Study and identification of the specific areas of interest in knowledge and knowledge management themes (chapter 2).
- Study of the existing means for collaboration, interoperability and data management in construction sector with specific reference to BIM (chapter 3).
- Identification of open issues and limitations in the state of the art with reference to the stated research questions (chapters 3.2, 3.3, 4.1).
- Comparison and validation of the evidences identified in the literature review through direct interaction with professionals, construction companies, and other stakeholders in the sector (chapter 4.2).
- Study of the framework of the knowledge network for the construction sector (chapters 4 and 6.2).
- Identification and study of critical parts of the framework for the inclusion of automated data and information processing procedure based on AI and machine learning[5] (chapter 5).
- Study of possible implication of the proposed framework and processes in providing domain-based knowledge (chapter 6).

[5] *This step includes the development of two case studies according to the process described in Figure 4. In detail, the two case studies are focused 1) on the management of text documents through a novel machine learning approach and 2) on the semantic enrichment of building information models through an image recognition process.*

1.3. Research questions

This research is based on four main research questions:

- How can we collect, share and generate knowledge with value in the construction sector domain?
- How can we collect and parse data and information provided by different sources along the construction process, addressing its spatiotemporal fragmentation and its long time distribution?
- How can we guide the construction sector in provide such data and information?
- How can we maximise the value of data and information collected and how can we use these assets in the generation of knowledge?

These four research questions define a broad area of research that have to be focused according to the objectives of the work, i.e. the development of a knowledge network platform in the construction domain. The questions follow the research methodology defining a progressive construction of the project solutions. Question one deals with the identification of the problem and is explored in chapters 2 and 3. Question two starts from the output of the former and deals with the development of the desired solution (chapters 4 and 5). Finally, questions three and four focus on the evaluation of the solution in the context of application (chapter 6).

Even if many of the analysis proposed in this research are based on a holistic view oriented to the entire process, the focus of the case studies and of the experimentations proposed is on the design and construction phases of the construction process. Moreover, the research is focused on the development of the process and do not address the possible issues related to the preparatory activities of the process itself such as contractual and legal issues.

1.4. Assumptions and clarifications

1.4.1. Assumptions

This research is based on the assumption that the construction sector is moving through a progressive transition that includes the introduction of digital technologies and techniques in its processes[6]. This assumption implies the increasing production of data and information in digital format that represent a critical point for the effective application of the proposed framework.

1.4.2. Clarifications

Digital Platform

This research includes as one of its main objectives the definition of the framework for a central platform for the construction sector (see chapter 4). However, the term “platform” nowadays is widely used and it includes different meanings. In the recent policy activities about the digital innovation related to the idea of Industry 4.0 a huge number of platforms have been developed at national level in the European context (Figure 5). In this context, the term platform refers to web spaces used to share and promote discussions, governance coordination, and national best practices. The main functions of these kind of platforms is very close to a public forum (Working Group 2, 2017), that does not represent the focus of this research. On the other hand, at European level policy activities are looking at the development of industrial digital platforms focusing e.g. on manufacturing, agriculture, and healthcare.

[6] *Looking at the evolution of the industry, this assumption may be perceived as trivial or self-evident. Nevertheless, the introduction of digital processes and techniques in the construction industry demonstrated to be a slow process with different breaks according to the market evolution. Hence the need of this preliminary assumption.*

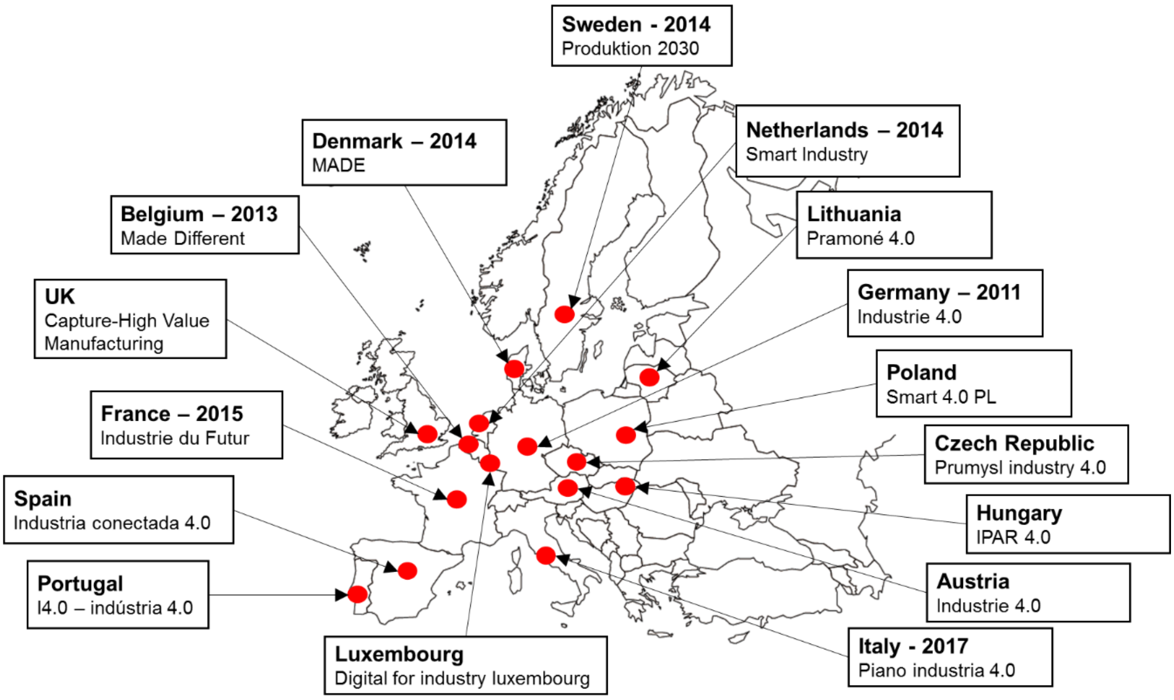


Figure 5. Map of the main platforms developed in the European context about Industry 4.0

In this research, the term "platform" does not follow the type above described. The specific focus is on the definition of a data platform framework[7] which use can promote the integration of data and information along the construction process limiting their spatiotemporal fragmentation. Hence, in this research the term "platform" is more close to a warehouse (Devlin, 1997; Nemati *et al.*, 2002) even if it is not limited to it. In fact, the platform framework that will be presented in this thesis can be interpreted from two perspective. On the one hand, the platform can be viewed as an environment[8] where data and information are collected and parsed according to defined rules and processes. In this case, the area of collaboration is not directly addressed. On the other hand, the platform can be viewed as the entire digital ecosystem where both the collaborative spaces and the spaces devoted to data management and analysis, knowledge management, etc. as well as the processes that define the passages from one space to another are integrated. In this thesis, the term collaborative platform (except in the cases where existing platforms are presented) is intended along this second interpretation.

Building Information Model

A building information model is defined in international standards as *"shared digital representation of physical and functional characteristics of any built object [...] which forms a reliable basis for decisions"* (ISO, 2016). Building information models can represent real buildings virtually over the whole life cycle (Eastman *et al.*, 2011; Watson, 2011). They are developed using object-oriented tools and consists of semantic-rich parametric objects including geometric and non-geometric information (Wong and Yang, 2010; Cerovsek, 2011; Nicolle and Cruz, 2011).

A building information model can be seen in a narrow and a broader perspective (Jernigan, 2007; Volk, Stengel and Schultmann, 2014). In a narrow sense, it is focused on the digital building

[7] As explained in the thesis (chapters 3.1, 4) this concept is broader than the concept of common data environment. The latter focuses on the project level while the former focuses on the sector level (including the project level).

[8] This do not means that the environment is represented by a single physical space. It can be interpreted as a distributed system where different spaces are related and linked.

model itself, interpreted as a central information management hub centered on technical information and model creation issues. In this vision spatial model with quantity take off (3D), construction scheduling (4D), and cost calculation (5D) are included (Cerovsek, 2011).

However, in an AEC project can be defined two main data categories, namely product data – or project engineering data (PED) – and process data – or project management data (PMD). These two categories are defined in (Jiao *et al.*, 2013) as follow:

- *PED are geometric presentations, parametric descriptions and legal regulations associated with the construction of a building, such as component position/layer/level, bill, and quota.*
- *PMD refer to control and communication information that are generated in and closely related to management activities throughout the construction life cycle, such as scheduling, monitoring, and work assignment.*

Hence, in a broader sense, building information model includes functional, informational, technical and organisational/legal issues dealing with both PED and PMD data category. Moreover, the needs of multiple model views[9] requires the inclusion of information provided from different sources and combined according to the specific user needs (Della Torre, Mirarchi and Pavan, 2017). Thus, a broader interpretation of building information model focuses on the entire digital environment where all the information is managed.

In the development of the research both these visions are included. In fact, many applications and experimentations are based on single building information models used as a tester. On the other hand, the framework proposed in chapter 4.3 refers to a broader vision of building information model.

1.5. Guide to read this thesis

The rest of the thesis is organised as follows:

Chapter 2 introduces the background about knowledge and knowledge management areas. Chapter 2.1 starts from the study of knowledge nature and of the literature review in knowledge management proposing at the end an integrated schema including the hypertext organisation schema, the cultural historical activity theory and the heuristic decision theory. This last is integrated with a spatio-temporal perspective on knowledge management in the construction sector. Chapter 2.2 introduces the technological possibilities and implications related to the management of knowledge in digital environments dealing with the area of knowledge discovery in databases (KDD) and decision support systems (DSS).

Chapter 3 introduces the concepts of collaboration, interoperability and data management in the construction sector. This chapter born from the identification of the main drivers of innovation in the construction sector that see the theme of collaboration at the centre of the debate. In addition to the analysis of the literature review regarding collaboration and interoperability, chapter 3.3 proposes a case study focused on data quality in the construction sector with particular reference to building information models.

Chapters 2 and 3, as background chapters, are completed by a “conclusion and focus of the research” subchapter that aims to help the reader in maintaining the focus on the project and explains the reasons behind the choice to present the proposed arguments.

Chapter 4 presents the path followed to develop the platform framework proposed in this chapter as one of the main results of the present research. The chapter comprises an analysis of the

---

[9] *Information models can contain information provided by different sources and related to different domains. Nevertheless, each stakeholder involved in the process needs only a subset of the information contained in a model, i.e. he or she needs a specific view of the model. Hence, from one model can be generated multiple model views according to the need of the subject that is using the model.*

requirements for collaboration environments and digital platforms identified in the literature and integrate these requirements with the theoretical schema proposed in chapter 2.1.

Chapter 5 proposes two experimentations developed with the aim of consolidate the processes described in the development of the platform framework. Chapter 5.1 presents a hybrid classification and clustering approach to organise heterogeneous datasets of digital documents paving the way for the development of dedicated applications and services. Chapter 5.2 presents machine learning image recognition process devoted to semantic enrichment of building information models with the objective of support the translation activities between the different levels identified in the framework.

Chapter 6 studies the implications derived from the proposed framework. This chapter presents on the one hand a possible evolutionary perspective on collaboration environments including the forecasts derived from the application of the platform framework. On the other hand, it proposes a comparative study between the existing national platforms and the proposed framework to underline how to practically apply the presented processes.

Finally, chapter 7 proposes the summary of the work highlighting the originality of the obtained results, the possible drawbacks, open issues and future research and illustrating the dissemination activities derived from this thesis.

All chapters, except for chapters 1 and 7, start with an aphorism that tries to represent the vision of the author in the development of the chapter's contents. The objective is to bring the reader near the perspective of the author defining a shared context that can facilitate the interpretation of each chapter. We are as dwarfs standing on the shoulders of giants (Mertone, 1965) and these starting points allow both the reader and the author to sit on the same shoulder and look together through the text.

Moreover, chapters 2 and 3 start with a paragraph (underlined using italic style) that explains the roles of the chapter in the overall thesis helping the reader in focusing on the objective of the work.

2. Knowledge and knowledge management

"Images of the future are shaped by the experience of the past."
(Tversky and Kahneman, 1973)

This chapter contains the theoretical basis to understand the societal context where the research is developed. It represents a fundamental contribution to the work addressing the societal dimension of knowledge management in the construction sector. Moreover, the first part of the chapter clarifies the concept of knowledge facilitating its comprehension in the future sections of the work. On the other hand, the second part of the chapter explains the technological context where the work is developed. It provides the instruments to understand and support the perspectives presented in chapters 4 and 6.

Dave and Koskela (2009, p. 896) stated that "*construction is an informative intensive industry where stakeholders communicate a large amount of information across various stages of the project lifecycle*". In every project, during each phase are generated hands-on experiences, problem solving capabilities, understanding of various means and methods and highly contextualised solutions, especially during the construction phase (Lin *et al.*, 2005). This knowledge represents one of the most important assets for architecture, engineering, construction, and operations (AECO) firms (Deshpande, Azhar and Amireddy, 2014).

In the construction process, the development of a project requires the aggregation of several stakeholders. Many times this group of stakeholders collaborate for the development of the project and once it is delivered, they disband moving on to the next project (Deshpande, Azhar and Amireddy, 2014). The temporariness and fragmentation characteristics of the construction sector lead to its identification as "*the temporary multiple organization*" (Cherns and Bryant, 1984) or virtual enterprise[10] (Kazi and Koivuniemi, 2006), and to the development of specific studies on its dynamics (Charbuck and Young, 1992; Browne, Sackett and Wortmann, 1994; Afsarmanesh *et al.*, 1997). This fragmented work environment creates a double issue. On the one hand, the continuous changes in the partners configuration hinders the creation of structured communicative processes with the consequence of possible information inconsistencies and regulatory compliance problems (Rezgui, Hopfe and Vorakulpipat, 2010). On the other hand, the knowledge generated during the process by a whole of stakeholders is disrupted at the end of the project due to the difficulties in creating an organisational knowledge. The experiences gained in the process are rarely, if at all documented with the consequence that knowledge remains stored in the minds of those who were directly involved (Kazi and Koivuniemi, 2006). Hence, the inability of document, store and use in a systematic way data, information and experiences (knowledge) can lead to the loose of most of their value. Nevertheless, the experiences of the individual may not capture the overview about the context where were generated losing in effectiveness. Fantino and Stolarz-Fantino (2005) highlighted the importance of context in decision-making processes. Thus, in the construction sector where the context of work changes dynamically, the theme of knowledge capture and re-use is still an area of open debate.

Due to the recognised importance of knowledge, the broad theme of Knowledge Management (KM) has gained a central role in the research filed and in the industry. There is a vast literature about theories and applications in different sectors. Among these Troxler and Lauche, (2006), and Wolf and Kazi, (2006) discuss the implementation of communities of practices respectively in three

---

[10] *"A virtual enterprise consists of a series of co-operating 'nodes' of core competence which form into a supply chain in order to address a specific opportunity in the market place" (Walton and Whicker, 1996).*

companies in the oil and gas industry and in the automotive industry. Dayan, Pasher and Dvir (2006) present the journey of an aircraft industry in the introduction of knowledge management, discussing the integration of several areas in complex organisations using competence centres. Blignaut (2006) proposes the use of community social network analysis and Cynefin categorisation for the identification of data silos in a large financial institution and the study of the critical areas in the company. Maree, Roux and Marais (2006) deal with the use of narrative techniques for the exploration of social processes in a R&D organisation and its use for the definition of change processes in the knowledge management field. Eidal, Hjorth and Kadal (2006) present the application of business modelling techniques for the development and the communication of effective knowledge management solutions in a maritime industry, dealing with both the ICT perspective and the intellectual capital theory. Grillitsch and Muller-Stingl (2006) deal with the effective introduction of Knowledge Management in project-oriented organisations discussing the use of project supervision meetings for cross-project knowledge transfer. Ruffino (2006) presents the results of a practical tool tested on a group of SMEs with the goal of supporting knowledge transmission and highlighting existing critical factors. Lewis and Diamond (2006) discuss the implementation of Knowledge Management in organisation dealing with critical incidents, referring to a practical case study about firefighters.

This brief list clarifies the wide range of fields and contexts where KM has been studied and applied in both research and industry. A systematic study of the applications of knowledge management in different sectors can be identified as a future and interesting line of research to understand possible points of knowledge transfer. Nevertheless, in the context of this research it would be too vast with the risk of concealing the main objectives of the work.

Despite the long debate on knowledge management, there is still no consensus about what it means. Some definitions are reported below.

- KM is the discipline of creating a thriving work and learning environment that fosters the continuous creation, aggregation, use and re-use of both organisational and personal knowledge in the pursuit of new business value (Cross, 1998).
- KM is *"any process or practice of creating, acquiring, capturing, sharing and using knowledge, wherever it resides to enhance learning and performance in organisations"* (Scarborough, Swan and Preston, 1998).
- *"Knowledge management is a process of acquiring, creating, sharing, utilizing and storing intellectual assets and other stimuli from the internal and external business environment that facilitates an organization to perform successfully"* (Kululanga and McCaffer, 2001).

However, all the definitions proposed in the literature highlight the continuity of KM, i.e. its development along the life of an organisation with a circular process. According to Lin and Tserng (2003) the construction KM life cycle can be divided in five phases, namely knowledge acquisition, knowledge extraction, knowledge storage, knowledge sharing and knowledge update. While Gupta and McDaniel (2002) identify five components for effective knowledge management, namely harvesting, filtering, configuration, dissemination and application. Hence, the study of KM technologies and techniques requires the study of different phases addressing the issues that arise in each one and in the relation between the different phases.

This chapter focuses on the definition and applications of knowledge and Knowledge Management in the construction sector. Starting from a general introduction about knowledge and its main characteristics, the chapter proposes a vision about knowledge in the construction sector. A literature review about KM applications in the construction industry is proposed, followed by the study of how knowledge management system can impact on decision making processes. According to the increasing production and availability of digital data and information, the chapter includes a brief introduction to Knowledge Discovery in Databases (KDD) clarifying the need of these techniques, the processes of data-to-knowledge transformation in databases manageable knowledge and the technological implications.

2.1. Knowledge dimensions and applications

There is not a unique definition of knowledge. Can be identified many areas, aspects and kinds of knowledge, including for this last inductive, a priori, perceptual, self, testimonial, memory, semantic, scientific, logical and mathematical, aesthetic, moral and religious (Bernecker and Pritchard, 2011). Studies related to the definition of knowledge are present as early as the ancient Greeks. In more recent years Lundvall and Johnson (1994) identified four different types of knowledge that can be used to define what knowledge concerns:

- Know-what
- Know-why
- Know-how
- Know-who

Know-what concerns the possession of information that is the knowledge of the facts. It refers to the information that can be transmitted through data and disseminated with the help of databases.

Know-why refers to the principles and the laws that govern nature, human mind and society. It is the theoretical knowledge that constitute the basis of the scientific and technological research. This type of knowledge allows the innovation of products and processes and reduce the frequency of procedural errors.

Know-how is related to the operative experience of individuals or group of workers. It represents the human capital of a company.

Finally, *know-who* concerns the knowledge about people able to carry out specific works and/or able to solve specific or complex problems.

In the literature, there are several classifications of knowledge, among which personal; shared and public; practical and theoretical; hard and soft; internal and external; foreground and background. However, the most used is the classification in tacit and explicit knowledge adopted by Nonaka and Takeuchi (1995). Tacit knowledge represents knowledge based on the experience of individuals, expressed in human actions in the form of evaluation, attitudes, points of view, commitments and motivations. This knowledge can be linked to know-how and know-who. Explicit knowledge is codifiable knowledge inherent in non-human storehouse including organisational manuals, documents and databases. This knowledge can be linked to know-what and know-why. An effective comparison between tacit and explicit knowledge is proposed by Nonaka and Takeuchi (1995) and reported in Table 1.

Table 1. Comparison between tacit and explicit knowledge by Nonaka and Takeuchi (1995)

Tacit Knowledge	Explicit Knowledge
Experiential knowledge (corporeal)	Rational knowledge (mental)
Simultaneous knowledge (here and now)	Sequential knowledge (there and then)
Analogic knowledge (practical)	Digital knowledge (theoretical)

Understanding the relation between tacit and explicit knowledge is fundamental to identify the proper conversion and transfer mechanisms (Section 2.1.1).

Focusing on the perception of knowledge in organisations, according to Alavi and Leidner (2001), knowledge may be viewed from five different perspectives as listed below.

- *Object perspective:* defining knowledge as a thing to be stored and manipulated.
- *Condition perspective:* emphasising access to knowledge.
- *State of mind perspective:* emphasising knowing and understanding through experience and study.
- *Process perspective:* focusing on knowing and acting.
- *Capability perspective:* viewing knowledge as a capability with the potential of influencing future action.

The object perspective focuses on the need of capture, store and manage knowledge. Both explicit and tacit knowledge need to be considered in this perspective (Fernie *et al.*, 2001).

Condition perspective emphasises the accessibility to knowledge. In this context, it is important distinguish between accessibility and availability. While the former implies the latter, this is not true in the other way round. In fact, knowledge can be available but difficult to access because its use requires complex instruments or procedure.

State of mind perspective highlights the needs of "digest" knowledge by subjects through experience and study. This bring to an important aspect of knowledge that is its association to the context in which is created and used. Nonaka and Takeuchi (1995) stated that information lost sense if it is abstracted from the associated emotion and specific context in which it is embedded. Thus, knowledge creation and sharing is constrained by social relations and it is context dependent (Fernie *et al.*, 2001). Moreover, its perception depends on the previous knowledge of the subject involved and it is influenced by the availability of previous information (Tversky and Kahneman, 1973).

Process perspective underlines the importance of action from knowledge. The use of knowledge requires human action to gain power in the process (Kamaruzzaman *et al.*, 2016).

Finally, the capability perspective focuses on the use of knowledge for the definition of future actions bringing to the areas of knowledge based decision-making processes in business.

Rezgui, Hopfe and Vorakulpipat (2010) used these perspectives to characterise an evolutionary map of knowledge management. In their work they defined three generation of knowledge management, that are:

- Level 1, knowledge sharing.
- Level 2, knowledge conceptualisation and nurturing.
- Level 3, knowledge value creation.

The evolution between the three levels is related both to the increasing capabilities of IT solutions and to the change of the management philosophies, that is the change in processes and social aspects. Nevertheless, the driver of change from one generation to the following one, is the increased consciousness of organisational and social aspects (second generation) and of social, human and intellectual capital as an asset (third generation).

The first level is perceived as the optimisation of information management activities, paving the way for future knowledge management aspects. The focus is on an intra-organisational dimension. At this stage, the objective is the optimal organisation of documents interpreted as information and knowledge carriers. However, the interpretation of documents content is still reserved to individuals. It must be noted that document sharing does not represent a direct transmission of knowledge between two subjects. According to Davenport and Prusak (1998) knowledge sharing includes the transmission process and the absorption process, i.e. the acquisition of knowledge in the individual's mind and its use. Hence the dynamic transformation of knowledge during time and the continuous process of knowledge creation.

The second level starts with the advent of firsts object oriented tools requiring an increasing change both in technologies and in processes and social aspects. The introduction of shared semantic structures for the definition of geometrical and non-geometrical information and the development of coordinated standards (chapter 3.1) allowed the optimisation of the inter-company dimension.

However, tacit knowledge cannot be handled through these technologies and techniques. Thus, in parallel with the development of object-oriented tools, the second generation of knowledge management includes the first attempts of social-based initiatives such as community of practices and discussion forums.

Finally, the third level represents the expected evolution of KM towards the understanding of the knowledge value perceived in a domain dimension. The increasing consciousness on social responsibility such as environmental and societal sustainability, health and safety of projects and corporate social responsibility of organisations underlines the progressive understanding of knowledge value at domain level. However, despite the technological and technical developments in the knowledge management field, Rezgui, Hopfe and Vorakulpipat (2010) argued that knowledge identification, capture, manage and reuse remain open issues that can hinder the processes of the third generation. Nevertheless, it can be argued that these issues are constant in all the levels described due to the continuous creation and co-creation of knowledge. They represent crucial points of analysis that must be developed in the specific context of application.

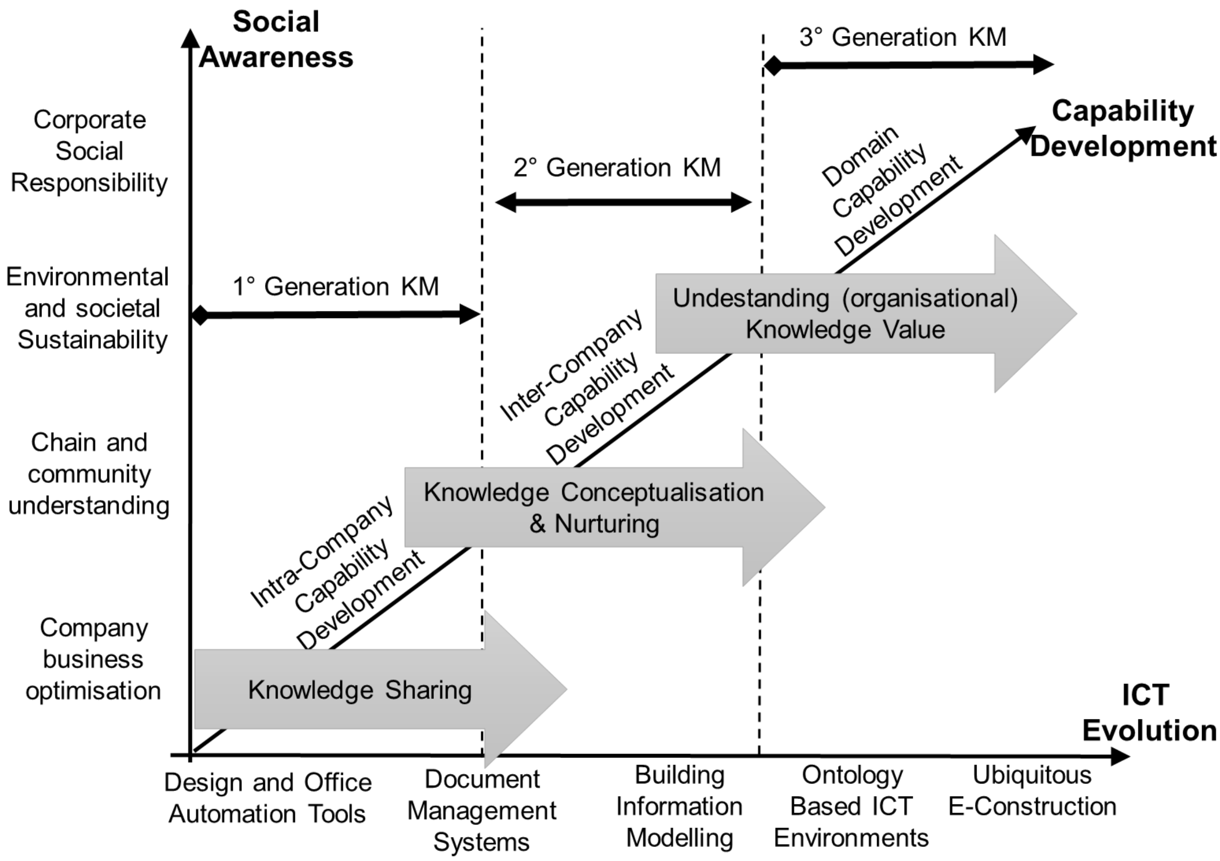


Figure 6. Generations of knowledge management (modified by author from Rezgui, Hopfe and Vorakulpipat, 2010)

Rezgui, Hopfe and Vorakulpipat (2010) proposed a graphical representation of the knowledge management generations. Figure 6 proposes a version of that representation modified by the author according to the context interpretation matured during the development of this thesis. The vertical axis originally representative of the management philosophies has been changed introducing the concept of social awareness that helps underlining the societal and procedural changes intrinsic in the evolution of knowledge management practices. According to Dourish and Bellotti (1992), social awareness can be defined as "*the understanding of the activity of the others, which provides a context of your own activity*". Social awareness represents a crucial prerequisite for good collaboration and its evolution can be supported by the use of digital technologies (Tollmar, Sandor and Schömer, 1996).

Thus, the vertical axis starts from a social awareness related to a single specific business (an organisation) focusing on a progressive understanding of intra-company relations isolated from the external environment. Social awareness then moves towards the understanding of the activities of subjects outside of the company and the impact of the activities of one company on the other companies involved e.g. in one project. Finally, the vertical axis indicates the progressive understanding of the impacts of the activity of one company to the society and environment promoting the concept of environmental and societal sustainability and of corporate social responsibility.

It must be noted that the chronological evolution of ICT technologies and techniques defined through the horizontal axis is not sequential. Technologies identified in the first generation are still under development and their progressive development impacts on the evolution of the newest ICT tools. Moreover, single organisations can adopt these technologies following a different path. However, the introduction of tools and techniques must be evaluated in terms of societal and relational consciousness to define the possible impact on knowledge management processes.

This evolutionary perspective identifies the connection between the management of data, information and knowledge and the corresponding required evolution of knowledge perception in different contexts of application (intra-company, inter-company, domain). The proposed representation can help in the definition of the research direction to promote the transition towards a possible domain-based generation of knowledge management.

2.1.1. Knowledge creation, transfer and conversion

"Sooner or later, every organization ends up creating knowledge" (Nonaka and Takeuchi, 1995). However, the identification of where the knowledge is created, how to extract, store, share and update this knowledge, who are the subjects involved, why do they learn, what do they learn, and how do they learn represent critical questions related to the specific environment where the learning process is introduces (Engeström, 2001).

According to Nonaka and Takeuchi (1995), the knowledge creation needs to follow a virtuous spiral moving back and forth between different mechanisms of knowledge conversion. In the middle-up-down approach proposed by Nonaka and Takeuchi (1995) it is highlighted the importance of a knowledge conversion layer between field operators, designers and top managers. In fact, each one of these subjects has a different perception of the work and consequently a different knowledge construction in his or her mind.

Figure 7 shows the "hypertext organisation schema" where the continuous conversion of knowledge between different levels of an organisation is highlighted. The knowledge base represents the place where the knowledge can reside and it can be categorised and contextualised into a more meaningful product for the organisation as a whole. The individuals acting in the team project level can acquire and interpret the knowledge derived from the business system level in a completely different manner in comparison to the individuals acting on this last level. In the same way, the knowledge created in the project level is interpreted in a different context at the business system level acquiring a different value. The mobility of the knowledge and its management in the knowledge base area allow the activation of a virtuous circle of knowledge conversion paving the way for the creation of organisational knowledge valuable for the whole organisation.

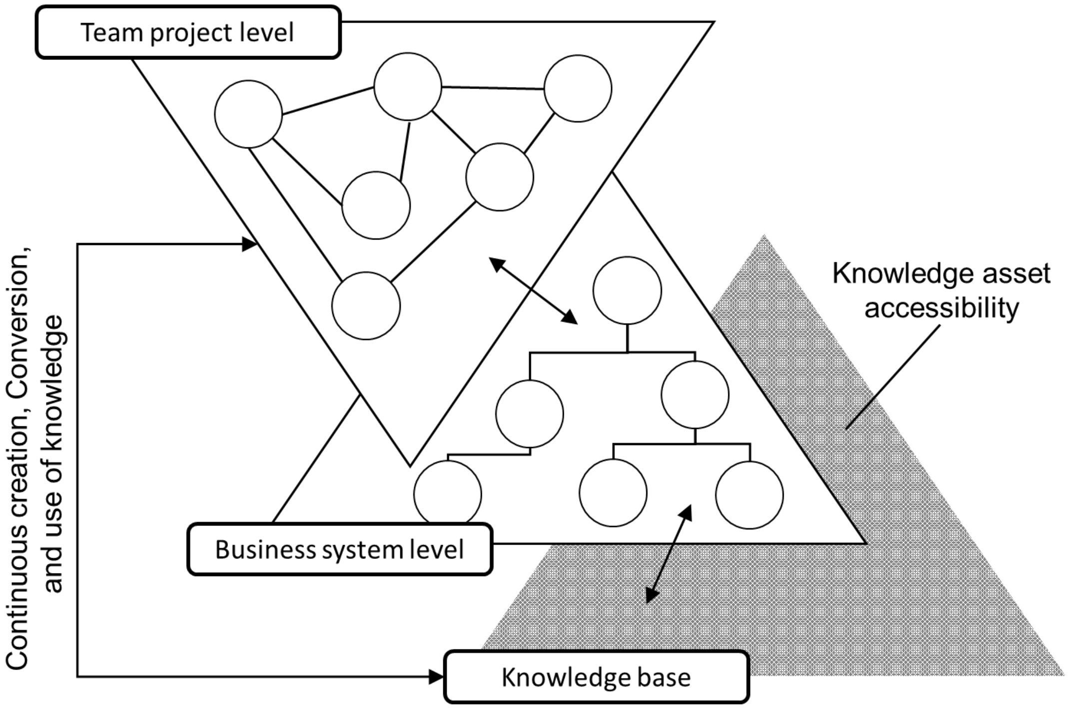


Figure 7. The hypertext organization schema by Nonaka and Takeuchi (1995)

The most popular model in terms of knowledge conversion (i.e. how subjects learn) is the SECI model (Nonaka and Takeuchi, 1995; Skyrme and Amidon, 1997; Nonaka, Toyama and Byosière, 2001). It outlines four mechanisms of knowledge conversion and/or transfer:

- *Socialisation:* tacit-tacit.
- *Externalisation:* tacit-explicit.
- *Combination:* explicit-explicit.
- *Internalisation:* explicit-tacit.

Socialisation is based on direct interaction between individuals that share their experiences simultaneously, usually with analogical and practical means.

Externalisation requires the identification of means to "translate" experiences in a codifiable way to allow future uses. The conversion realised in a socialisation process is different from the one obtained through an externalisation-internalisation process because there is a change in time and context of knowledge usage. In fact, when tacit knowledge is shared, it still requires to be decoded by individuals (Bolisani and Scarso, 1999) and this passage is constrained by the context of interpretation.

In combination, individuals share and combine knowledge through different means including documents, meetings, and computer networks. The reconfiguration of information sorting, adding, combining and categorising explicit knowledge can produce new forms of knowledge. However, highlighting the human factor in the knowledge conversion process, Roos *et al.* (1998) stated that combination mechanism cannot really exist, because in this form it is a simple transfer of data and/or information without a real involvement of knowledge. For admit this mechanism, an explicit-tacit-explicit mechanism is required.

Internalisation implies the conversion of the experiences gained through socialisation, externalisation and combination in shared mental models or technical know-how for an individual. The explicit to tacit knowledge conversion is facilitated when the former can be re-experienced e.g. using documents, graphical representation, or stories.

2.1.2. Knowledge in the construction sector

Unfortunately, in the construction sector the above-mentioned principles cannot be directly applied. The temporary aggregation between clients, designers, construction companies, field operators and the other stakeholders makes difficult the identification of the effective hierarchical roles and consequently the definition of structured processes.

Furthermore, the conversion processes that represent the learning activities, need to be managed in a cross organisational context. This represents a critical point due to the unwillingness of the stakeholders in sharing information with other partners that are seen as competitor entities.

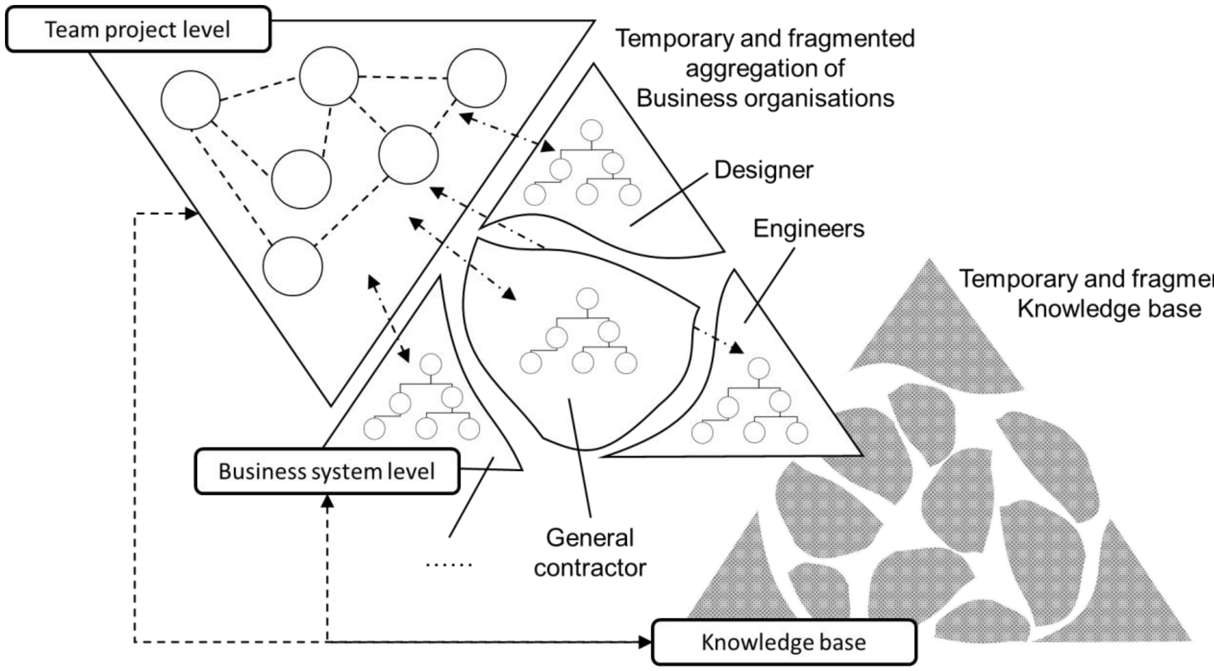


Figure 8. A simplified vision of the hypertext organisation schema for the construction sector

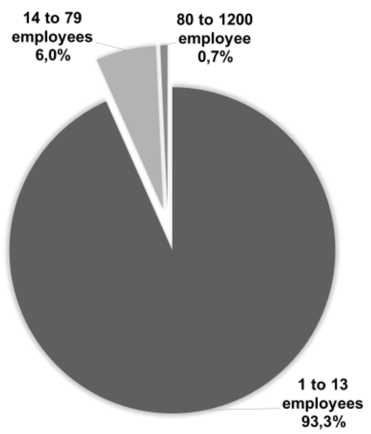
Figure 8 shows a possible simplified reinterpretation of the hypertext organisation proposed by Nonaka and Takeuchi. In this representation, the business system level is disrupted to admin the identification of the different stakeholders involved in the development of a project, each one with its own hierarchical and organisational structure[11]. The knowledge base level is in turn disrupted

[11] To comprehend the effective fragmentation of the construction sector and its possible impacts on the analysed model, it is possible to study the market composition, i.e. the organisations that composes the construction sector in terms of dimensions. In Italy more than 96% of the construction companies have less than 9 employee (ANCE, 2017). This configuration which is often associated with the Italian context, it is indeed a characteristic valid at world level. The 95% of OECD[11] enterprises are SMEs (OECD, 2000) and in the UK construction sector more than 93% of the enterprises in 2005 had less than 13 employees (DTI UK, 2006). The following images show the configurations of construction sector in UK and Italy. In both countries up to 99% of the companies in the construction sector are small companies, i.e. companies with less than 50 employees (or 79 in the case of UK).

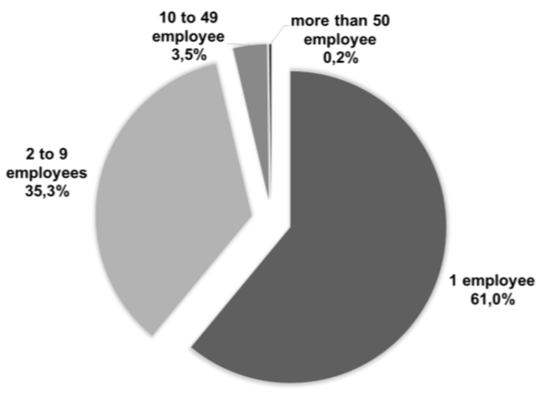
representing both the knowledge embedded and generated by each subject in the business system level and the knowledge generated by the subjects as a whole in the team project level, highlighting its volatility in this framework. Moreover, the coexistence of a temporary aggregation - i.e. the team project - that impose defined and specific rules and of companies that can have long histories and defined processes and hierarchies can generate points of friction and consequently difficulties in the creation of an ideal learning environment. However, as pointed out by Engeström (2001), the knowledge conversion model proposed by Nonaka and Takeuchi assumes that the assignment for knowledge creation is given from above without conflicts. Hence, the SECI model requires the creation of a friendly environment where knowledge and learning processes are defined in a top-down approach (Engeström, Miettinen and Punamäki, 1999).

This assumption is in contrast with the typical project environment in construction processes. Thus, the proposed model needs to be expanded in order to understand the societal and interpersonal relations in a distributed, conflictual and fragmented environment. As proposed in other studies (Hartmann and Bresnen, 2011; Miettinen *et al.*, 2012), cultural historical activity theory (CHAT) can be used to explicit the complex of relations and factors that rise in a construction project environment. CHAT was introduced between 1920s and 1930s by Vygotsky (1978) that formed the concept of *cultural mediation*. According to Engeström and Escalante (1996), from the Vygotsky's principle CHAT has evolved through three generations of research.

The first generation, centred on the cultural mediation idea, is commonly represented through a triangle of interaction where the connection between subject (stimulus in the original representation) and object (or response) is mediated by cultural artefacts (Figure 9, left). According to this schema, individuals and society cannot be understood without their mutual interaction and cultural means. The second generation, based on the work of Leont'ev (1978), extended the first model explicating the difference between individual action and collective activity. This concept was crystallised by Engeström (1987) that proposed the graphical representation reported in Figure 9 on the right. In this expanded framework, an individual subject is immersed in a complex of interrelations centred on his or her community creating a human activity structure.



Configuration of the UK construction sector in 2005 – (DTI UK, 2006)



Configuration of the Italian construction sector in 2015 - (ANCE, 2017)

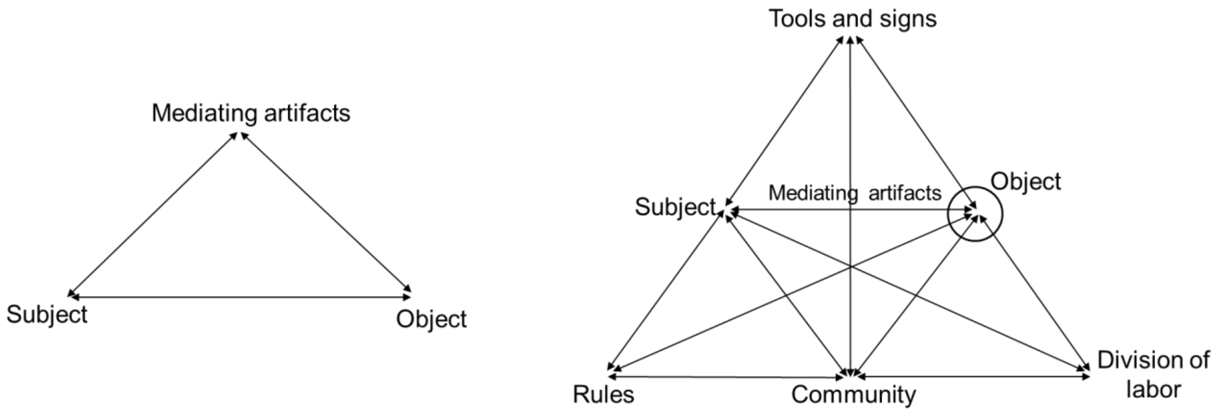


Figure 9. Reformulation of Vygotsky's model (left), and the Engeström's graphic representation of a human activity system (right)

The third generation of CHAT deals with the interaction between different activity systems that can include different traditions and/or perspectives. Figure 10 reports the graphical representation of two interacting activity systems proposed by Engeström (2001). This representation highlights the movement, the evolution and the different perspective of the object in the interaction between the two activity systems. In the Engeström words, *"the object of activity is a moving target, not reducible to conscious short-term goals"*.

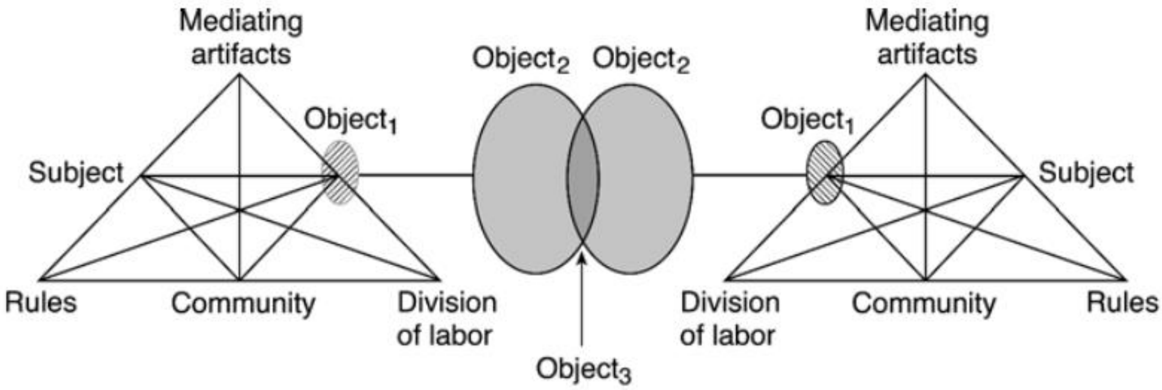


Figure 10. Representation of two interactive activity systems (Engeström, 2001)

Figure 10 represents a minimal model of interacting activity systems. Nevertheless, it can be expanded to represent the structure of an articulated project that can include multiple activity systems. Including this paradigm in the reviewed structure proposed in Figure 8, it is possible to define an integrated vision of the hypertext organisation schema for the construction sector (Figure 11). At the business system level, each triangle represents a specific entity (e.g. a company) that participate to the development of the project. The business system level highlights the independencies of each entity in their subjectivity and their interrelation by the means of the project terms and objectives. At the team project level, each part of the entities identified in the business system level can be represented as an activity system that collaborate on a shared object (e.g. a building). According to the principles of CHAT and following a recent interpretation in the construction sector proposed by Miettinen *et al.* (2012), the object is interpreted by each subject in a different way according to his or her specific interest and background. Moreover, the norms and rules that regulate each activity system at the team project level are generated by the integration of the directions defined at the business system level and the directions imposed by the external environment where the construction project is embedded.

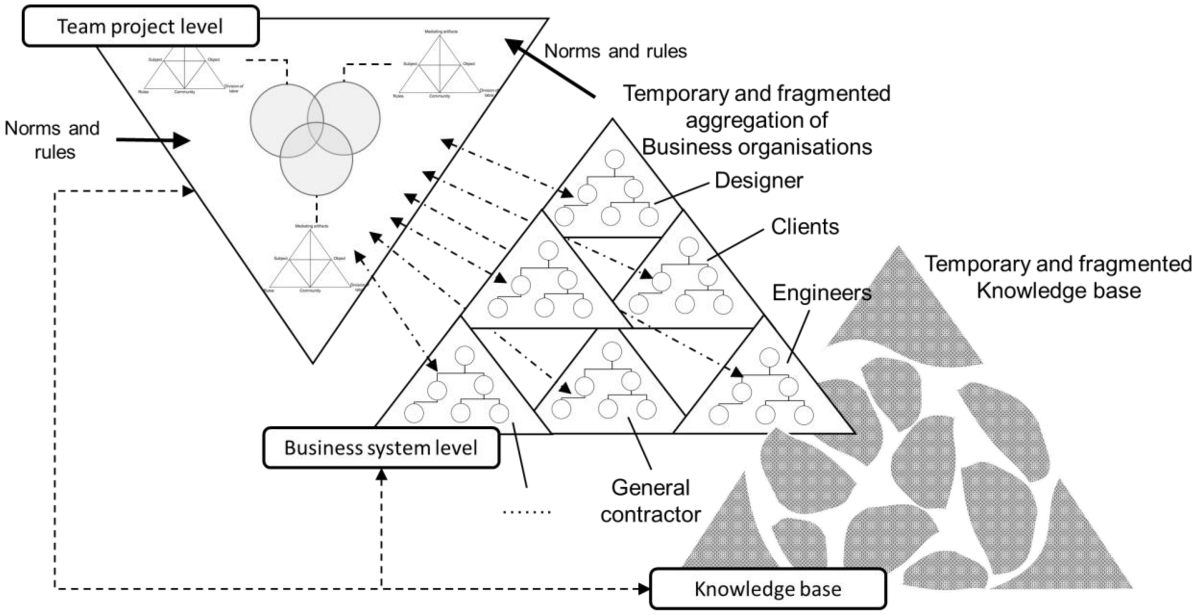


Figure 11. An integrated vision of the hypertext organisation schema for the construction sector

This interpretation highlights the complexity of interaction and the interpretation of the knowledge generated during the project activities. In fact, the coexistence of different communities and perspectives shape the way in which subjects act at the project level and the way in which they interpret and convert the generated knowledge.

Due to these factors, the proper application of KM technologies and techniques in the construction sector requires the understanding of relations between stakeholders (subjects), places and phases, taking also into account the societal aspects generated due to the temporary aggregation of different entities. While the framework proposed in Figure 11 can help in the understanding of the organisational structure in construction projects and of the interaction between subjects, it is not able to capture the spatial and temporal distribution that characterise the construction process. Thus, it can be integrated with a spatiotemporal perspective able to express these dimensions. Understand the spatial and time distribution of knowledge represents a critical point to understand its management and transfer processes. In fact, while knowledge moves from one place to another and while it moves forward in the time, its perception changes due to the change in the context and in the availability of the surrounding information.

Figure 12 reports a qualitative representation of the impact in the production and use of knowledge during time in the different locations involved in the construction process. The graph was empirically defined combining the evidences found in the literature and the observations of real construction processes gained during the collaboration with construction companies and engineering organisations. The proposed representation includes three locations, namely construction site (i.e. the place where the construction product is produced), product (i.e. the construction product such as a building or an infrastructure, once it is completed), and offices or factories (i.e. the places external to the physical location of the product where this last is designed and monitored and/or where its components are designed, produced and monitored). The horizontal axis reports the time correspondent to the three main phases of the construction process, namely design, construction, and operation and maintenance. The vertical axis reports the impact in terms of knowledge production and consumption, defined in a qualitative fashion.

According to the literature (Lin *et al.*, 2005) in the construction phase - and especially in the construction site - there is an intensive production of knowledge as a consequence of the concentration of efforts devoted to the production of the good. Towards the end of the production, the construction site is progressively transformed in the product that will produce knowledge during its entire life cycle. Hence, the impact of construction site decreases while starts the impact of the product as a physical place[12].

The time scale is not coherent to the real life-cycle of the building due to scale problem. Actually, operation and maintenance phase is longer than it is in the graph. Thus, even if the impact appears lower, it is extended on a long period of time covering a bigger area in comparison to design and construction phases. The oscillation of the graph in the operation and maintenance phase highlights the dynamicity of the impact along the life cycle that can change according to specific events including maintenance, restoration, and/or change in the destination of use.

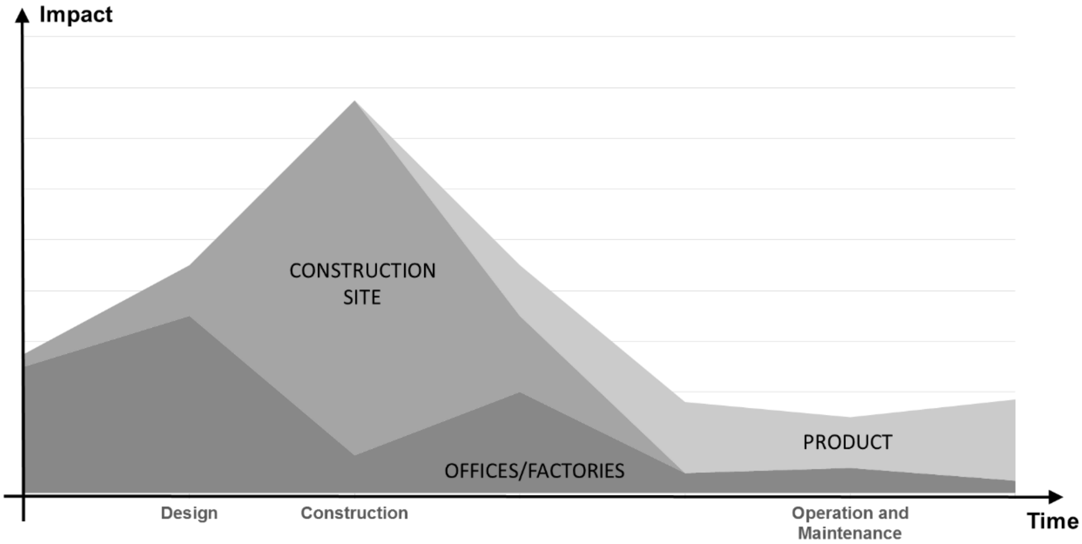


Figure 12. Spatiotemporal perspective in knowledge generation and use

In every phase and in every location can be recognised examples of both tacit and explicit knowledge. The design phase is characterised by an intensive use of regulations (e.g. national, local, hygiene), standards (e.g. fire safety, acoustic, management), and other requirements (e.g. client's requirements). Many of the solutions that can be used to fulfil these requirements can be converted in an explicit form of knowledge and can be stored and used through digital tools. For example, several studies explored the use of IT solutions to handle this explicit dimension. These include the use in energy simulations (Cheng and Das, 2014), the use of ontology-based approaches (Yurchyshyna and Zarli, 2009; Zhong *et al.*, 2012), the evaluation of permissions from public administrations (Pavan, Odorizzi, *et al.*, 2017), and the representation of regulatory knowledge through open standards (Dimyadi, Pauwels and Amor, 2016). Nevertheless, the identification of the correct solution between all the possible ones that can be proposed by an algorithm, the interpretation of design solutions in their context in a human perspective and the management of design teams are only some examples of the variety of tacit knowledge that can be experienced in the design phase.

[12] This representation follows the traditional process organisation that sees a huge impact of the on-site production. A change in the production paradigm, following a progressive industrialisation of the construction chain could change this impact distribution moving it from the construction site to the factories (off-site production). In this case, not only the impact of the knowledge production and consumption will change according to the spatiotemporal representation but also the organisational context according to a different delocalisation of the personnel involved in the process. In fact, while the on-site production sees a clear distance between the business level and the project (operational) level that are separated both in the physical context of knowledge interpretation and in the knowledge background, in the off-site production this "distance" is reduced. Nevertheless, the need to coordinate different and highly specialised works and tasks as well as several (fragmented) businesses in the on-site environment remains a crucial activity in the off-site production paradigm (Arashpour et al., 2016) maintaining acceptable the organisational structure proposed in Figure 11.

During the construction phase, the stakeholders matures genuine experiences about the constructability of a specific solution, the effective applicability of the solution in the context and its alignment with time and costs hypothesised in the design phase. Hence, during the construction phase are generated problem solving, know-how, know-what and innovation (Lin, Wang and Tserng, 2006). In this phase mostly of the knowledge is tacit and its collection and transmission is a critical issue in the process. In fact, in a general contractor environment, the site work is subcontracted to various trade contractors on a competitive tendering basis. Therefore suppliers have no incentives to share learning experiences (should read knowledge) for the sake of reapplying them on future project of the main contractor (Koskela and Vrijhoef, 2001). Furthermore, in the construction phase the interaction between designers, engineers, clients, authorities, contractor, and subcontractors produces a dynamic environment of knowledge generation and conversion. However, there is not a unique business level where can be established techniques and technologies to document this knowledge asset.

The operation and maintenance phase is distributed on a long time span. Furthermore, once the design and construction phases are concluded, it is difficult identifying who is the subject interested in the collection, management, use and update of the information and experiences produced during the life cycle of the construction product. In other sector, like the automotive one, the manufacturer can act as central collector in all phases including design, construction and monitor of the product. In the construction sector this is hardly the case. Nevertheless, there is a substantial learning opportunity from the analysis of a construction product (i.e. building, infrastructure) during its operation. For example, the durability of a technological solution can be assessed along its life cycle, and the effectiveness of a design solution in satisfy the client's requirements can only be evaluated following the experiences of the client.

2.1.3. Knowledge Management in the construction sector

Recognised the central value of knowledge, the interest in knowledge management is growing in the construction research community over the last decades (Deshpande, Azhar and Amireddy, 2014). Several initiatives in the knowledge management domain have been developed with reference to the construction sector. These includes the following projects:

- *CLEVER:* with the main aim of developing a framework for knowledge transfer in a multi-project environment in the construction sector (Kamara, Anumba and Carrillo, 2002).
- *IKON:* focused on projects and their organisational context with the aim of understanding how the organisations can learn from projects (Scarborough *et al.*, 2004).
- *B-Hive:* deals with inter-organisational learning demonstrating that problem structuring methods are by their nature appropriate for routine use within multi-organisational partnerships (Franco, Cushman and Rosenhead, 2004).
- *KnowBiz:* focused on the relation between knowledge management and business performances in construction (Carrillo, Anumba and Kamara, 2000).
- *KLICON:* focused on the role of IT in capturing and managing knowledge for organisational learning on construction projects (McCarthy *et al.*, 2000).
- *C-Sand:* focused on the impact of knowledge management in promoting sustainable practices in the construction sector (Wetherill *et al.*, 2007).
- *e-Cognos:* focused on the improvement of knowledge sharing across projects and organisations through a service based knowledge management for the construction sector (Rezgui, 2006).
- *IMPaKT:* focused on the perception of business impact of KM strategies for AEC organisations (Carrillo *et al.*, 2003).
- *Capri.net:* a web based KM system for "live" capture of knowledge and reuse in ongoing and future projects (Tan *et al.*, 2007).

Furthermore, Rezgui and Miles (2011) presented an activity based KM system to capture knowledge generated in the construction phase. Lin *et al.* (2005) proposed the use of knowledge maps to capture and reuse knowledge in construction projects and developed a framework for a web-based portal, which enables the user to search and read construction documents in different formats.

The majority of the literature seems to be concentrated on the aspects related to explicit knowledge (Whitley, 2000; Egbu and Sturges, 2001) that can be effectively managed through IT solutions. However, any KM approach that is purely based on IT solutions is bound to be less successful because people issues including organisational and social interaction as well as the willingness to share knowledge, are not directly solved by IT (Pathirage, 2007). A process-based view is required to critically consider the human factor in the equation of knowledge (Kogut and Zander, 1992). The relation between the management of tacit and explicit knowledge through the use of IT solutions has been critically analysed in several studies. Bresnen *et al.* (2003) argued on the difficulties in codifying construction knowledge using technological means. Egbu *et al.* (2003) analysed the IT's inability to transform tacit into explicit knowledge, underlining a possible hindering effect of IT in the activation of social mechanisms. Furthermore, Jimes and Lucardie (2003) presented the imbalance between the proliferation of KM technologies and the effective solutions of KM issues.

Nevertheless, the advent of novel technological solutions such as machine learning algorithms and Artificial Intelligence is moving the effectiveness of the IT allowing its integration in the process-based perspective. Furthermore, the ability to manage an increasing volume of data and information allows the exploration of new possibilities in the generation of knowledge (see chapter 2.2).

2.2. Managing knowledge in digital environments

2.2.1. Knowledge discovery in databases

The traditional method of turning data into knowledge relies on manual analysis and interpretation (Fayyad, Piatetsky-Shapiro and Smyth, 1996b). Due to the increasing availability of data (Soibelman and Kim, 2002) this type of manual application is becoming completely impractical in many domains (Fayyad, Piatetsky-Shapiro and Smyth, 1996b). The increasing use of digital technologies allows humans to gather more data than we can digest. Hence, analysis need to be automated, at least in part.

Knowledge Discovery in Databases (KDD) is defined as the nontrivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data (Fayyad, Piatetsky-Shapiro and Smyth, 1996a). Historically, the concept of finding useful patterns in data has been identified with various names, including knowledge extraction, information discovery, information harvesting, data archaeology, and data pattern processing (Fayyad, Piatetsky-Shapiro and Smyth, 1996a). Starting from 1980s KDD replaces them all referring to methods to find patterns and similarities in raw data (Soibelman and Kim, 2002).

KDD process comprises several steps, namely 1) selection of a target data set, 2) data cleaning and pre-processing, 3) data reduction and projection, 4) matching the goal of KDD to a particular data mining technique, 5) exploratory analysis and model and hypothesis selection, 6) data mining, 7) interpretation of mined patterns, 8) action on the discovered knowledge. These steps need a first activity of domain analysis where previous knowledge and goals of KDD process are identified (Fayyad, Piatetsky-Shapiro and Smyth, 1996b). The above mentioned steps can be classified in four main phases, namely business objective determination, data preparation, data mining, analysis and assimilation (Cabena *et al.*, 1998). Figure 13 reports the efforts required for each KDD step expressed in percentage. The bar chart highlights the central role of data preparation that requires more than a half of the entire efforts.

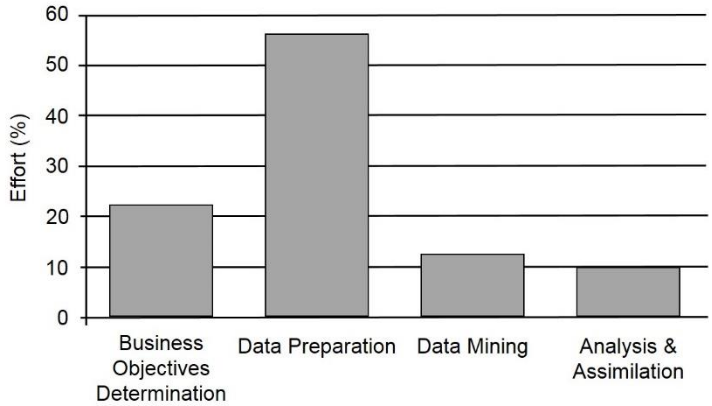


Figure 13. Required effort for each knowledge discovery in databases step – from (Cabena et al., 1998)

The first step of KDD is the objectives determination. This step is defined before any data analysis and it contains intrinsically domain knowledge. This last can be obtained from two different sources (Soibelman and Kim, 2002):

- Domain experts can provide knowledge such as historical data or description language and can evaluate the results of generated knowledge.
- Literature review.

The identification of prior knowledge represents one of the most important consideration in KDD processes (Fayyad, Piatetsky-Shapiro and Smyth, 1996b). KDD as well as KM are domain dependent. For this reason, their application cannot be generalised to an entire sector, especially if the sector is characterised by high fragmentation as it happens in construction. The proper introduction of KDD and KM techniques and technologies needs a priori study of the application domain taking in consideration e.g. context, human perception, needs, and existing issues.

Data preparation includes data cleaning and data pre-processing (or transformation). Data cleaning is *"the process of identifying and eliminating inconsistencies, discrepancies and errors in data in order to improve quality"* (Naumann and Rolker, 2000) and it deals with the evaluation and improvement of data quality (the concept of data quality will be further explained in chapter 3.3.1). Data cleaning includes the identification and quantification of data quality, the identification of kinds and source of errors, the standardisation and normalisation of data, error localisation and corrections, duplicates[13] detection and merging, final data quality assessment and feedback (Naumann and Sattler, 2006). Usually, database management systems (DBMS) provide means to limit errors in the generation of data (e.g. domain constraints, check, not null value, foreign key, and primary key) (Atzeni *et al.*, 1999). However, there are several reasons that can bring to data dirty, including lack of knowledge, typos, not DBMS sources, and heterogeneity. Furthermore, these issues are exacerbated in the case of building information models where the data structure has a limited number of constraints (see chapter 3.3).

Data mining is the application of specific algorithms to extract patterns from data (Fayyad, Piatetsky-Shapiro and Smyth, 1996b). There are several techniques and functionalities that can be applied in data mining including: characterisation and discrimination (i.e. the summarisation of the general characteristics or features of a target class of data); frequent patterns, association and correlations (i.e. mining patterns that occur frequently in data to discover interesting associations and correlation within data); classification and regression (i.e. finding a model or function that describes and distinguishes data classes or concepts); clustering (i.e. generate class label for a group of data); outlier analysis (i.e. find objects that do not comply with the general behaviour or model of

[13] Duplicate detection is the discovery of multiple representations of the same real-world object (Naumann, 2013)

the data). Data mining tasks can be classified in two main categories, namely descriptive and predictive. The first characterise properties of the data in a target data set, while the second performs induction on the current data in order to make predictions.

A data mining system has the potential to generate thousands of patterns or rules. Hence, it is fundamental to understand what patterns are interesting and in which way they can be used. *"A pattern is interesting if it is (1) easily understood by humans, (2) valid on new or test data with some degree of certainty, (3) potentially useful, and (4) novel. A pattern is also interesting if it validates a hypothesis that the user sought to confirm. An interesting pattern represents **knowledge**"* (Han, Kamber and Pei, 2012). There are several ways to objectively measure patterns interestingness, including support (i.e. the percentage of transactions from a transaction database that the given rule satisfies), confidence (i.e. the degree of certainty of the detected association), accuracy (i.e. the percentage of data that are correctly classified by a rule), and coverage (i.e. the percentage of data to which a rule applies). Furthermore, patterns can be evaluated in a subjective way according to human belief. In this case patterns are interesting if they are 1) unexpected, contradicting a user's belief; 2) actionable if they offer strategic insight that the user can use to act providing power to the generated knowledge (Kamaruzzaman *et al.*, 2016), and 3) expected in situations where the user need confirmation to his or her hypothesis.

2.2.2. Knowledge based decision making

The development and use of KDD applications to discover patterns and to extract knowledge is often a time-consuming procedure. Individuals need to access information and knowledge in limited time to effectively use it in everyday activities. The volume of information and knowledge sources available in digital format is today higher than human capability to use it all. Retrieving the correct source can be a time consuming, labor-intensive and inefficient work reducing the possibilities to learn from past experiences (Zou, Kiviniemi and Jones, 2017). Hence, the effective use of information and knowledge in decision making activities needs the introduction of technologies and techniques able to systematise the available sources and retrieve the essential ones.

Decision support systems (DSS) are defined as *"interactive, computer-based systems intended to provide support to the decision makers engaged in solving various semi- to ill-structured problems involving multiple attributes, objectives and goals"* (Nemati *et al.*, 2002). Starting from 1980s data warehouses have been studied as an integral part of DSS. A data warehouse empowers knowledge workers with information that allows them to make knowledge based decisions (i.e. decision based on fact) (Devlin, 1997). It provides an infrastructure that enables businesses to extract, cleanse, and store vast amounts of corporate data from operational systems to provide efficient and accurate responses to users queries (Inman, 1996). Data warehouses are the natural environments to apply KDD and other Artificial Intelligence (AI) techniques for knowledge creation, storage, dissemination and management processes (Nemati *et al.*, 2002). However, a data warehouse does not necessarily provide adequate support for knowledge intensive uses due to its computer-based nature that makes hard the management of non-codifiable (tacit) knowledge. Nevertheless, the ability to process and empower the accessibility of high volumes of data, information and knowledge, provides effective means to the application of knowledge conversion processes, including externalisation and internalisation (Nemati *et al.*, 2002). In DSS, the process devoted to solve new problems based on experience of similar past problems is generally known as Case-Based Reasoning (CBR) (Jonassen and Hernandez-Serrano, 2002). It is based on the exam of what happened in the past to reuse it in a new situation (Kolodner, 1993). CBR is a branch of AI born in the early 1980s (Schank, 1983, 2014; De Mantaras *et al.*, 2005) mainly based on four processes, namely retrieve, reuse, revise, and retain (Aamodt, 1994), (Figure 14).

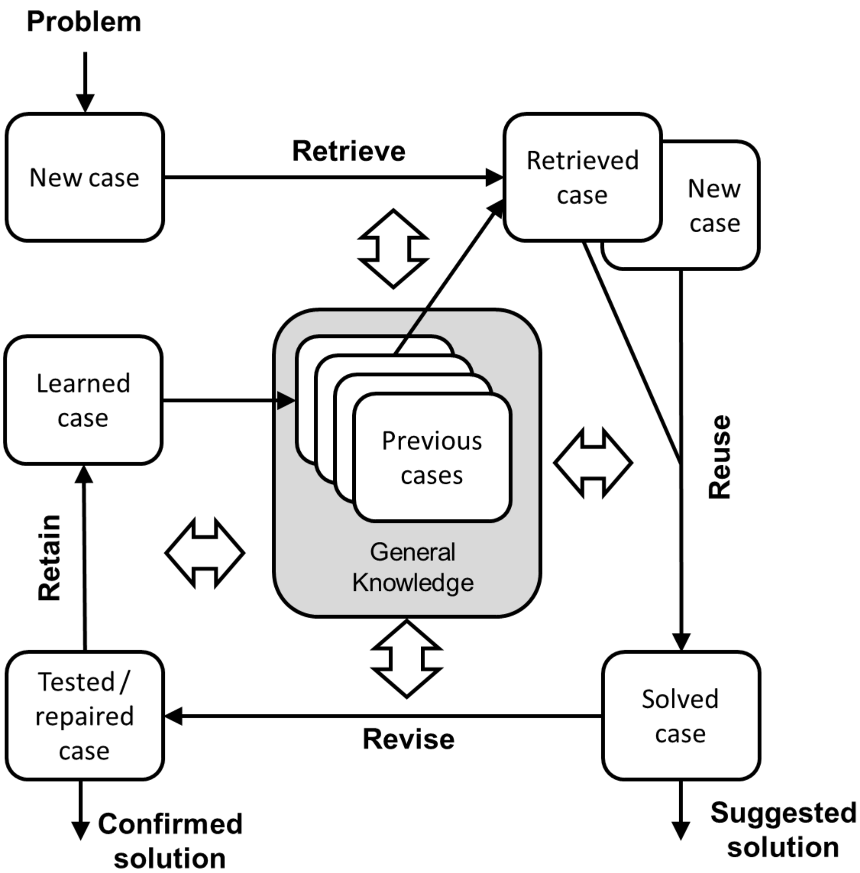


Figure 14. CBR model (Aamodt, 1994)

There are several examples of CBR applications in the construction sector, especially in risk identification and management. Construction hazard identification (Goh, 2009; Goh and Chua, 2009), safety risk analysis in subway operations (Lu, Li and Xiao, 2013) and construction supply chain risk management (Kumar and Viswanadham, 2007) are only some examples.

The retrieve process is the most important one in CBR. It is a process of searching and determining the most relevant sources (Aamodt, 1994; De Mantaras *et al.*, 2005) and it must guarantee the accessibility to the right sources in a feasible time. Most of the information produced in the construction sector are reported in natural language. This impose the use of technologies and techniques able to manage this type of information and provide the required services. This brings to the use of Natural Language Processing (NLP), an interdisciplinary topic that deals with the interaction between computer and human languages (Bar-Hillel, 1960; Chowdhury, 2003). Indeed, the issues related to the classification of documents in the construction sector have been studied long since. For example Caldas, Soibelman and Han (2002) and Caldas and Soibelman (2003) proposed an automated classification of construction project documents according to their related project components, Caldas, Soibelman and Gasser (2005) presented a methodology to integrate project documents in AEC/FM model-based information systems, and Lin and Soibelman (2007) worked with the classification of web based product information. Nevertheless, the advancements in the research field of AI are opening new possibilities introducing the ability to parse the content of documents exploring the embedded information and extracting knowledge usable by humans (e.g. Fortune Knowledge Group, 2018).

2.2.3. Technological implications

While the increasing volume of data and information imposed the introduction of new techniques for their interpretation like KDD, the explosion of Big Data (Laney, 2001) changed the requirements in the infrastructures used in the data and information management. Traditional relational database management systems (RDBMSs) provide powerful mechanisms to store and query structured data under strong consistency and transaction guarantees (Atzeni *et al.*, 1999). Furthermore, thanks to the long period of use in the market, they have reached an unmatched level of reliability, stability and support.

However, traditional RDBMSs may not be the right solutions to handle Big Data bringing to the development of novel data storage systems subsumed under the term NoSQL databases (i.e. not only SQL databases). Unstructured data, increasing use of sensors, and user-generated contents, are only some example of the Big Data phenomenon (Gessert *et al.*, 2017). There is a wide range of NoSQL databases systems and it is not trivial understanding where they differ from each other and where they excel. Nevertheless, NoSQL databases can be classified according to their data model into four mayor categories, namely key value stores, document stores, column family stores, and graph databases (Hecht and Jablonski, 2011; Bicevska and Oditis, 2016; Makris *et al.*, 2016; Gessert *et al.*, 2017).

In key value stores, the store values correspond to specific keys. Like maps or dictionaries, data are addressed through a specific key that represents the only way to retrieve the associated value that is stored as un-interpreted data array[14]. These databases are completely schema free and are particularly suitable for simple operations, speeding up the time of response. Examples of key value databases are Redis (Redis, 2018) and Dynamo (DeCandia *et al.*, 2007).

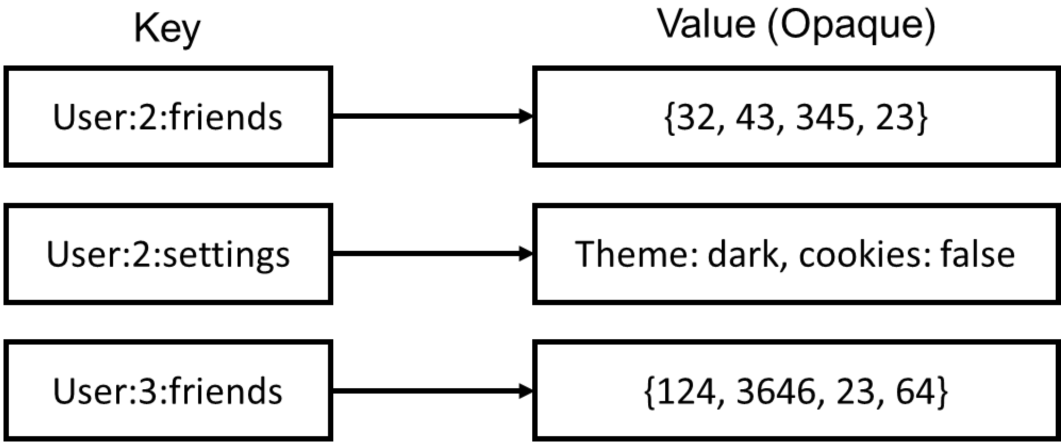


Figure 15. Example of a key value store

A document store encapsulates key value in pairs in semi-structured formats such as JSON (ECMA, 2017) documents (Gessert *et al.*, 2017). Document stores are completely schema free. Furthermore, stored values are not opaque[15] (in contrast to key value stores) and can be queried allowing querying operations on keys, IDs and contents. Thus, it is possible to access the entire document or some parts and retrieve the specific information required. These systems are convenient in the case of data integration and schema migrations tasks. Practical use cases include real time analytics, logging and storage layer of small and flexible website like blogs (Hecht and Jablonski, 2011). Examples of document stores are CouchDB (Apache, 2018a) and MongoDB (MongoDB, 2018).

[14] *This means that the stored data are not directly interpretable (i.e. are opaque) for the database and the only way to retrieve the value is using the related key. The direct consequence is the impossibility to define specific queries based on the stored values. In fact, key-value stores have no query languages and cannot perform complex operations on the stored data.*

[15] *In contrast to key value stores, in document stores the data are not opaque, meaning that the value can be searched according to specific queries based on the database content.*

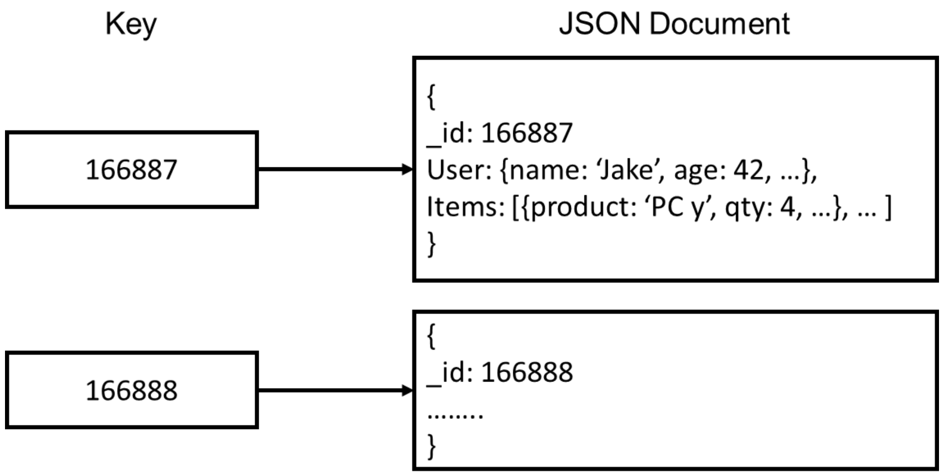


Figure 16. Example of a document store

All column family stores are inspired by Google BigTable, and the data model is described as *"a sparse, distributed, persistent multidimensional sorted map"* (Chang *et al.*, 2006). In this map, an arbitrary number of key values can be stored within rows and multiple versions of the values can be stored in chronological order. This data schema supports the organisation and partitioning activities grouping columns in columns families. Column family store databases are suitable for application in the case of huge amount of data stored on very large clusters (Makris *et al.*, 2016). Examples of column family stores are HBase (Apache, 2018c), Hypertable (Hypertable, 2014), and Cassandra (Apache, 2016).

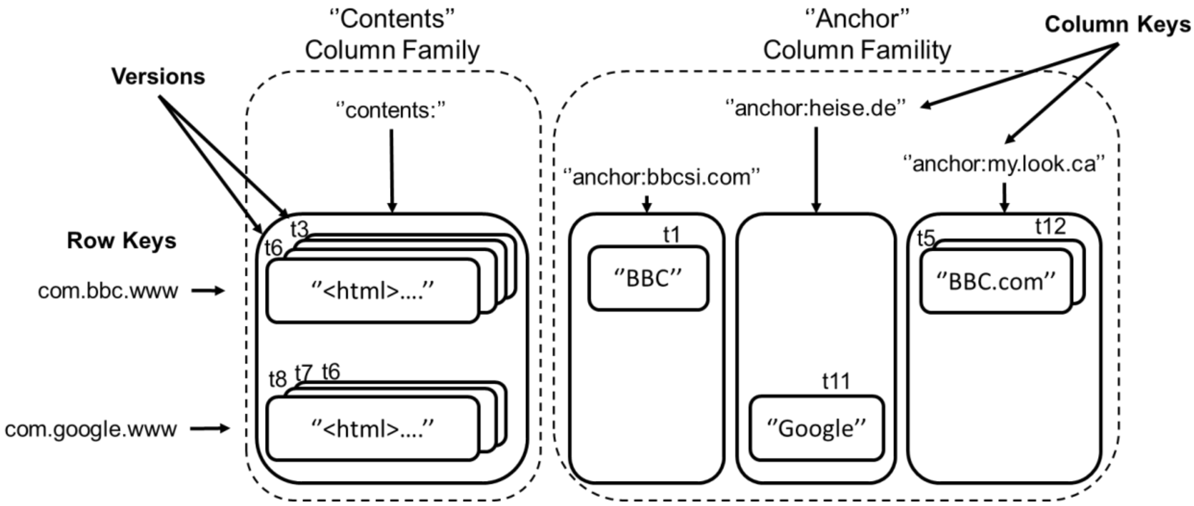


Figure 17. Example of a column store

Graph databases are characterised by a flexible graphical representation that provides powerful scalability performances. They are specialised on efficient management of heavily linked data[16] and optimised for highly connected data (Miller, 2013). Hence, graph databases are suited for applications with many relationships because cost intensive operations like recursive joins can be replaced by efficient transversal graph operations. An extensive comparison between relational and graph databases performance is proposed by Vicknair *et al.* (2010) and highlights the ability of graph databases in handle complex and linked queries on real data.

[16] The term linked data refers to the publication (in web environments) of structured data that can be linked one to each other from different sources. This term is related to semantic web principles introduced by Berners-Lee (2009) and nowadays at the centre of several research initiatives. In the construction sector research field the use of semantic web technologies and linked data is gaining increasing interest and represent a central area of research as presented in chapter 3.1.3.

Table 2. Query results on Character databases, in milliseconds – from (Vicknair et al., 2010). Rel: Relational database, Neo: Neo4j (graph) database, d: length of the search string, 1000, 5000, 10000, and 100000: number of nodes, 8k and 32k: dimensions of random string generated for the test

Database	Rel	Neo	Rel	Neo	Rel	Neo	Rel	Neo	Rel	Neo
	d=4	d=4	d=5	d=5	d=6	d=6	d=7	d=7	d=8	d=8
1000char8k	26.6	35.3	15.0	41.6	6.4	41.6	11.1	41.6	15.6	36.3
5000char8k	135.4	41.6	131.6	41.8	112.5	36.5	126.0	33.0	91.9	41.6
10000char8k	301.6	38.4	269.0	41.5	257.8	41.5	263.1	42.6	249.9	41.5
100000char8k	3132.4	41.5	3224.1	41.5	3099.1	42.6	3077.4	41.8	2834.4	36.4
1000char32k	59.5	41.5	41.6	42.6	30.9	41.5	31.9	41.4	31.9	35.4
5000char32k	253.4	42.3	242.9	41.5	229.4	35.3	188.5	38.5	152.0	41.5
10000char32k	458.4	36.3	468.8	41.6	468.3	41.6	382.1	41.5	298.8	36.3
100000char32k	3911.3	41.4	4859.1	33.3	6234.8	37.3	4703.3	41.5	2769.6	41.5

Examples of graph databases are Neo4j (Neo4j, 2018), and HyperGraphDB (Iordanov, 2010). A special form of graph databases is the RDF that will be presented in chapter 3.1.3. The application of this framework in the construction context represents nowadays a central area of research in the field with specific reference to the themes of collaboration and interoperability. For this reason, it will be extensively explored in the next chapter.

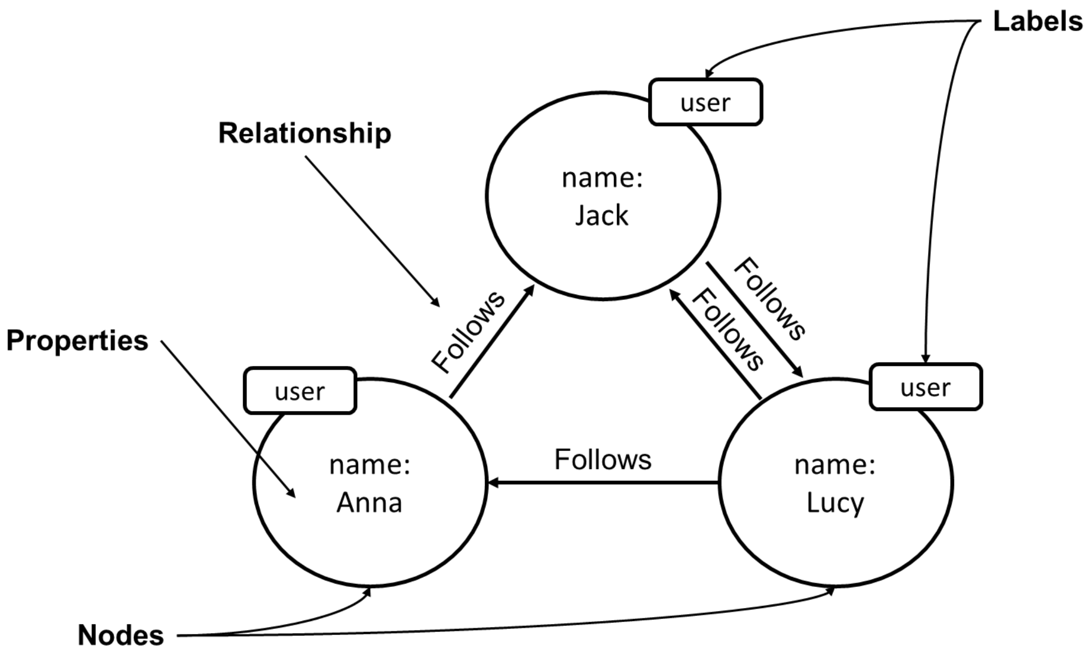


Figure 18. Example of a graph database in social networks

In addition to database solutions, the Big Data phenomenon imposed the change in use of IT infrastructures to handle the increasing requirements in computational power and storage. In 2004 Google presented the MapReduce programming model and its implementation (Dean and Ghemawat, 2004). From there, Hadoop (Apache, 2018b) was created as a framework to process, store and analyse massive amounts of distributed, unstructured data. It defined a change in the data analytics approach switching from one, huge block of data analysed with a single machine to a disruption of Big Data into multiple parts so each part can be processes and analysed at the same time drastically reducing execution times (Kelly, 2014). Hadoop includes two main technologies to address storage and processing issues. The Hadoop Distributed File System (HDFS) provides storage function, while the MapReduce programming model provides processing. Figure 19 shows an example of the data flow using map reduce techniques.

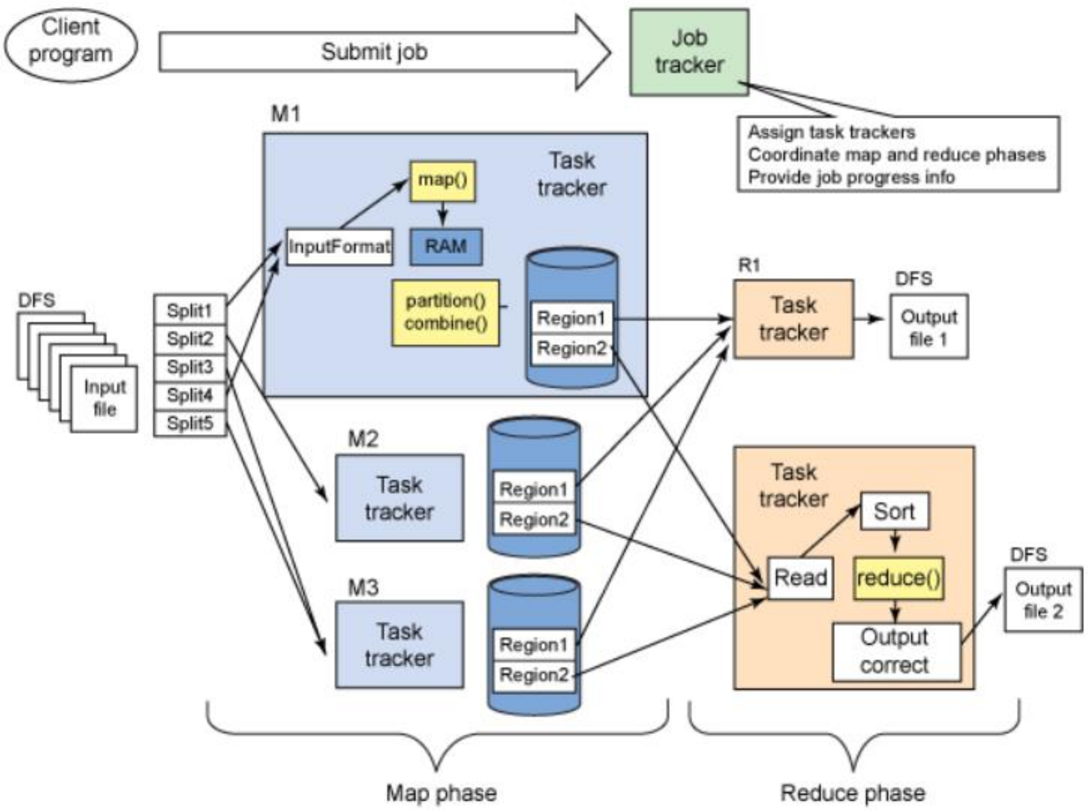


Figure 19. MapReduce data flow diagram – originally from (Markey, 2012)

2.3. Conclusions and focus of the research

This chapter provides an overview on the vast area of knowledge and knowledge management focusing on its possible interpretation and application in the construction sector. The proposed analysis highlights the main peculiarities of these lasts providing a guide to understand their impact on knowledge management applications. In the first section of the chapter, emerge the variety of type (explicit, tacit) and sources (fragmented stakeholders) of knowledge and the distributed localisation in terms of both space (offices, factories, construction site, products) and time (design, construction, maintenance and operation). The framework proposed in Figure 11, i.e. the reinterpretation of the hypertext organisation schema integrated with CHAT, provides the theoretical basis to develop the frameworks presented in chapter 4. The high impact of interactions between different subjects (or activity systems as proposed in Figure 11) underlines the need to explore available means to collaborate. Thus, chapter 3 presents an extensive review on this area.

The literature explored in this chapter highlights the central role of knowledge and culture in decision processes. Tversky and Kahneman (1974) extensively explored how people base their decision on a limited number of heuristic principles that can lead to severe and systematic errors. The limited ability of people in correctly evaluating and using the available information (in combination with previous knowledge) underlines the importance of means to support decision making, i.e. technologies and techniques that can help individuals in the evaluation and decision processes (explored in chapter 2.2.2). This principle represents one of the main guidelines that orientated the development of the proposed research.

The literature review highlights other two critical points. On the one hand, manage, classify and retrieve the correct source of data, information, or knowledge seems to be a critical point for the effective deployment of knowledge management processes. This implies the ability to process non-structured data such as documents written in natural language and/or provided by web sources to guarantee their accessibility in a reasonable time. On the other hand, the collection of information and knowledge still represents a critical activity. Due to the lack in the use of the available information and knowledge assets, individuals tend to underestimate their value hindering the collection process. Moreover, due to the fragmentation and spatiotemporal dislocation of sources, the collection of all possible data, information and knowledge represents a critical point. In fact, recover these assets in a second time may be too onerous or even impossible. In this context, this research focuses on the identification of technologies and techniques that can pave the way to collect these assets from heterogeneous sources following the time distribution of construction processes.

3. Collaboration, interoperability and data management in the construction sector

"Whatever the device you use for getting your information out,
it should be the same information"
(Tim Berners-Lee)

Chapter 2 focuses on knowledge management and defines the societal context where the research is developed. Moreover, it reviews the existing technologies that can be used in knowledge management process in digital environment. The analysis highlighted the need to shared data, information and knowledge to promote knowledge conversion and generation processes. Chapter 3 aims to explore the existing technologies and processes related to collaboration and interoperability in the construction sector. This study provides the basis to understand the state of the art about collaboration environments that constitute a central component of the framework described in chapter 4. Moreover, the chapter presents the existing research gaps and highlights some existing issues that need to be further explored (see chapters 3.2 and 3.3).

Collaboration, interoperability and data management are three synergic aspects in construction sector processes. Collaboration is defined as a *"creative process undertaken by two or more interested individuals, sharing their collective skills, expertise, understanding and knowledge (information) in an atmosphere of openness, honesty, trust, and mutual respect, to jointly deliver the best solution that meets their common goal"* (Wilkinson, 2005). This definition involves the creation of a friendly environment without conflicts and with shared and common interests between the subjects involved. According to the analysis proposed in chapter 2.1.2 this is hardly the case in the construction sector. Hence, the notion of collaboration as well as its definition need to be reinterpreted in the context of application. The representation proposed in Figure 11 can be used to understand the organisational framework (i.e. the context) where the collaboration processes can be studied and embedded. This includes different perspectives and means to activate collaborative processes in traditionally conflictual environments that can be approached e.g. from the technological side (as extensively explored in this thesis) or from the contractual side as proposed in other relevant studies as well as acting from both sides.

The Institute of Electrical and Electronics Engineers (IEEE, 1990) defined interoperability as *"the ability of two or more systems or components to exchange information and to use the information that has been exchanged"*. Konstantas *et al.* (2005) included a vision on users defining interoperability as *"the ability of a system or a product to work with other systems or products without special effort from the customer or user"*. Data management can be defined as an *"administrative process by which the required data is acquired, validated, stored, protected, and processed, and by which its accessibility, reliability, and timeliness is ensured to satisfy the needs of the data users"* (WebFinance Inc, 2018a). As such, collaboration and interoperability need the integration of proper data management procedures, but they cover a wider area dealing with human interaction and sharing processes including not only data but also information and knowledge. Nevertheless, data management definition underlines the importance of accessibility, reliability and timeliness of the data to guarantee their usability. Hence, data management is a fundamental prerequisite in the study of collaboration and interoperability processes.

In the construction sector, BIM represents one of the latest approaches to bridge the existing interoperability gaps promoting the introduction of collaborative processes. In BIM, building information models can be viewed as collectors of data, information and knowledge. Hence, BIM can be identified as a critical area in the study of collaboration, interoperability and data management in the construction sector.

3.1. Collaboration and interoperability

Among its main processes, collaboration includes sharing information activities. Indeed, sharing information includes both collaboration and interoperability. Information is shared along a defined process involving different stakeholders (collaboration dimension) that must be able to read the information using different tools (interoperability dimension). Can be listed four main types and/or technologies to share information along the construction process, namely file exchange, central repositories, container exchange, and industrial data space (Törmä, 2017). In addition, cloud computing solutions are introducing new possibilities, including the use of software and/or specific applications directly from the cloud.

File exchange is the basic way to share information. File exchange is based on the transmission of entire files through digital means like e-mails and/or physical supports. Its simplicity and the few infrastructure requirements have led to its spreading. It is still the main form of information exchange in the construction sector. Figure 20 shows an example of file exchange using e-mails.

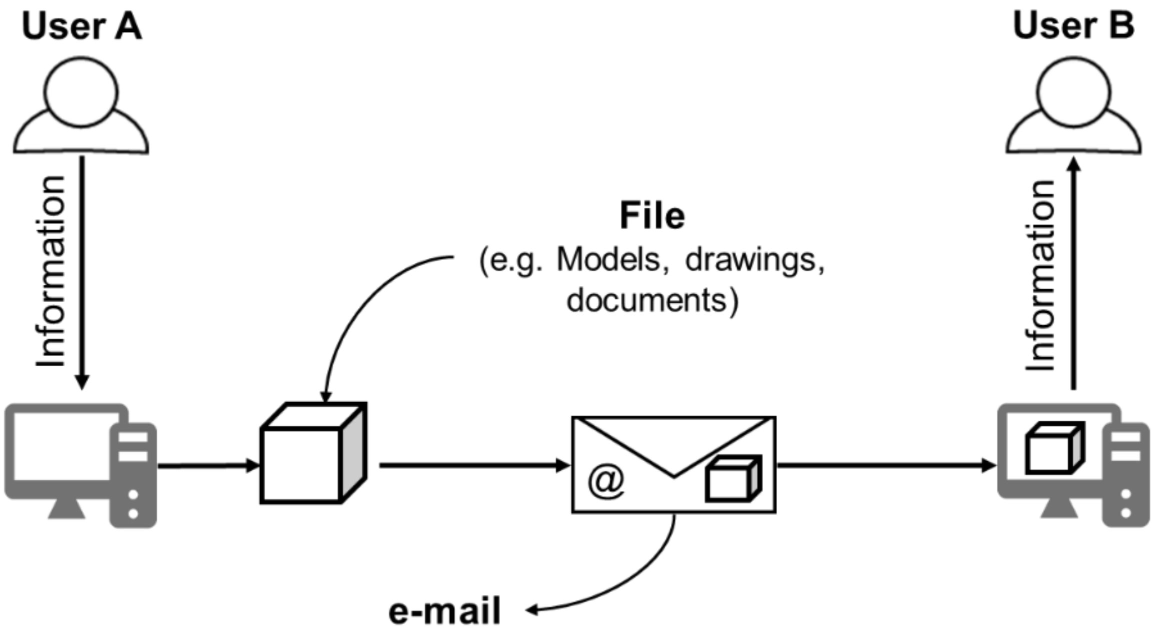


Figure 20. File exchange schema (e.g. using e-mails)

In file exchange, the exchanged units are entire file. This means that the receiver (User B in Figure 20) must read the entire file and extract from it the only information that he or she needs. Furthermore, the sender (User A in Figure 20) may need to filter information that do not want to share. Individuals - according to their roles in the process - develop files using their own computer. This produce a fragmented working environment where it is difficult to understand the effective state of development of the project and of each file. Hence, in file exchange the receiver has no control on the effective update of the file that he or she received and use.

Central repositories are based on the idea of centralising the information in the process. Information can be centralised on a file-based (e.g. defining a shared structured directory where the file is uploaded) or through advanced collaboration platforms where data can be managed at object level and can be defined collaboration and verification processes. Centralising the information, can limit the fragmentation and timeliness issues of point-to-point file exchange. Figure 21 shows a file-based central repository schema. Every individual produces information and uploads the information in the central repository allowing other to read it. In recent central repositories applications, individuals need to upload (or download) the entire file only once in (or from) the central repository, while exchanges and updates are managed in a star-like topology. However, the use of a central repository usually does not guarantee a granular access to the information.

In BIM, the use of central repositories is increasing thanks to the development of several commercial solutions. For example, Autodesk developed a set of cloud-based tools that can be integrated to manage different types of file including models, drawings and other documents and to define structured collaboration in the development of building information models (Autodesk, 2018b, 2018a). Similar solutions have been developed by Nemetschek Allplan (2018) with the BIM plus platform, Nemetschek Graphisoft (2018) with the BIM Cloud platform, Trimble (2018) with Trimble connect, Bentley (2018) with Project Wise, and Acca Software (2018) with usBIM platform. However, even if these solutions usually admit the use of open formats (see chapter 3.1.2) most of them are focused on specific proprietary formats. This can hinder the definition of effective collaboration processes due to a limited interoperability dimension.

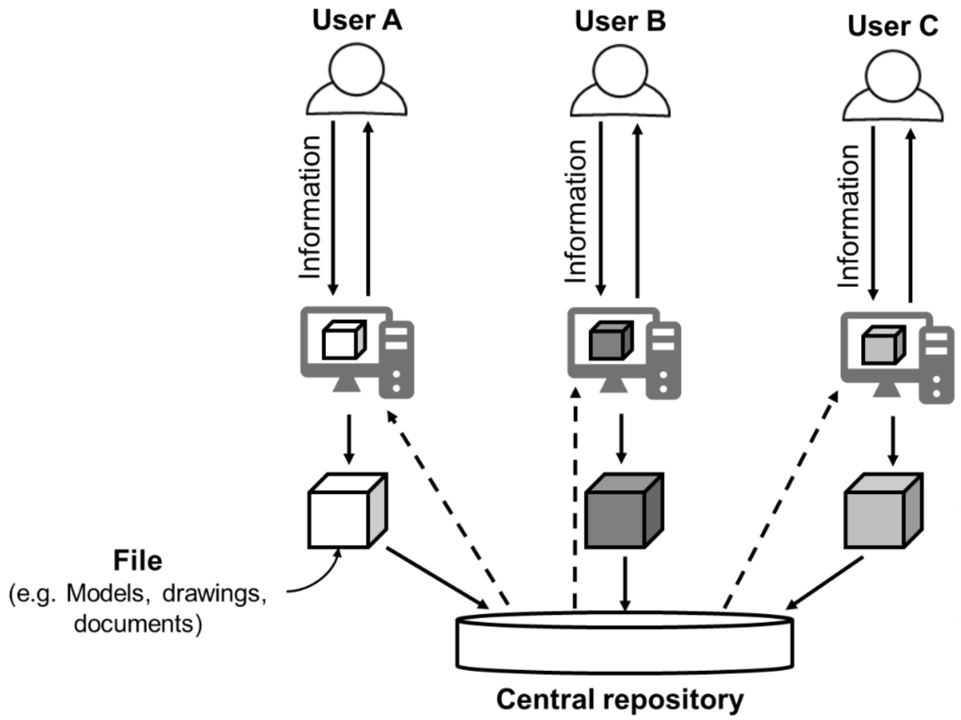


Figure 21. File-based central repository schema

Central repositories are defined according to the development of one project including a specific group of stakeholders and specific rules and requirements in a specific time-frame. In fact, in defined phases of the construction life cycle (e.g. design and construction) it is possible defining the “centre” of the phase (e.g. the architect or the construction company), at least in terms of responsibility for the collaboration processes and technologies. However, the centre may change along the construction life cycle requiring the definition of means to transfer the required information from one centre to the other. Furthermore, several data sources cannot be embedded in central repositories including e.g. geographic data, and product data.

Due to the limits showed by file exchange procedure and by central repositories, another direction of research has been the definition of container repository. The use of container repositories shifts the paradigm from a centralisation of information to a distribution of the information between multiple resources linked to each other. A container repository can be interpreted as a digital environment composed by datasets such as building information models, schedules, and documents linked to each other through defined linksets. A well-known application of this principle is the multi models paradigm (Fuchs, Katranuschkov and Scherer, 2010; Katranuschkov *et al.*, 2010).

Finally, in manufacture, with the advent of Industry 4.0 paradigm, can be found several efforts for the development of industrial data spaces (EU, 2016d; Working Group 2, 2017). However, as discussed in chapter 2.1.2 there are several differences between manufacture and construction industry. In industrial data spaces developed in the manufacture it is usually possible defining a central party who has the interest in managing the entire product life cycle from design and production up to its use, maintenance and disposal. For example, in the automotive industry the implementation of industrial data spaces is already in action. During an audition about Industry 4.0 the CEO of Porsche Consulting s.r.l. stated that *"[…] Magneti Marelli receives periodically the orders through production programs. Only 5 days before the assembly in Stuttgart, it receives the "start" for the production of the exact sequence of instruments that will be mounted on the car. Drawings and schedules are on-line in a shared platform accessible by every interested supplier […]"* and *"[…] customer profiling allows understanding which accessories could be of interest and it provides suggestions to the customer a few days before the production […]"*. However, as it is difficult to identify the centre of the construction life cycle, it is also difficult to identify a central party that has the interest in governing the entire process monitoring a product with a life cycle that can last more than 100 years[17].

3.1.1. Common Data Environment

In the construction sector, sharing information processes are nowadays materialised in the concept of Common Data Environment (CDE). One of the first definition of CDE is contained in the UK standard BSI-BS-1192:2007 (BSI, 2007) where point 4.1.2 says *"a 'Common Data Environment' (CDE) approach should be adopted to allow information to be shared between all members of the project team. This is a repository, for example a project extranet or electronic document management system."* The proposed definition reflects the cultural and technological advancement of the sector at the time of publication of the standard, focusing on the idea of document management (See Figure 6).

Nevertheless, the main schema is similar to the actual one presented in BSI-PAS-1192-2:2013 (BSI, 2013) including four main areas, namely "work in progress", "shared", "published documentation", and "archive" (Figure 22). BSI-PAS-1192-2:2013 extends the concept of CDE as a *"single source of information for any given project, used to collect, manage and disseminate all relevant approved project documents for multi-disciplinary teams in a managed process"*. In this last definition, the main structure is maintained but extended including specific rules defined for the management of digital environments (e.g. Employer Information Requirements[18], and BIM execution Plan[19]), the cooperation between several professionals, and the connection from the design and construction phase (Capex) to the operation and maintenance phase (Opex) - this last described in BSI-PAS 1192-3:2014 (BSI, 2014).

[17] *This aspect introduces also a possible technical issue that is the distance between the life-cycle of the tools used to generate and read the digital models (e.g. BIM authoring tools) and the life-cycle of the good. The lifecycle of a BIM tool is typically 12-24 months, whereas the lifecycle of FM legacy systems vary between 10-20 years and building lifespan could be up to 100 years.*

[18] *pre-tender document setting out the information to be delivered, and the standards and processes to be adopted by the supplier as part of the project delivery process BSI-PAS-1192-2:2013 (BSI, 2013)*

[19] *plan prepared by the suppliers to explain how the information modelling aspects of a project will be carried out BSI-PAS-1192-2:2013 (BSI, 2013)*

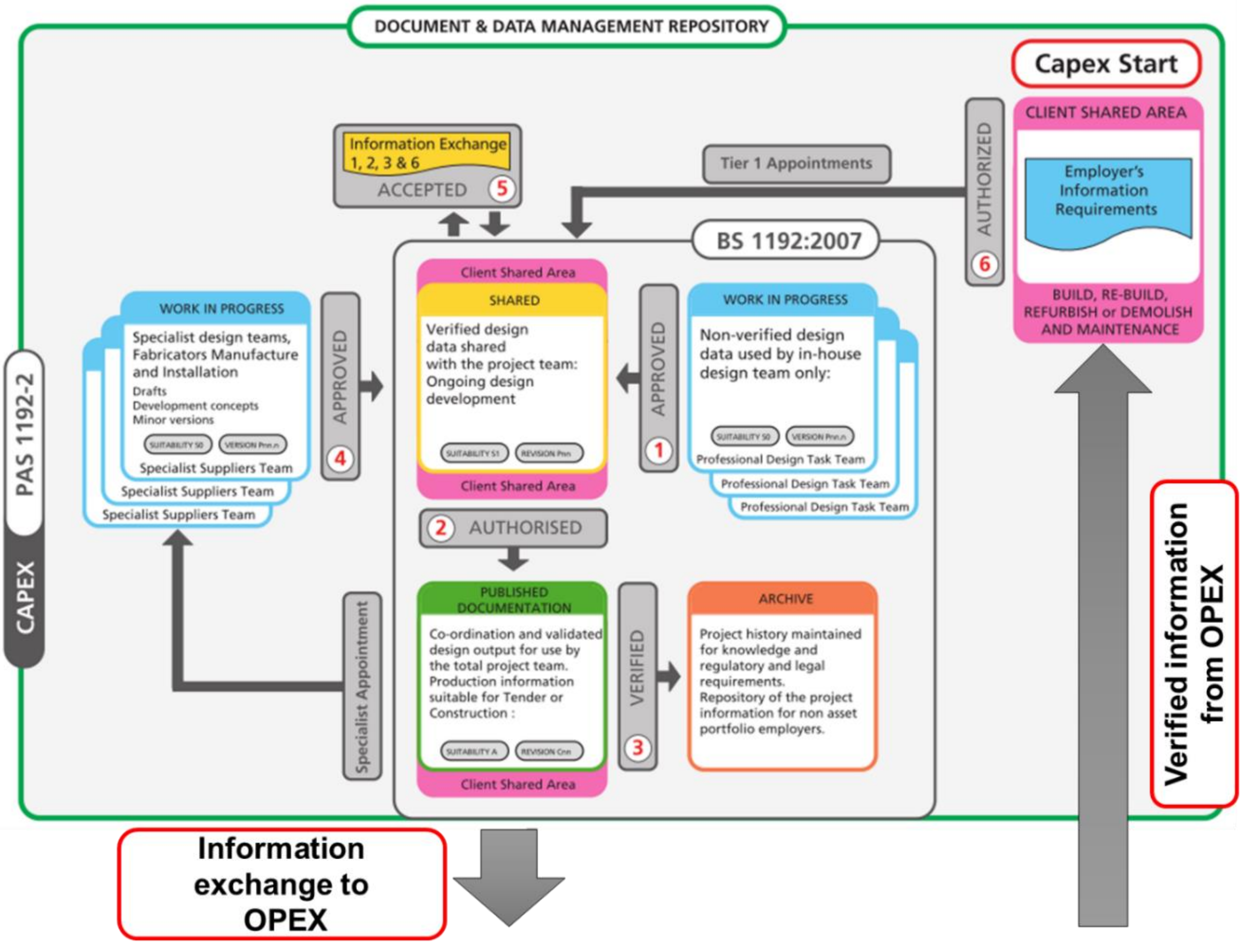


Figure 22. Common data environment schema (BSI, 2013) – Modified by the author

The Italian standard UNI 11337:2017 (UNI, 2017b, 2017c, 2017d) inherited the concept of CDE (called ACDat in the Italian version) widening its definition according to the latest technical and technological advancements. In UNI 11337-1:2017 CDE is defined as *"an environment for an organized collection and sharing of data related to digital models and output for a single work or a single complex of works"*. Nevertheless, UNI 11337-1:2017 introduces also the concept of digital collaboration platform as *"a digital environment for an organized collection and sharing of data, information, models, objects and output for the building industry: resulting products, component products and processes (objects, parties, actions)"*. Figure 23 reports a visual schema of the information flow association to the use of a CDE according to UNI 11337:2017. In the lower-left part of the graph it is identified the CDE where all the stakeholders involved in the process collaborate, including clients. In the upper-left part of the schema, it is identified the digital platform, connected to the client's repository in a logic of use and re-use of data and information.

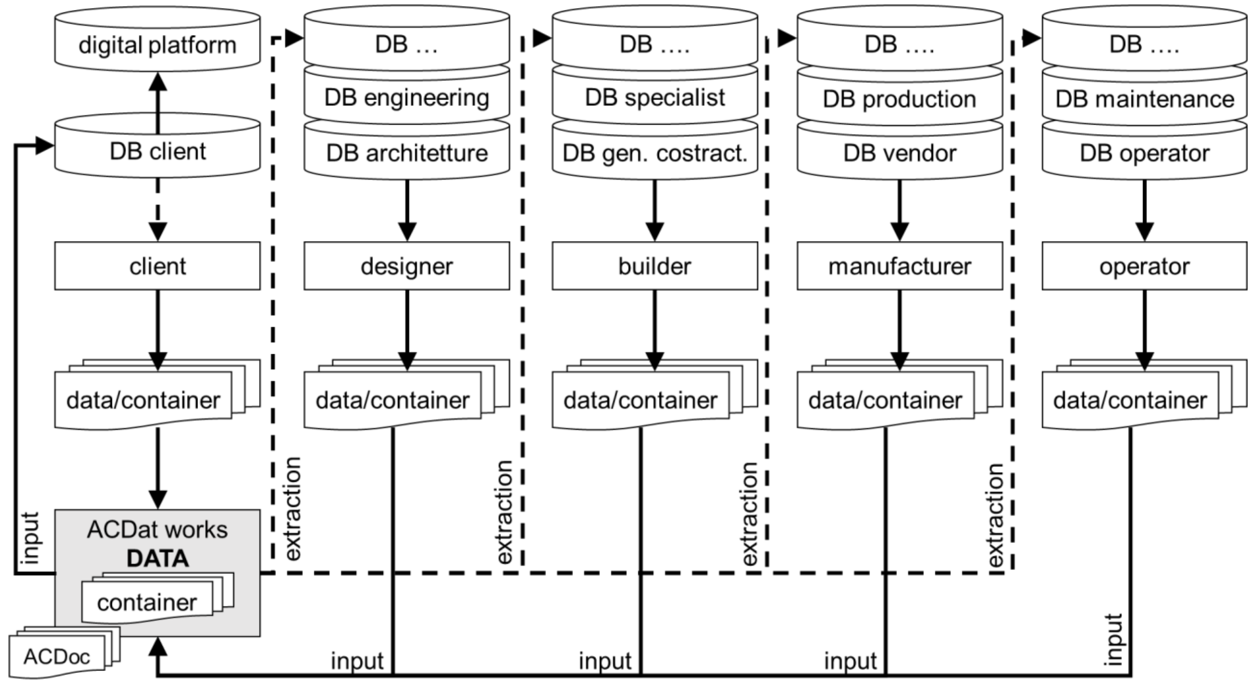


Figure 23. Common data environment schema according to UNI 11337-5, (Pavan, 2017; UNI, 2017d)

In the practice, nowadays CDE constitutes one of the main components in the implementation of BIM processes. However, in this context it cannot be seen as a simple repository as proposed in BSI-BS-1192:2007. The gates identified in BSI-PAS-1192-2:2013, i.e. the layers where it is defined the passage of a file and/or an information from one space to the next one (e.g. from shared to published), must be defined through a managed process. This requires the introduction of structured and integrated means to monitor, verify, and validate documents and contained information to guarantee the quality, traceability, and transparency of the process. However, due to the temporary and fragmented work environment (see chapter 2.1 for details) and the one-of-a-kind product that characterise the construction sector, rules and processes must be defined for each project. Furthermore, in an iterative design process the identification of what information can (or must) be shared and with who is not trivial. This issue has been pointed out by Pavan (2017) that proposed the logical duplication of the CDE allowing a more dynamic collaboration process. Figure 24 shows the proposed schema focused on the design phase. Both CDE can be divided in the four main areas defined in BSI-BS-1192:2007, constructing a link between the published area in the "Internal CDE" and the shared area of the "work CDE". The client has the need to maintain all the project information (including the information generated in the structured processes embedded in the CDE) between the different project phases and during the life cycle of the product. Hence, the "work CDE" is intended as owned by the client allowing the reproduction of the same process in the different phases of the project including construction and operation and maintenance.

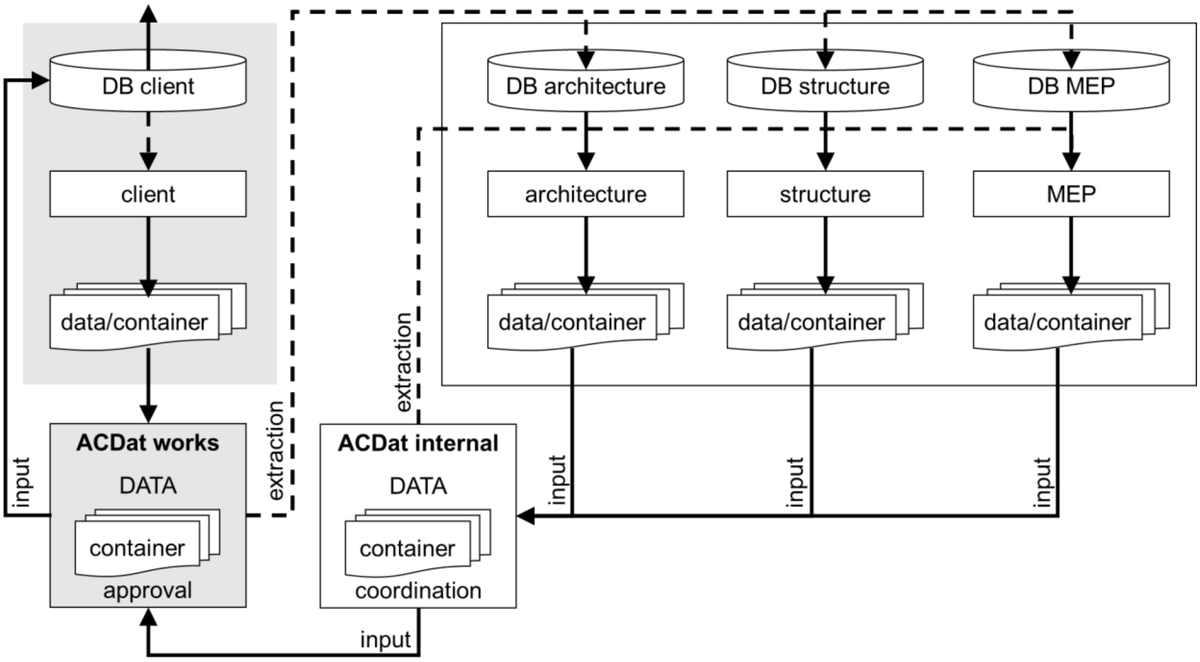


Figure 24. CDE application in design contracts proposed in (Pavan, 2017)

According to these evidences, the international standard ISO 19650-1 (ISO, 2018), extends the principles related to the concept of CDE explaining how the whole information model[20] might not be included in one place but can be defined according to a distributed environment where the information are related according to defined principles established at the beginning of the process. Moreover, ISO 19650 introduces the concept of organisational information requirements (OIR) highlighting the need of integrate information provided by different projects to satisfy organisational needs such as strategic business operation, portfolio planning, policy-making, etc. integrating the information requirements in the overall organisational management structure and promoting the activation of processes to capture and reuse lessons learned (ISO 19650-2 - ISO, 2018).

However, one of the main declared obstacle in the introduction of new IT instruments in existing organisations is represented by the misalignment between the competencies of the personnel involved and the competencies required by the IT system (Enkovaara, Heikkonen and Tapionen, 1998). Focusing on the CDE structure, it is clear the difficulty that a fragmented sector with a high number of micro and small enterprises, as the construction one, can encounter in its right definition. The result is that the CDEs actually created and used in real cases reflect the competencies of the market with incomplete and many times inconsistent structures.

3.1.2. Industry Foundation Classes

All the described Information exchange types require the introduction of effective means to guarantee the interoperability between the tools involved in the collaboration process. In this direction, open and neutral data schemas have been developed to enhance interoperability (Venugopal *et al.*, 2012). Among all, Industry Foundation Classes (IFC) (ISO, 2013) can be seen as the reference standard in BIM processes.

Since 1995, BuildingSmart consortium (formerly known as International Alliance for Interoperability until 2008) is working on IFC, a common data model to present and describe building processes (Liebich and Wix, 1999; Laakso and Kiviniemi, 2012). IFC has been defined as a standard common language. In fact, IFC schema provides the building blocks for interoperability

[20] *Defined in the ISO 19650 as "set of structured and unstructured information containers" where an information container is defined as "named persistent set of information retrievable from within a file, system or application storage hierarchy".*

through its open and neutral data schema (Venugopal *et al.*, 2012). IFC represents geometry, relations, processes and material, performance, fabrication, and other properties, using the EXPRESS language.

However, in several cases, the adoption of IFC is considered a hard task due to several IT related issues that can be encountered, in both import and export of IFC files. Lack of information in sharing building information models through IFC formats is registered in several cases, e.g. when models are exported and imported in the same BIM authoring tool, from architectural models to structural models (Hu *et al.*, 2016; Sibenik, 2016), from BIM authoring tools to analysis software (e.g. for energy analysis) or from architectural models to facility management tools.

Different reasons have been pinpointed for this issue. Although IFC is a rich product-modelling schema, it is highly redundant[21] (Shafiq, Matthews and Lockley, 2013). Moreover, objects, relations, and attributes can be defined in several ways. Thus, data exchanges are often unreliable due to inconsistencies in the different assumptions and interpretations in expressing information (Sacks *et al.*, 2010). Furthermore, IFC has been developed with a focus on file exchange principles (Törmä, 2017), thus instance-level interoperability is not supported (Törmä, 2013).

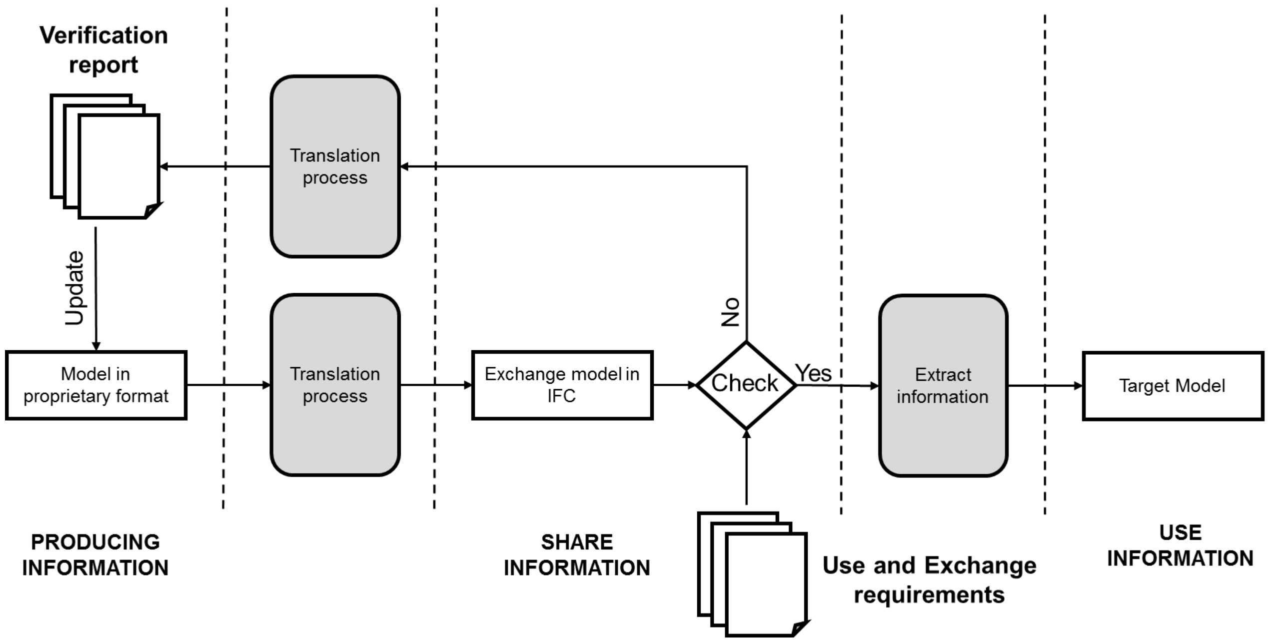


Figure 25. IFC-based information sharing use process

Generally, the exchange of IFC format does not follow a conversion roundtrip. Therefore, even if models can be exported in IFC (e.g. through BIM authoring tools), the IFC version of the model generally can only be used in a read-only mode. Information is lost in a roundtrip and the result is a restructured native model that generally needs to be further developed through extensive manual repairs (Törmä, 2013). This imply the introduction of laborious checking activities as shown in Figure 25. The model in its native format is translated in IFC. The IFC model is checked according to the use exchange and use requirements defined for the target model[22]. If the IFC model do not satisfy the requirements, it is defined a verification report used to update the original model in its native format, restarting the process.

Each discipline involves specific domain requirements and a defined set of information that need to be shared. Furthermore, commercial BIM authoring tools define information and data structure in different ways that remains embedded in the IFC producing syntactic incoherence (Lee,

[21] *In IFC files, it may happen that during the translation process a number of equal entities are generated, causing redundancy – i.e. repetition of the same entity – in the IFC.*

[22] *The target model represents the model in which is required the use of the information contained in the shared IFC model (e.g. the energy model developed starting from the IFC model)*

Eastman and Lee, 2015) and low data accuracy (Sibenik, 2016). Nevertheless, acting on the IFC data structure it is possible to overcome these issues operating directly on IFC models as shown in Figure 26 (Datacubist, 2018).

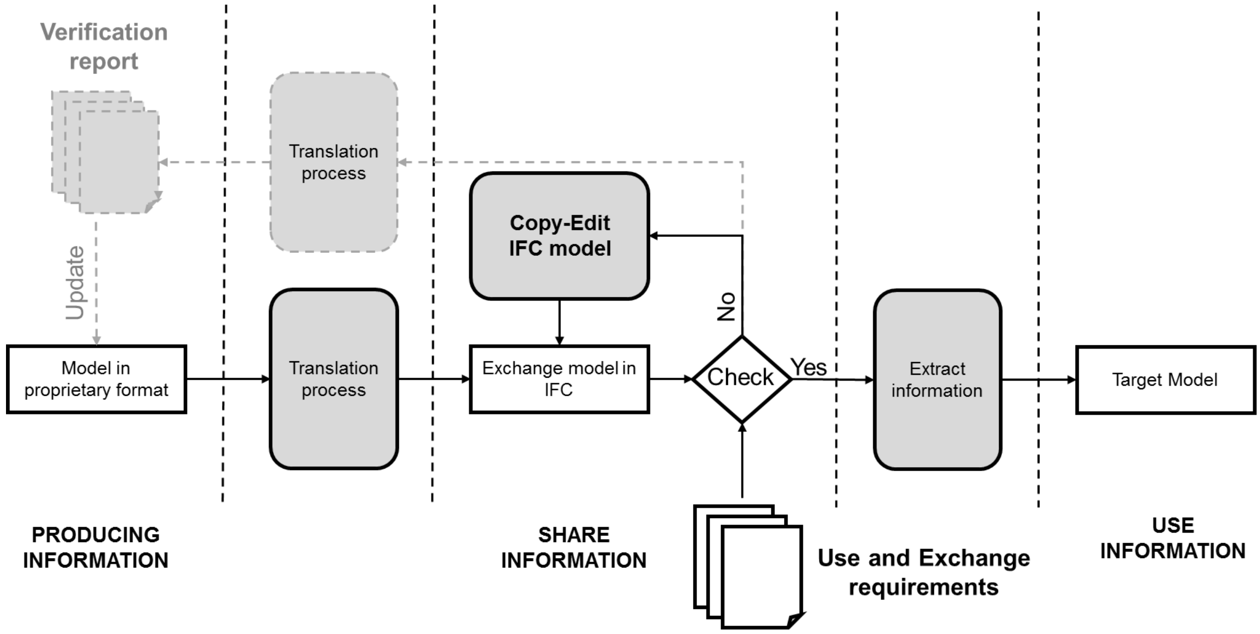


Figure 26. Process schema of IFC use with direct operations on IFC models – (adapted from (Datacubist, 2018))

However, this approach can produce the reversal of responsibility from the developer of the model in its native format (usually the designer) to the individual that use and modify the IFC model (usually the client or other specialised technicians involved in the process). Hence, the identification of proper rules about what can be modified directly in the IFC model and what need to return back to the native model remains a critical aspect in the information flow.

IFC covers only the transport of data and information in the process. Processes, standard terminology and technical requirements are not included in the IFC schema. Indeed, Building Smart includes five basic standards, namely Information Delivery Manual (IDM) (ISO, 2016), international Framework for Dictionaries[23] (IFD) (ISO, 2007), BIM Collaboration Format (BCF), Model View Definition (MVD), and Industry Foundation Classes (IFC) (ISO, 2013).

IDM focuses on processes that require the exchange or share of information between project participants and the information required and resulting from the execution of these processes (Building smart, 2010). It is strictly related to MVD that is the identification of the required model views, i.e. the *"subset of the IFC schema that is needed to satisfy one or many exchange requirements"* (Building Smart, 2018c), needed for a specific use of the model. IFD provides a systematic collection of terms, vocabulary and attributes to establish a standard semantic in the construction sector.

[23] *Including the buildingSMART Data Dictionary (bsDD)*

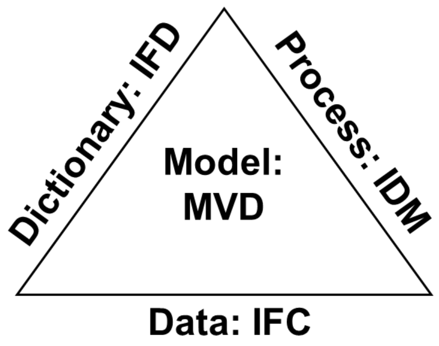


Figure 27. Building Smart basic standards (Building Smart, 2018a)

3.1.3. Semantic Web and Linked Data

Due to the limitations of the above-described information sharing techniques and technologies, and in accordance to the increasing need of combining information provided by different sources (e.g. BIM and GIS), researchers started proposing the use of semantic technologies in the construction industry (Pan, Anumba and Ren, 2004; Elghamrawy and Boukamp, 2008). The advantages of introducing semantic web technologies have been extensively explored in the literature. For example Rezgui *et al.* (2011) and El-Diraby (2013) highlighted the reasons to switch from a model centric approach to a distributed semantic approach. According to Pauwels, Zhang and Lee (2017) these advantages can be included in three main topics, namely interoperability, linking across domains, and logical inference and proof.

Semantic web technologies have two main components, namely Resource Description Framework (RDF)[24] and Web Ontology Language (OWL). In addition SPARQL represent the query language to navigate in RDF (Harris and Seaborne, 2013).

RDF is a graph based representation schema that can be used to represent both physical entities and digital resources (Klyne, Carroll and McBride, 2014). RDF representation is based on triples (i.e. a 3-place tuple) composed by three fields, namely subject, predicate, and object (Figure 28). The result is a graph based representation where each nodes of the graph presents a concept or object in the real world identified with a Unique Resource Identifier (URI) (Pauwels, Zhang and Lee, 2017).

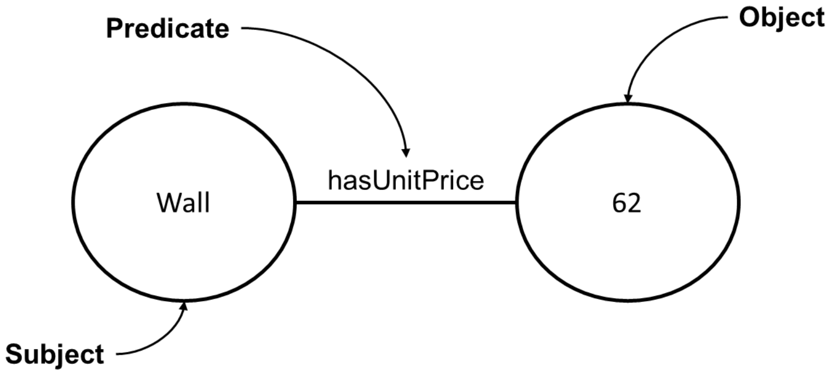


Figure 28. Representation of a triple

Compared to property graph schemas (chapter 2.2.3), RDF graphs offer the possibility to define a more complex and expressive schema (Hecht and Jablonski, 2011).

[24] *RDF can be seen as a special form of graph database (labelled oriented graph)* (Bouhali and Laurent, 2015)

OWL provides the basis to allow the integration of rules and proofs in RDF defining representational mechanisms for a detailed description of concepts and properties related to real world objects (Motik *et al.*, 2012). Practically, as an ontology[25], using OWL are defined what kind of things exist in a specified domain.

The definition of statements according to OWL allows the activation of inferences[26] from the machine that can automatically infer information that are not directly imputed in the graph. However, the reasoning behind the definition of all the inferences inside a graph schema can require too much time in comparison to the needs of use of the data in the specific domain of application. Hence, even if OWL can be defined in a full schema with powerful capabilities in the expression of concepts, there is the need to reduce this expressivity according to an increase in the performance of the system. It has been defined a different syntactic subset of OWL that are suitable for different purposes (Figure 29)

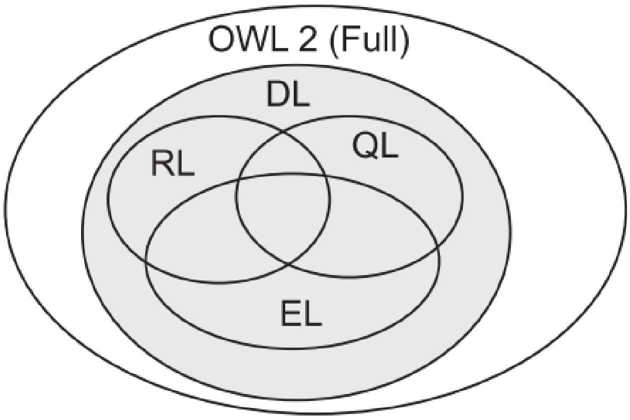


Figure 29. OWL syntactic subsets – from (W3C OWLWG, 2009)

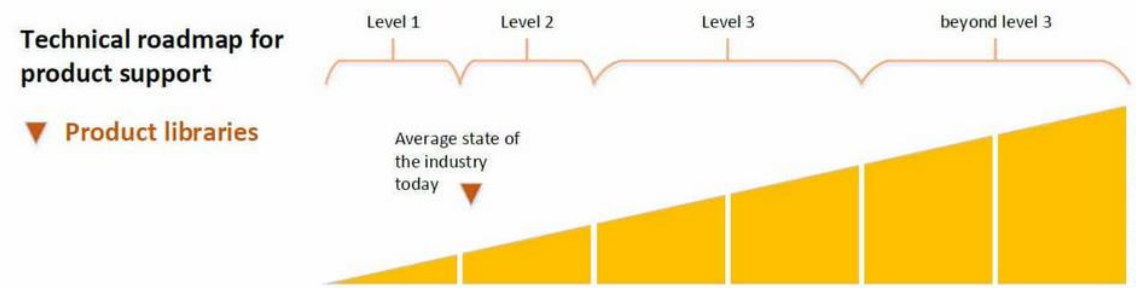
In several sectors including the AEC/FM one, exist many domain ontologies designed to satisfy the information requirements of specific use cases. For example the LOV (2018) website is a repository where are collected ontologies used worldwide. However, the identification of effective means to reuse the existing ontologies to fulfil the requirements of the sector is still a working area (Rasmussen *et al.*, 2017; Schneider, 2017)

The central role of semantic web technologies in the construction sector is highlighted in the Building Smart technical roadmap for product support that includes in its future development the use of these technologies (Figure 30). Several studies starting from the work of Beetz, Leeuwen and De Vries, (2009) investigated the possibility to translate the IFC standard into OWL. Nowadays, Building Smart (2018b) proposes a reference version of an ifcOWL ontology. Moreover, Pauwels and Terkaj (2016) described an automated method to generate the Building Smart's ifcOWL ontology starting from the IFC EXPRESS schema (IFC-to-OWL converter). Nevertheless, other studies are under development to explore different data schemas and different technologies, e.g. using other kind of NoSQL graph databases (Ismail and Scherer, 2018).

[25] *An ontology O is a tuple $O = (S, A)$, S being its signature (conceptual entities used for knowledge representation) and A its associated set of axioms (expressed in a specific ontology language or knowledge representation formalism).*

The signature of an ontology is defined as the union $S = C \cup I \cup P$ of the ensemble of classes C, instances I and properties P (Stephan *et al.*, 2011)

[26] The mechanism that allows deriving new assertion based on existing axioms according to rules (World Wide Web Consortium (W3C), 2015)



Main theme	**sheets**	**vendor/market specific**	**static open library**	**parametric open library**	**cloud library**	**...**
Working means	2D / 3D drawings	downloadable components	open, online product libraries		semantic search in the cloud	
Standards, open formats	dxf, dwg, pdf, (skp, 3ds)	none	ifc's mvdXML	parametric ifc's mvdxml / RDF	OWL / RDF	

Figure 30. Building Smart technical roadmap for product support (Building Smart, 2014)

3.1.4. Cloud based solutions

According to the National Institute of Standards and Technology (NIST) cloud computing can be defined as *"a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g. networks, servers, storages, applications and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction"* (Mell and Grance, 2011). Cloud computing can be divided into four main categories: namely private cloud (i.e. cloud for exclusive use by a single organisation comprising multiple consumers); public cloud (i.e. cloud infrastructure provisioned for open use by the general public); mixed or hybrid cloud (i.e. cloud infrastructure as a composition of two or more distinct cloud infrastructures); and community cloud (i.e. a *"cloud infrastructure provisioned for exclusive use by a specific community of consumers from organizations that have shared concerns")* (Mell and Grance, 2011). Furthermore, can be identified three service models, namely software as a service (SaaS), platform as a service (PaaS), and Infrastructure as a service (IaaS).

Nowadays, the majority of the advanced solutions for information exchange are cloud based e.g. (Allplan, 2018; Autodesk, 2018a, 2018b; Graphisoft, 2018). However, these solutions are usually based on an IaaS concept and the information production (e.g. the modelling activities in BIM) still happens on specific software that runs on individual devices.

A different line of research is based on the development of SaaS solutions where the visualisation and production of the information can be done in cloud environments (Cloud-BIM), (Chuang, Lee and Wu, 2011). Cloud-BIM technologies can be considered as enabling tools that can deal with the standalone nature of traditional BIM (Chuang, Lee and Wu, 2011; Wong *et al.*, 2014). One of the main advantages of these solutions is the sharing costs of software facilitating the inclusion of SMEs in processes that involve the use of onerous instruments as in the case of BIM. However, there are several open issues due to the variety of tools that can be used in the process and the share of information in the working phase that can cause misunderstandings and reduce the willingness of individuals in using the solution.

Some private companies are working on the development of solutions based on this principle. An example is the Project Quantum announced in 2016 by Autodesk (Day, 2017). The project aims to define a cloud environment offering specialised services that can be used by sector stakeholders instead of monolithic programs installed on personal machines. The development of the project foreseen the use of git technologies to address the issues related to sharing working in progress information, allowing individuals to hibernate the working data and upload the updated one only when the working activity is concluded. However, the project is still under development and no detailed descriptions of the architecture and processes are available. Furthermore, it is still focused on the project level (at least as a service for the sector) without a vision about the possible applications on data integration between multiple projects and multiple data sources.

The introduction of cloud-based solutions can open different interpretations in the role of digital information models along the processes of the construction sector. In fact, the use of SaaS paradigm, usually requires the centralisation of information models (i.e. set of structured and unstructured information containers (ISO, 2018)) used and managed through dedicated applications that can act on different levels (granularity) of the information containers. The introduction of cloud-based solutions for the development of a CDE can generate ambiguities in this interpretation. In fact, the information models are generated with stand-alone software and shared in cloud environments where the information are centralised. Nevertheless, according to the principles of a CDE - presented in chapter 3.1.1 - this environment must be used to generate useful information (e.g. approval procedures, review, lesson learned, etc.) that enriches the information models. Hence, in both cases building information models can be interpreted as components of a shared and related environment where other information containers can exist to constitute the information models of the project, the asset, etc.

3.2. Standardisation and management through code

3.2.1. Communication and codes

IFC covers only the transport of data and information in the process. Hence, the only use of IFC is not sufficient to manage the variety of information produced in a construction project (Fuchs, Katranuschkov and Scherer, 2010). The introduction of specific classification systems based on codes is required. A code is a *"sequence of characters, often a mnemonic, having defined meaning when interpreted in the context of the field in which it is entered, used to concisely convey meta-data"* (BSI, 2007). This definition highlights the context-based character of codes and the inclusion of specific meta-data in the code.

In the introduction of codes supported by digital tools, it is fundamental distinct between classification and coding activities based on the context of application and the code embedded in the digital tools and used by these last to run. In fact, in the generation of object base models, such as building information models, every object is characterised by a unique code or ID. This characteristic is maintained also in the use of IFC where a Globally Unique Identifier (GUID) is associated to each object in the translation between native format and IFC format. These codes are generated and managed in an automated way by the machine following the rules defined in the data schema embedded in the specific tool.

However, these codes are not related to the characteristics of the specific object in the reference domain. Consequently, an object with the same characteristics in two different projects (or models) will have a different code. In the same way, similar objects in the same project will have different codes[27]. Hence, even if there is an inherent codification automatically provided by digital tools, it cannot be used to share information according to the needs of a shared context based semantic structure.

[27] *This statement is true for BIM authoring tools that are not type-based. In the case of type-based tools, there are two inherent codes, one defined for each instance and one for each type, allowing the identification of all the instances of the same type. However, this do not guarantee the usability of this code for context based purposes because they are not scalable according to the project requirements.*

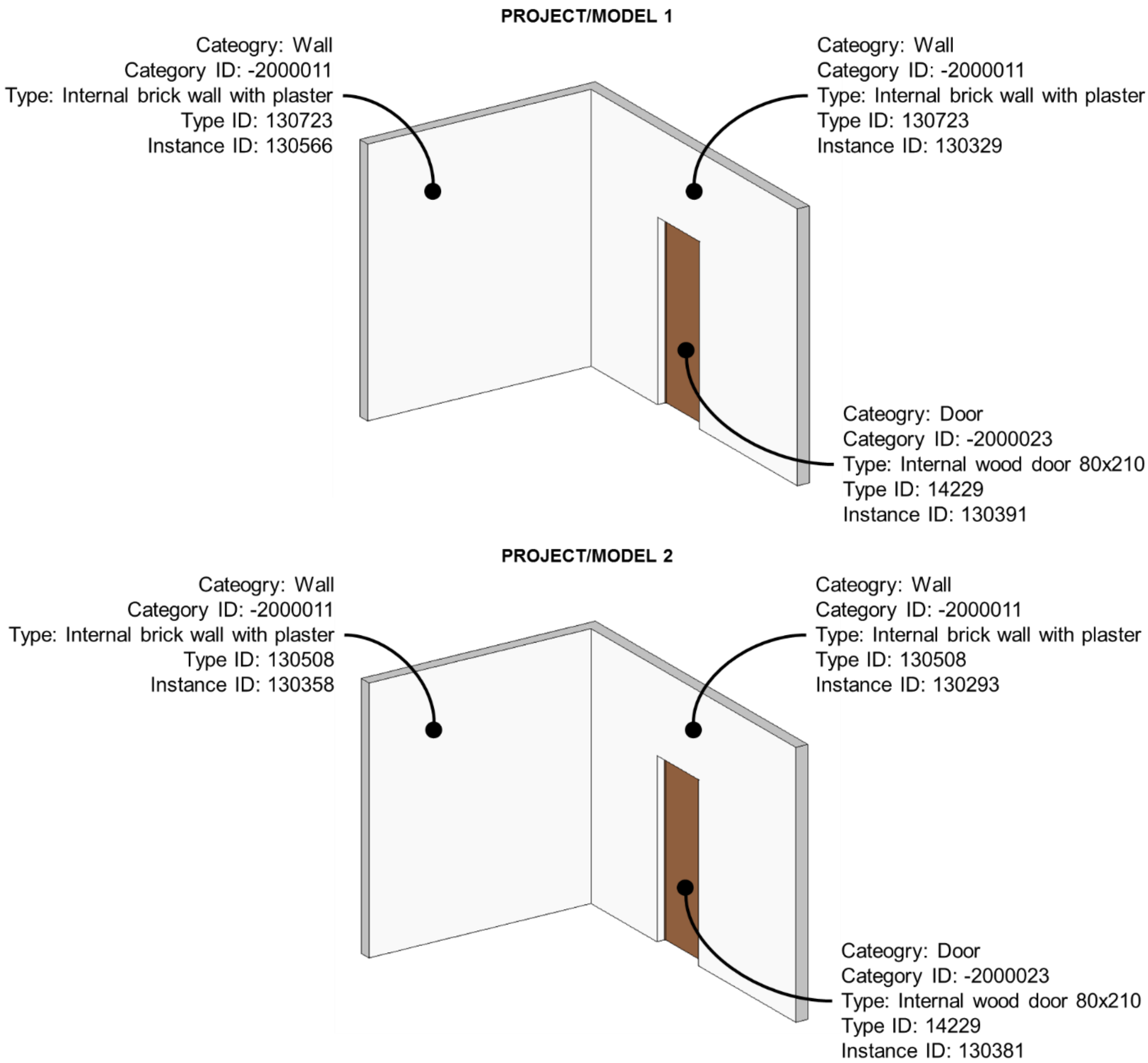


Figure 31. Visualisation of codes embedded in a type based BIM authoring tool

Figure 31 shows the behaviour of codes generated in a type-based BIM authoring tool[28]. The images show two identical design reproduced in two different models. Categories are imposed and fixed in the tool, thus the same category has the same IDs in every project and/or model. However, this is not applicable in the case of types and instances highlighting the semantic issues above mentioned[29]. Note that door type IDs remain the same in the two models because has been used the standard door type provided in the tool. However, this kind of objects are rarely used in the practice.

Hence, even in the implementation of BIM remains the need of overcome these limits. The reference solution in the industry is the introduction of one or more shared classification systems used to identify the elements composing a project. The effective use of a classification system allows the definition of a parallel language that can be used to guarantee the information exchange between different stakeholders involved in a project.

At international level has been defined several classification systems, developed to address specific requirements of the construction sector in the different phases of a project. Starting from the

[28] *Autodesk Revit has been used for the representation.*

[29] *Walls are special objects in Autodesk Revit that cannot exist as stand-alone type objects outside of a project. Hence the type IDs are different. Nevertheless, users do not have control on these codes that are managed directly by the machine.*

ISO 12006-2 (ISO, 2015) where it is defined a framework for the development of a classification system for the construction industry, several specifications have been defined.

The implementation of ISO 12006-2 in the UK is published under the name of Uniclass (also known as Uniclass2 in its last major release in 2015) (Delany, 2018). Uniclass is divided into a set of tables defined accordingly to a hierarchical structure (Figure 32). In the tables are included complexes (i.e. the description of a project in overall terms), entities (i.e. the area where different activities occurs), activities (i.e. the definition of activities to be carried out in complexes, entities or spaces), spaces/locations (i.e. spaces where one or more activities can take place), elements (i.e. the main components of a structure), systems (i.e. the collection of components that go together to make an element or to carry out a function), and products (i.e. the individual products used to construct a system).

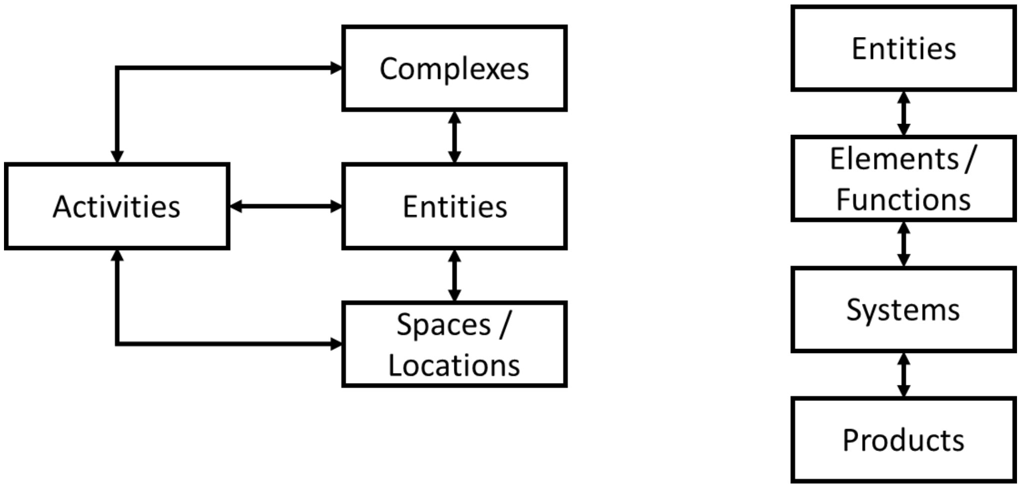


Figure 32. Table structure of Uniclass 2015 from (Delany, 2018)

On the other side, OmniClass is a classification system for the construction industry based on ISO 12006-2, created and used by the North American AEC industry (OmniClass, 2013). It consists of 15 tables, namely construction entities by function, construction entities by form, spaces by function, spaces by form, elements, work results, products, phases, services, disciplines, organisational roles, tools, information, materials, and properties. OmniClass incorporates other extant systems as the basis of many of its tables, namely MasterFormat for work results (CSI and CSC, 2016) and UniFormat for elements (CSI and CSC, 2010).

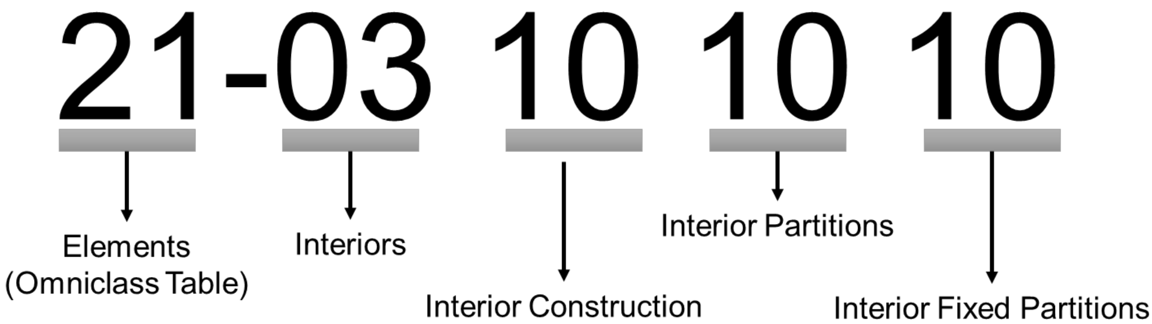


Figure 33. Example of code generated using OmniClass (Elements Table)

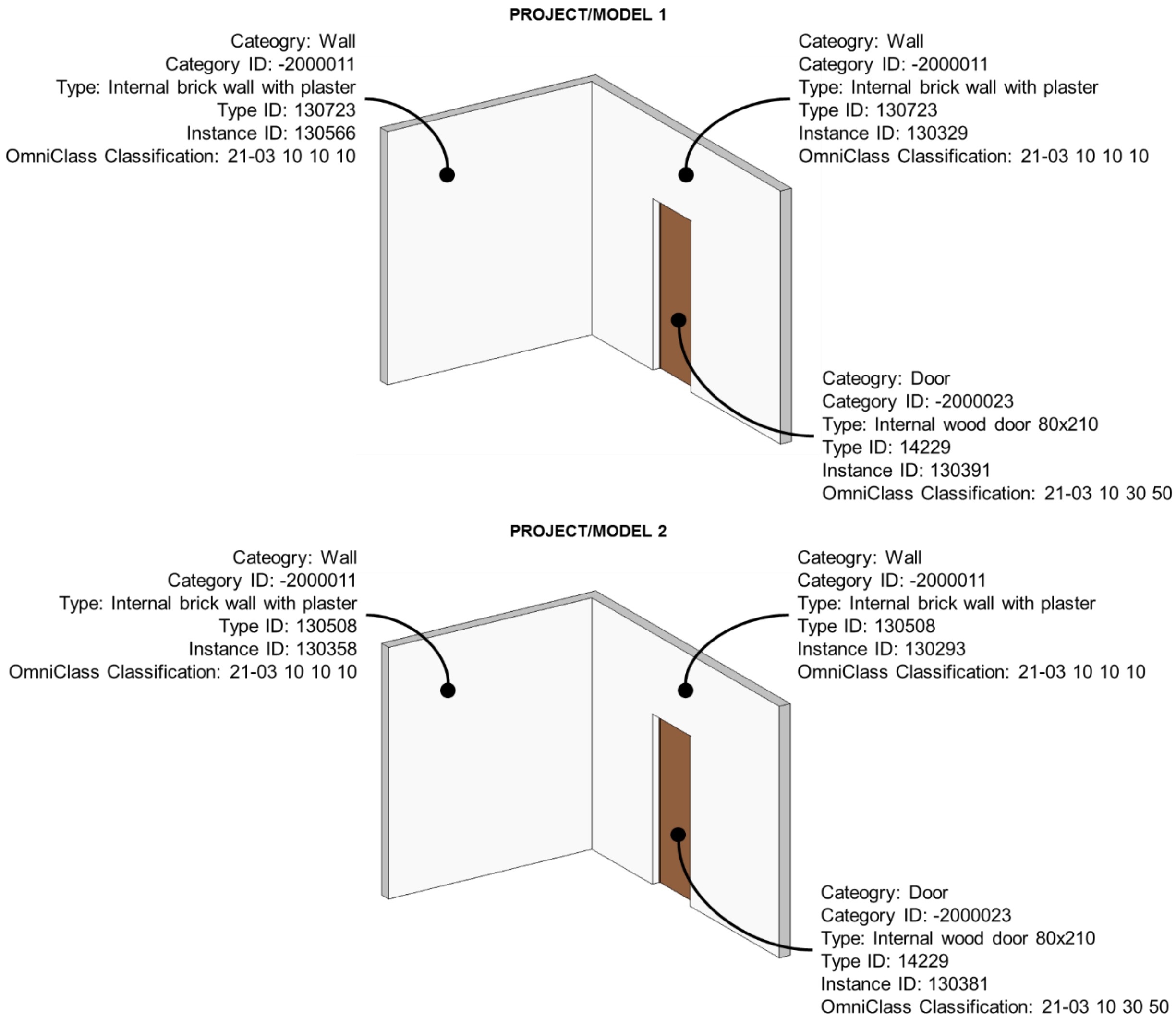


Figure 34. Visualisation of tools embedded codes and of OmniClass code (this last remain stable between the different models)

However, according to the use and objective of each project, the context usually requires a specific adaptation of the general structure offered by the existing classification systems producing a myriad of "ad hoc" versions that are usable only in restricted contexts (or in the specific company). Furthermore, classification systems cover only a restricted part of the project information with the consequent definition of project specific codes (BSI, 2007). This statement introduce a not negligible critic point. While the codification systems can be implemented in a specific project for the information exchange process, they can become useless in the management of company or domain information where there is the need to aggregate information provided by multiple projects.

Usually, in a BIM process the definition of classification procedures requires the introduction of specific informative fields compiled with the classification codes and associated to the objects used for the development of building information models. This activity is usually developed manually and can produce several problems including information inconsistency and incoherence. Figure 35 represents the classification schema highlighting the process where existing attributes associated to objects, models, and documents are translated in a defined code that is used in the information exchange process (including the coordination of the information in the CDE).

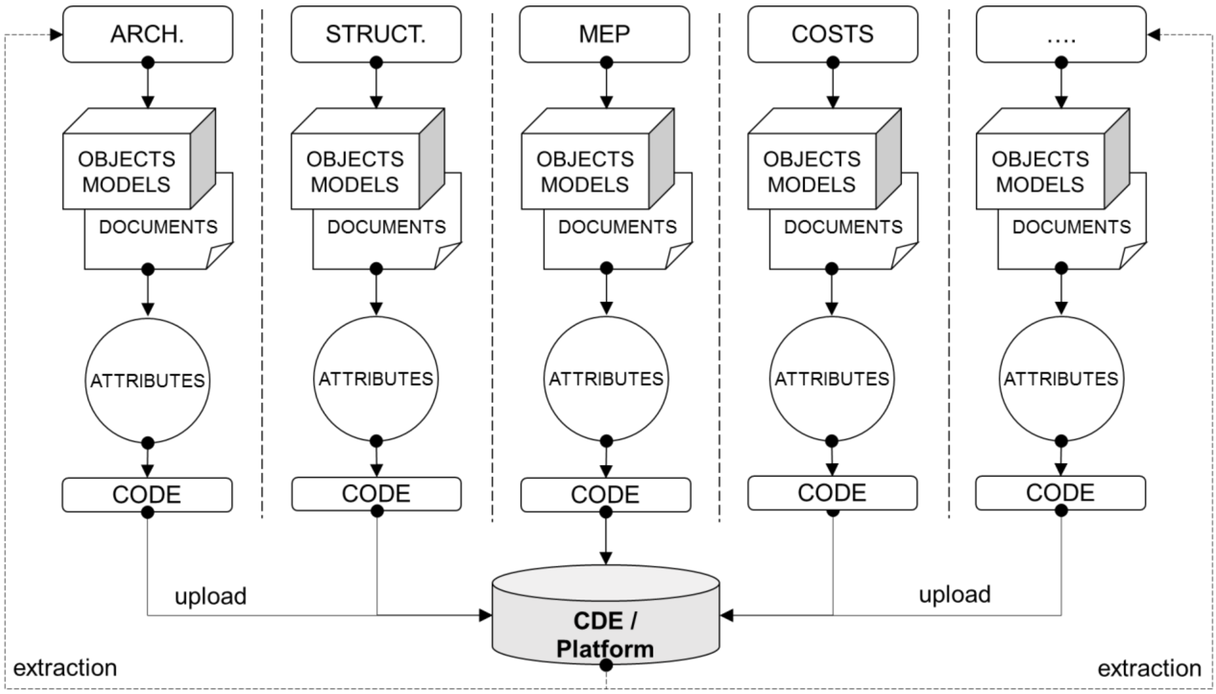


Figure 35. Actual code-based schema

3.2.2. Semantic enrichment and rule-based approaches

Nowadays, thanks to the progressive introduction of advanced techniques in the information and communication technology field, such as expert systems and machine learning, there are several reasons to explore different means to code and/or classify objects, automating the process. Recent research started exploring the concept of semantic enrichment for building information models. Semantic enrichment of information models refers to the process that automatically or semi-automatically introduce new facts about building objects in an original information model by means of an expert system inference rule engine. The inference rules that characterise the expert system embed a domain-based knowledge that allows the characterisation of objects in the context of use. Belsky, Sacks and Brilakis (2016) proposed the Semantic Enrichment Engine for Building Information Modelling (SEEBIM) that represent one of the first system developed in this direction. This study - further developed by Sacks *et al.* (2017) - focuses on the concept of made explicit information that is implicit in the building model allowing e.g. the automatic classification of objects according to their characteristics. Thus, instead of introducing manual codification procedure, an expert system is able to interpret the attributes of a specific object and characterise the object according to specific needs (e.g. classification or querying). It is worth noting that the attributes can be distributed between different sources of information and interpreted by the expert system enriching the building information model as explained by Sacks *et al.* (2017).

Figure 36 represents these principles incorporating it in a process based on the use of a CDE. The subjects involved in the process, and the instruments that these subjects use, require specific views on the information produced. Hence, the introduction of MVD (see chapter 3.1.2) and/or codes. The limits of these approaches - as explained in previous chapters - can hinder the effective interoperability and collaboration in BIM processes. The use of expert system can limit the need of manual classification of objects, identifying these last automatically according to the information contained in the building information model.

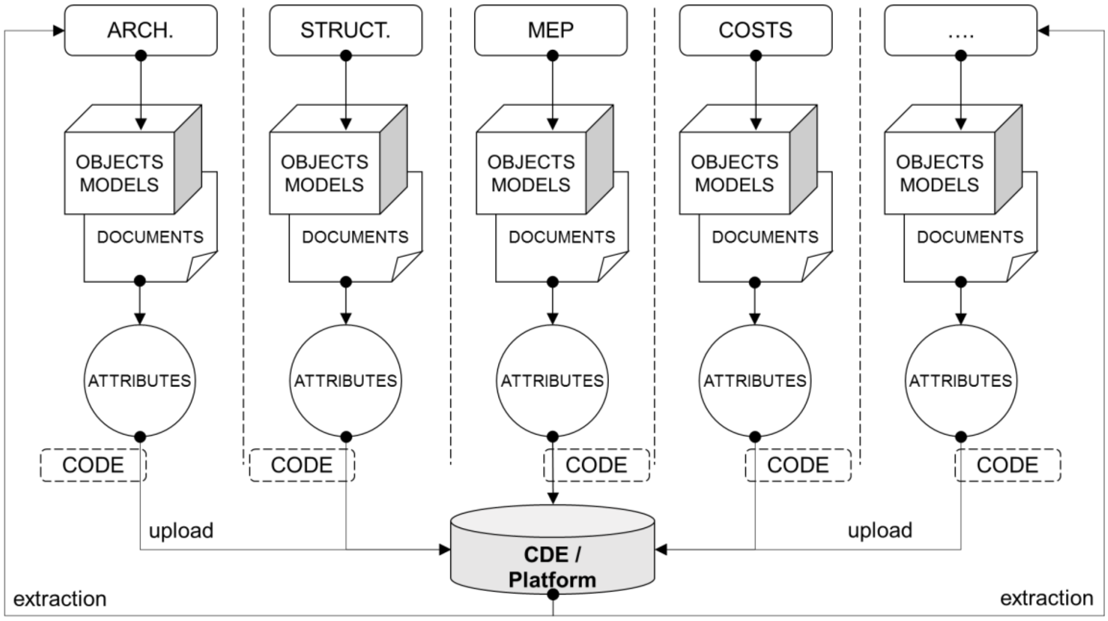


Figure 36. Attribute base code/classification schema

Starting from the same issues, researchers explored the use of semantic web technologies to generate building model views. For example, Farias, Roxin and Nicolle (2018) explored a rule based methodology to extract a sub-set of a building information model according to specific needs. The main difference between this approach and the one proposed by Belsky, Sacks and Brilakis (2016) is that in the former the building model views are generated querying the model, i.e. the information are inferred following the specific query and are not materialised in the model. In the second, it is defined a "hard" passage that materialise implicit information in an enriched model. Nevertheless, both methods are based on rules-based reasoners highlighting a common issue that is the reliance on the quality of the input data. Modelling deficiencies and/or errors in the introduction of meta-data can produce drastic effects in the performance of the reasoners as demonstrated by the case studies presented in the cited articles. Moreover, the system can use only computable data and is not able to interpret the information contained in text documents developed in natural language.

To overcome these issues, in the context of the Italian standard UNI 11337, is under evaluation the introduction of hashtag technologies. In an experimentation presented by the author (Mirarchi, Pavan and Cianciulli, 2018), we demonstrated how the introduction of hashtag in text documents wrote in natural language can provide an effective means to link the relevant information with objects contained in building information models. Practically, the introduction of hashtags allows the materialisation of computable information in a text document that can be used by a semi-automated system to create the above-mentioned link. However, this approach does not resolve the issues related to data quality that remain a critical problem. In fact, as demonstrate in the following chapter 3.3, data generated during construction processes can present several quality issues.

3.3. The impact of Data Quality in the construction sector

3.3.1. Data quality

Real world data is dirty (Hernández and Stolfo, 1998). In general dirty data means either missing data or wrong data or non-standard representations of the same data (Kim *et al.*, 2003). The management of data quality represents a fundamental area to guarantee the quality of analysis and data uses (see chapter 2.2). Data quality can be defined as *"data that are fit for use by data consumers"* (Wang and Strong, 1996). Hence, the extent for which the quality of data needs to be evaluated depends on the context of use of the data. This evaluation can be based on more than 170 dimensions described among others by Wang and Strong (1996), Naumann and Rolker (2000), Delone and McLean (2003). Nevertheless, the most used objective dimensions are accuracy (i.e. the extent to which data are correct, reliable and certified); completeness (i.e. the degree to which a given data collection includes the data describing the corresponding set of real-world objects); consistency (i.e. the satisfaction of semantic rules defined over a set of data items); and timeliness (i.e. the extent to which data are sufficiently up-to-date for a task).

The implementation of BIM can provide effective means in the management of data quality, especially in terms of consistency and timeliness. However, the inclusion of geometrical and non-geometrical information using digital instruments impose new issues related to the verification and validation data and information quality. Building information models can provide information about the shape of an object, its identity, and its relationship between other objects in the project (e.g. wall and floor slabs are perpendicular to each other) (Tang *et al.*, 2010). The complete omission of this information would bring essentially to a CAD model (Tang *et al.*, 2010). Hence, the assessment of data and information quality is a critical task in BIM processes.

However, in the construction sector - in different project phases - the requirements associated to data quality dimensions can change. Focusing for example on completeness at the design and construction level, the absence of some non-geometrical information associated to the geometrical one can be accepted in terms of data quality. In fact, the information is contained in the drawings and usable by the final users who can interpret the drawings. However, when this information is transferred to a database environment for data analytics applications, the graphical information is lost due to the difficulties of the machine in interpreting in computable terms the complex data structure behind the graphical representation. Hence, in data analytics applications the perception on data quality must change to focus on the inclusion of data that can be omitted e.g. in a specific phase. The same perception can be required in the case of interoperable processes as already mention in chapter 3.2.2. In that case, the integration of expert systems requires a specific level of data quality to guarantee the performance of the process.

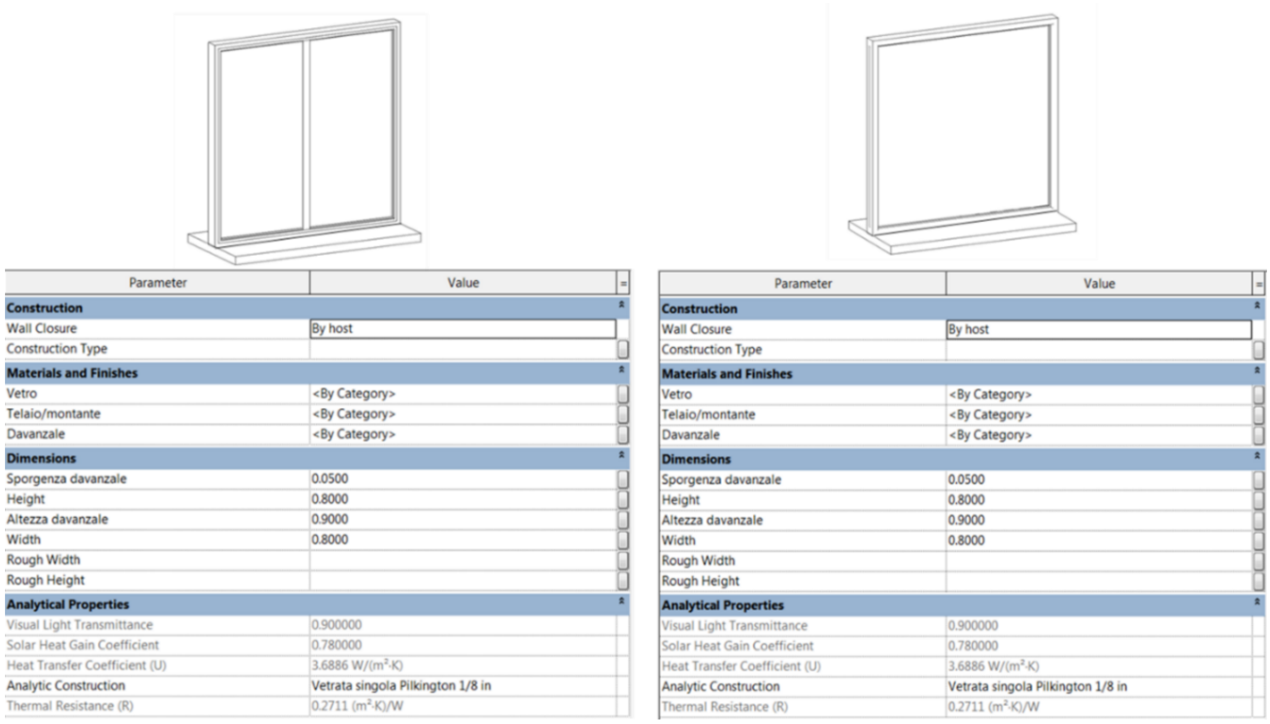


Figure 37. Representation of two windows geometrically different but with the same non-geometric (and parametric) information

Figure 37 shows an object developed using a BIM authoring tool. The two objects are characterised by the same non-geometric information (including the information associated to geometrical parameters). However, they are different in the graphic representation.[30] Humans can easily understand the difference between the two objects, thus in the design and construction phase the information is clear. However, reading the computable information the machine would recognise the two objects as equal providing erroneous results in possible analysis developed on the project data.

The definition of a model can be viewed as the simulation of real-world information in a digital environment. The real-world view derived from the observation, should not be in contradictions with the derived user view represented using the information contained in the model (Figure 38).

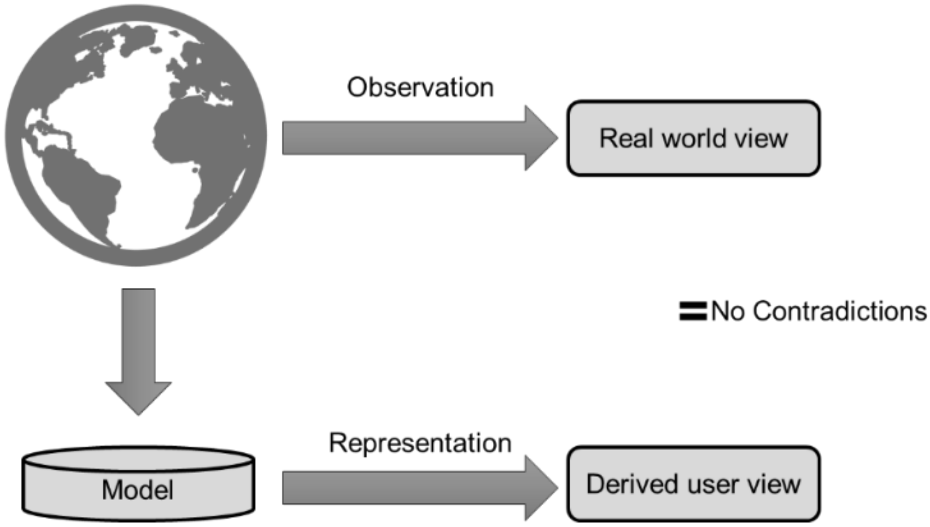


Figure 38. Data quality representation

However, the use of specific tools for the development of digital models can limit the ability to represent real world objects producing possible contradictions (or incoherence) between the real and the digital world. For example, in the development of building information models the classification[31] of objects is bounded by the inherent data structure of the specific BIM authoring tool. Table 3 reports the main object classes that can be defined in Autodesk Revit. All the objects that do not fit in the classification require a posteriori manual classification using specific means such as codes and/or rule-based approaches (chapter 3.2). This issue can prejudice the quality of the model due to the distance between the richness of real world objects and the limits of the available instruments used in the development of models.

Table 3. Main object classes in Autodesk Revit

n°	Class	n°	Class
1	Air Terminals	24	Nurse Call Devices
2	Cable Tray Fittings	25	Parking
3	Casework	26	Pipe Accessories
4	Ceilings	27	Pipe Fittings
5	Columns	28	Planting
6	Communication Devices	29	Plumbing Fixtures
7	Conduit Fittings	30	Railings
8	Data Devices	31	Roofs
9	Doors	32	Security Devices

[30] *Distinguish between geometric and non-geometric data in BIM might seem useless. One of the milestone points of BIM is that both geometry and other information are attributes of the model. Nevertheless, the data structure of some BIM authoring tool does not follow perfectly this principle and it may happen that the geometry and the related information are not aligned.*
[31] *Intended as the assignment of one object to one of the classes defined in the BIM authoring tool.*

10	Duct Accessories	33	Site
11	Duct Fittings	34	Specialty Equipment
12	Electrical Equipment	35	Sprinklers
13	Electrical Fixtures	36	Stairs
14	Entourage	37	Structural Columns
15	Fire Alarm Devices	38	Structural Connections
16	Floors	39	Structural Foundations
17	Furniture	40	Structural Framing
18	Furniture Systems	41	Structural Stiffeners
19	Generic Models	42	Telephone Devices
20	Lighting Devices	43	Topography
21	Lighting Fixtures	44	Walls
22	Mass	45	Windows
23	Mechanical Equipment		

3.3.2. Case study

With the aims of understanding the quantity, and the variety of data that can be encountered in building information models developed in the industry, was defined a practical case study. The experimentation was designed as follow:

- Selection of the concept design[32] for a civil building (the concept design was defined using a traditional design process).
- Definition of a general Employer Information Requirement (EIR) defining Level of Developments according to the USA BIM Forum (BIMForum, 2016), models required (namely architectural and structural), and type of coordinated information (namely schedule and costs).
- Selection of four design teams to develop the design according to the concept design and to the EIR.
- Analysis of the models developed by the different teams and comparisons.

The four design teams shared the same objective, i.e. propose the best solution in terms of costs, performance, design, etc. to win the competition. In this case study, the organisational structure defined by each design team was considered as a black box focusing on the final outputs (i.e. the building information models) and not on the generation process. This is in line with the scope of the case study that is the analysis of the data quality in building information models that can be generated in real applications in the construction sector.

To maintain the privacy of all the subjects involved in the case study, every reference to the original models has been omitted and the models are numerated from 1 to 4. Each design team has defined more than one model according to different internal rules. Hence, in the term "model" are comprises all the models composing the design developed by a specific team (e.g. "Model 1" includes the architectural, structural, and destination of use models).

All the models have been defined using the same BIM authoring tools. Thus, the first analysis focuses on understanding the number of informative fields that each design team has introduced in the model to satisfy the EIR. Table 4reports the comparison between the number of fields that are inherently defined in the BIM authoring tool in four classes of elements, namely walls, windows, pillars and materials, and the number of fields derived from the analysed models.

[32] *Are used the terms derived from the UK RIBA (RIBA, 2013), however the design development stage have to be interpreted according to the Italian legislation.*

Table 4. Identification of the number of information per object in the different models

	Walls			Windows			Pillars			Materials		
	S	**M**	**Δ**	**S**	**M**	**Δ**	**S**	**M**	**Δ**	**S**	**M**	**Δ**
Model 1	62	89	27	58	75	17	59	75	16	42	47*	5
Model 2	62	87	25	58	76	18	59	89	30	42	50*	8
Model 3	62	62	0	58	70	12	59	61	2	42	42*	0
Model 4	62	67	5	58	65	7	59	64	5	42	42*	0

Legend

S: number of parameter in a standard model
M: number of parameter in the analysed model
Δ: M-S
*: the number shows the material completed with physical and thermal properties (these are not included in all the materials included in the model)

This first analysis highlights the needs of introduce personalised fields to satisfy the project requirements. Moreover, the variety of the number of fields introduced by every design team demonstrates how the same objective can be satisfied following different paths. This underlines the complexity of the comparison between different models developed by different design teams even if referred to the same project.

The second step of the analysis consisted in the understanding of the quantity of data contained in each model. This information can provide a quantification of the efforts required in the development and in the maintenance of a model as well as the requirements in the verification and validation of the information by the client.

Figure 39, Figure 40, Figure 41, and Figure 42 show the relation between the number of objects and the number of fields respectively in models 1, 2, 3, and 4. Number of objects and number of fields are divided according to a logical (and technical) subdivision of the model. In the graphs are reported data for the architectural models (circle), the structural models (X), the materials (square), and the spaces (plus). Focusing for example on the architectural models, the number of objects per model ranges from around 3.400 to around 5.500, while the number of fields range from around 110.000 to around 200.000.

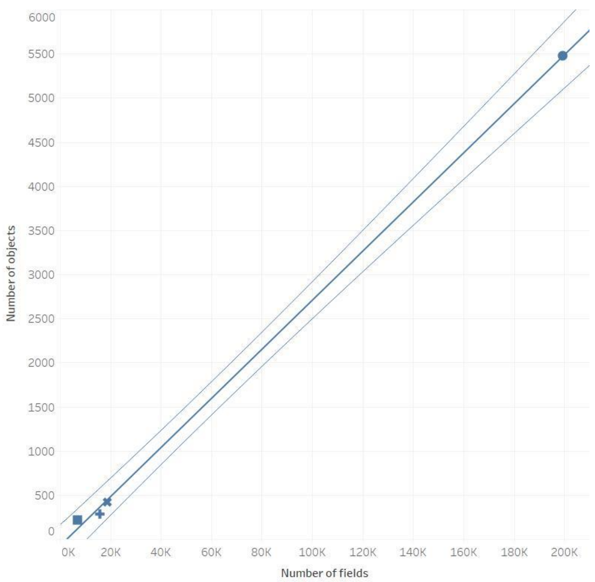


Figure 39. Number of objects and number of fields identified in Model 1

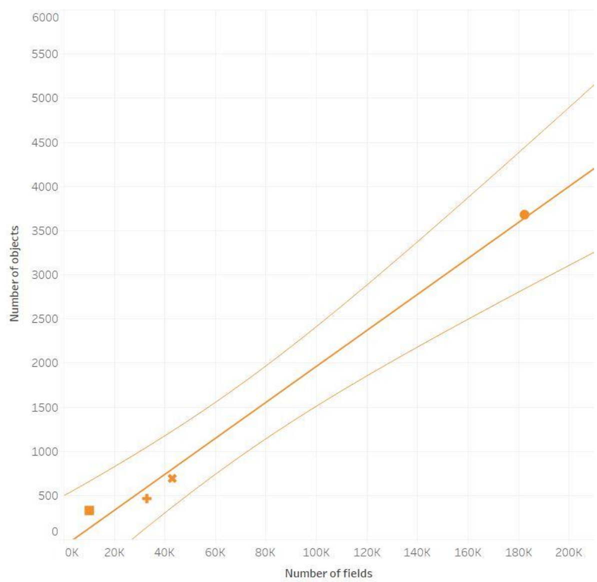


Figure 40. Number of objects and number of fields identified in Model 2

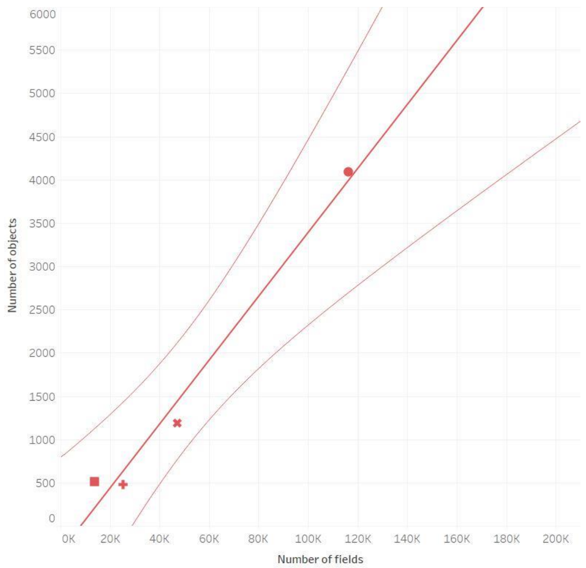


Figure 41. Number of objects and number of fields identified in Model 3

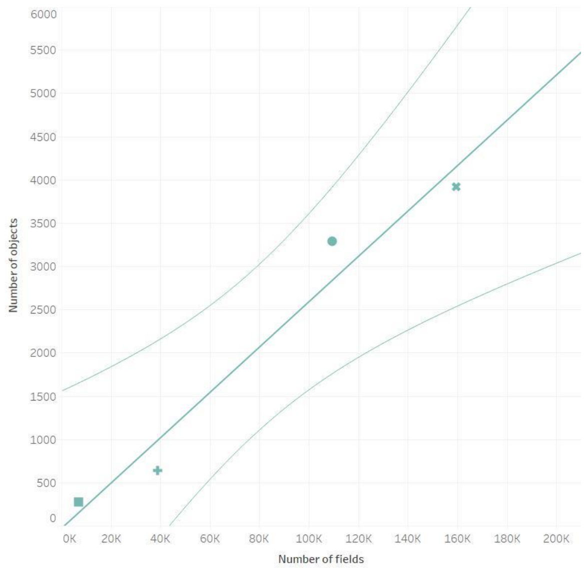


Figure 42. Number of objects and number of fields identified in Model 4

Defined the number of fields introduced by each design team, the total number of objects and the total number of fields in each model, the third part of the analysis focuses on the identification of the completion percentage of the fields. Figure 43 shows the results obtained analysing the architectural models. The graph compares the number of objects and the number of empty field for each model identifying the percentage of completion. The number of empty fields ranges from around 15.000 to around 87.000.

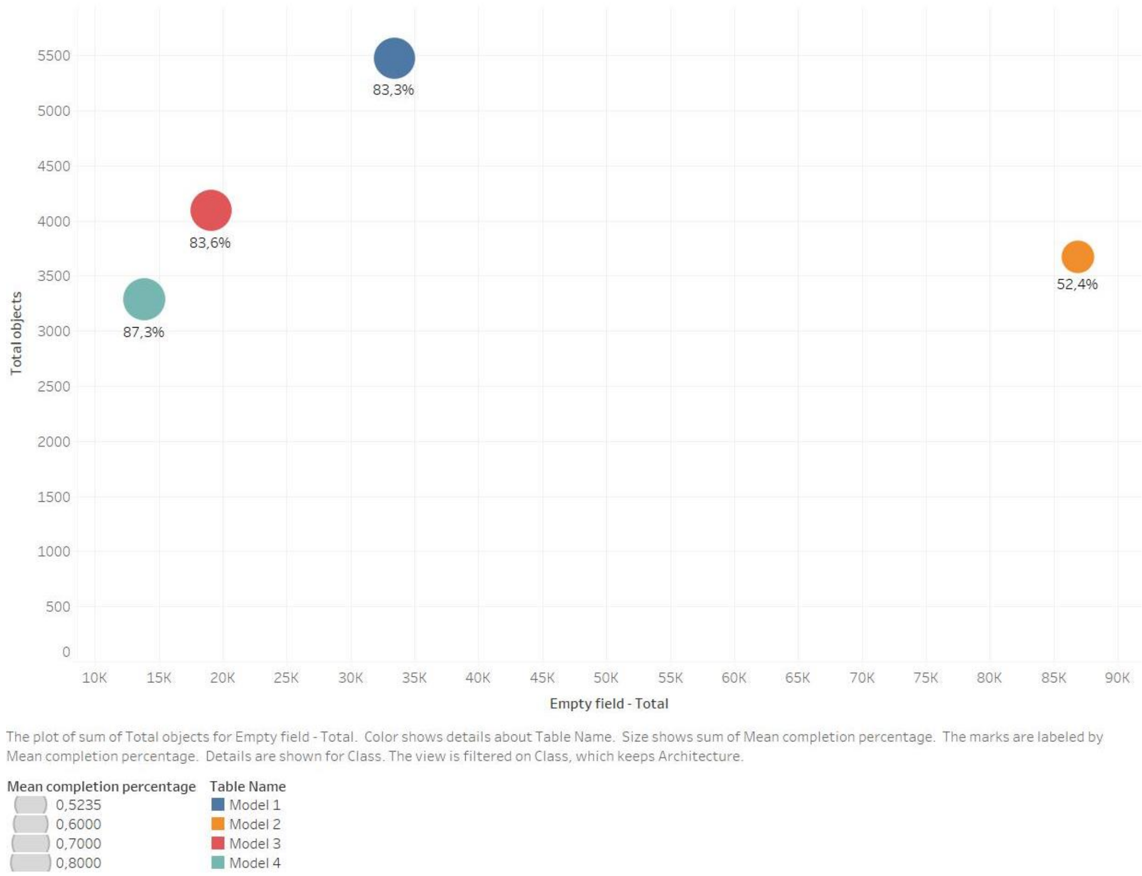


Figure 43. Number of objects and information fields in the architecture models

Empty fields can represent a problem in databases and, therefore, in building information models. On the one hand, all the fields - even if empty - need to be verified and validated. Hence, a high number of empty fields imposes an increase in the efforts needed in the activity of quality control on models. On the other hand, the evaluation of empty fields is not trivial. In the database research field, the "null value" problem has been studied long since. Zaniolo (1984) proposed two interpretation for null values. The *unknown* interpretation, i.e. the value exists but it is not known and the *nonexistent* interpretation, i.e. the value does not exist. Furthermore, understand if the null value falls in the former class or in the latter or if it has been generated due to an error from the designer is an issue with still no unique answer (the persons that develop the models do not have direct control on this interpretation).

3.4. Conclusions and focus of the research

The proposed analysis shows how the quality and reliability of data and information exchanged in a collaboration process represent a critical aspect. Furthermore, it underlines how this factor must be evaluated in the specific context of use of data and information. However, there are some general considerations that have to be taken into account in every use. Figure 44 shows the relation between data (or information) accuracy and its value. It highlights how a poor accuracy can bring to an inverse action in the use of data and/or information with a consequent lost in the value.

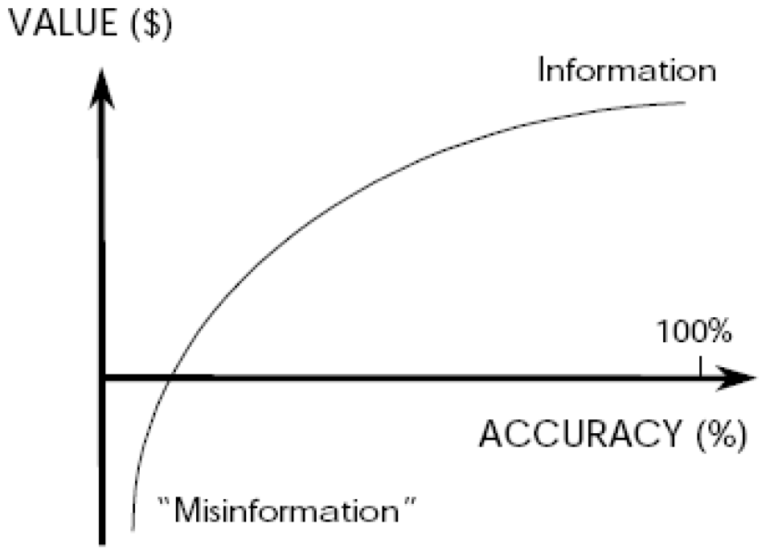


Figure 44. Relation between value and accuracy – originally from (Moody and Walsh, 1999)

Furthermore, the value of data and information changes according to the ability to combine different sources and find new relations (see KDD chapter 2.2.1). Figure 45 shows the relation between data integration and data value. The ability to integrate data provided by different sources augment exponentially the value of the data. Hence, analyse integrated datasets has more value than analyse single isolated datasets. This integration can be interpreted in terms of both data fusion (Shahandashti *et al.*, 2010) and data integration from different sources (e.g. customer information and sale information). This last interpretation is nowadays a central point of research as presented in chapter 3.1.3.

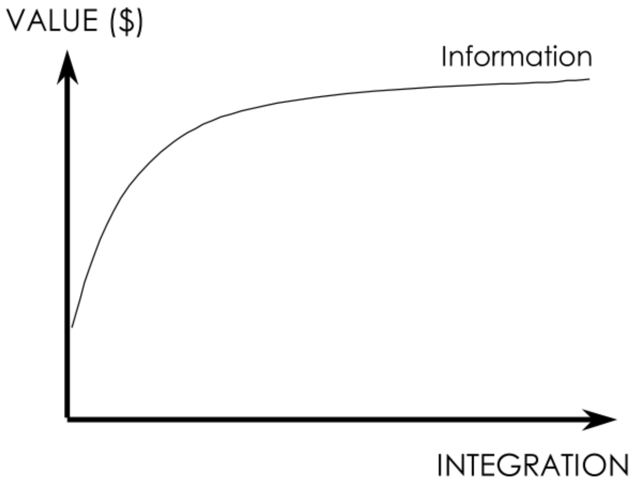


Figure 45. Value increase with integration – originally from (Moody and Walsh, 1999)

Furthermore, the Garbage-In-Garbage-Out (GIGO) principle must be considered. It express the relation between input, analysis process, and output in data analytics processes. The application of perfect algorithms to a garbage dataset, will produce a garbage result. In the same way, the application of garbage algorithms to a perfect dataset, will produce a garbage result.

Finally, the volume of data and information must be considered. Figure 46 shows the relation between information volume and value. It underlines the need to identify the right amount of information required for the specific use, because exceeding in the information collection can bring in a loss of value. This is mainly caused by the needs to process several useless information imposing unnecessary efforts and consequently unnecessary costs (see chapter 3.3.2).

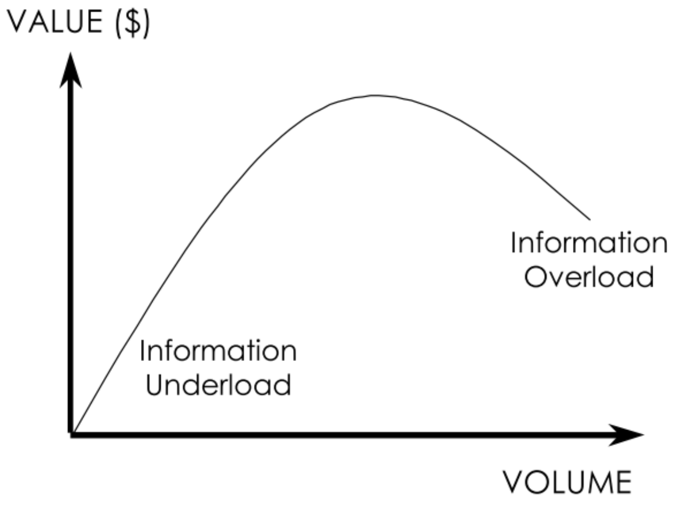


Figure 46. Relation between value and volume – originally from (Moody and Walsh, 1999)

However, this approach can change passing from information to data. While the collection of information must be fitted in the specific context of use, the collection of data can follow a different path, i.e. gathering as much data as possible identifying possible uses for the data in a second time. With the progressive decreasing of data storage costs this approach is nowadays widely applied by the major digital companies like Google, Facebook, and Twitter.

Finally, it is worth noting that information value, in contrast with other physical assets, increase with the increase of information use. On the other hand, unused information represents a lost in value due to the costs required to its generation and collection and the lack of value generation because it is unused.

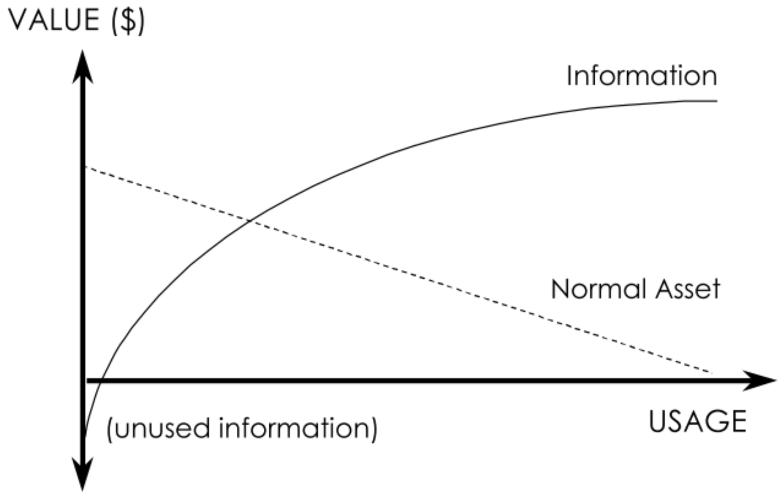


Figure 47. relation between value and information use – originally from (Moody and Walsh, 1999)

The framework presented in the next chapter is based on the optimisation of data quality procedure, on the integration of data provided by multiple sources and on the collection of the maximum volume of data with the aim of introduce this last in data analytics applications.

The analysis contained in chapters 2 and 3 paves the way for the definition and development of the framework described in chapter 4. Chapter 2 (with particular reference to the representation of Figure 11) clarifies the context where collaboration processes and the related technologies are embedded. The theoretical structure composed by three main level, i.e. project, company, and knowledge, is the main input to generate the multi-level approach proposed in chapter 4. Chapter 3 provides an extensive review of the existing technologies and processes related to collaboration and interoperability in the construction sector. Moreover, the detailed analysis on both the semantic enrichment of models and the data quality area provides the instruments to understand one of the existing research gaps in the passage from the project environment to the company or sector environment. This analysis creates the background to understand the motivations behind the case study described in chapter 5.2.

4. Framework of the knowledge network platform

"It is a capital mistake to theorize before one has data sources.
Insensibly one begins to twist facts to suit theories, instead of theories to suit facts"
(Arthur Conan Doyle, Sherlock Holmes)

Chapter 3 concludes the analysis of the state of the art for the development of the first objective of this research, i.e. the definition of the framework for a *knowledge network in the construction sector*. As described in chapter 1.2 the development of the proposed framework starts from the evidences retrieved from the literature review and from the theoretical background. In particular, the conclusions reported in chapter 2.1.2 constitute the theoretical basis that guided the development of the high-level framework and its refinements. This chapter proposes a further literature review focused on the requirements of collaborative environments critically revised according to the main theoretical evidences matured in the development of this thesis. Finally, through the integration between the theoretical background, the specific literature review and the observations of the industry, chapter 4.3 presents the researched framework.

4.1.Main vision

To provide a better understanding about the path that conducted to the definition of the framework presented in chapter 4.3, this chapter introduces the main vision of the project, retracing the logical path from the first intuition to the general architecture. The theoretical evidences reported in chapter 2.1.2 supported this path.

4.1.1. Single project environment

Starting from the model in its native form[33], it is possible to translate the database that constitute the model itself extracting (with a deconstruction of the schema in e.g. a simple .csv format) all the data referred to the model and reorganising the data itself in a different database schema (Pavan and Mirarchi, 2017). For example, any standard relational database can be transformed in a graph schema (DellaValle, Celino and Cerizza, 2009) as already demonstrated with the IFC-RDF translator developed by Pauwels and Oraskari (2015) (chapter 3.1.2). This process can lead to a new database that results more effective and efficient. The translation of the geometrical information is still an open issue due to its complexity as demonstrated e.g. in the translation from IFC to an RDF schema (Pauwels *et al.*, 2017). The new database can follow a different structure that is not bounded to the existing ones proposed by actual software (traditionally a relational database with a complex and inefficient structure). For example, starting from the use of ifcOWL, Pauwels and Roxin (2016) proposed a simplified structure to optimise the performance. This procedure can be performed through the introduction of a translation process – identified in Figure 48 as *"layer of deconstruction"* – able to explode the model, extracting, deconstructing and then reconfiguring all the data in it contained. In this first logical representation, it is assumed that the process works in only one direction, from the model to the data environment and not vice versa. This is coherent with the actual collaboration processes and with respect to the responsibilities of the individuals that develop the models in its native format. Furthermore, the focus of the research is on the analysis and the use of the data provided without the need of acting to modify the data in the virtual environment.

33 This implies that there are not changes in the model. These changes can be introduced during translation processes (for example from the native form to the IFC format) that are used improperly. See chapter 1.4.2 for clarification about the concept of "model".

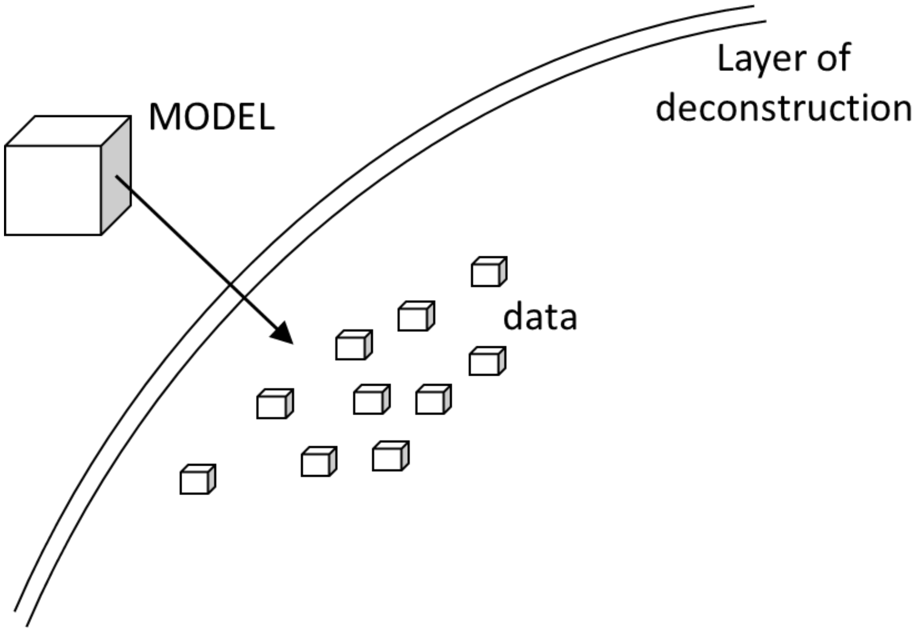


Figure 48. Model deconstruction[34]

The use of this structure provides two main implications:

- The deconstruction of the model allows organising data in a more efficient way, improving the performances in querying the model data and combining the queries itself;
- Data extracted and deconstructed are stored in an ad-hoc environment developed for sharing and linking data from different sources. This approach allows the association of information provided by sources that can be external to the model.

This last point is critical in the progressive digitalisation of the construction sector and of the entire industry chain connected to it. Indeed, there is the need to overcome the concept of BIM as a closed environment including a wider range of data interlinked to each other. Digital models that are often identified as BIM, to today, represents a single island of data in a wider project context that contains information coming from a myriad of different sources (Radulovic *et al.*, 2015). There are several examples such as information from the construction site, context, performances, standards, regulations, etc. For example, looking at the use of BIM in the field, there are several studies where the possibilities offered by the introduction of digital processes are demonstrated (Leung, Mak and Lee, 2008; Babič, Podbreznik and Rebolj, 2010; Wang *et al.*, 2016; Park and Cai, 2017).

However, the proposed solutions are often defined ad-hoc for the specific application and/or for a specific part of the process and are not inclusive with reference to the entire life cycle of the work. In many examples, the information from the construction site can be updated directly from the personnel on site improving the communication process and the collection of important information. Introducing this information in a more opened context, the same information can be enriched with data provided by different sources classified as *"reliable"*. Some experimentations in this direction have been developed by the Linked Building Data Community Group (W3C, 2018) linking data provided by building information models to data provided by Wikidata (Wikidata, 2018) or other open data sources (Rasmussen, Hviid and Karlshøj, 2017).

[34] In this image, the "model" box represents an information model such as a building information model. The "data" boxes represent the data extracted from the model. Focusing on a building information model, the data should not be confused with objects. Objects are still complex models while data represents the atoms that constitutes the overall model as well as the objects that compose the model.

Indeed, this vision is not new. Figure 49 shows the process involved in the translation of information in its native format to information in RDF schema then enriched from other information sources, proposed by Curry *et al* (2013).

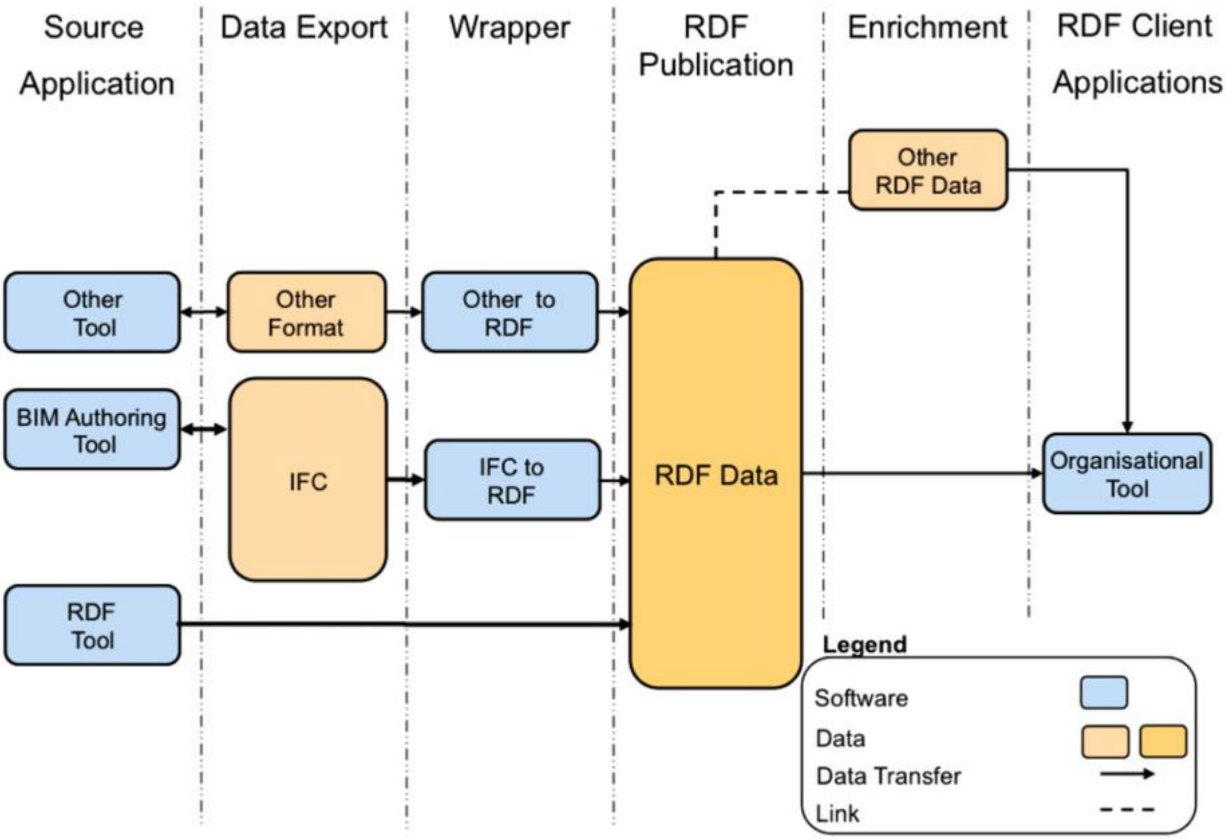


Figure 49. Process for the use of an RDF based environment for the information integration from different sources (Curry et al., 2013)

The translation process from the model in its native format to the final RDF based environment involves two passages that are from native to IFC and from IFC to RDF. In a dynamic work environment characterised by an high number of changes that many times need to be performed in restricted period of time, this approach seems to be not suitable. In fact, the effective definition of collaboration processes requires that the process can be executed in a time laps that is coherent to the requirements of the industry. Furthermore, the introduction of two consequent translation passages, without an intermediate verification step can produce incoherence between the native model and the final information published in the RDF format.

Figure 50 shows another example that is the simplified architecture of Bimserver.org (opensourcebim, 2016) proposed by Beetz *et al.* (2010). In comparison to the previous schema, the Bimserver platform introduces an interesting vision that is the definition of a dialog between the platform and other databases structures for the development of data analysis. However, the issues highlighted in the previous schema remain unsolved.

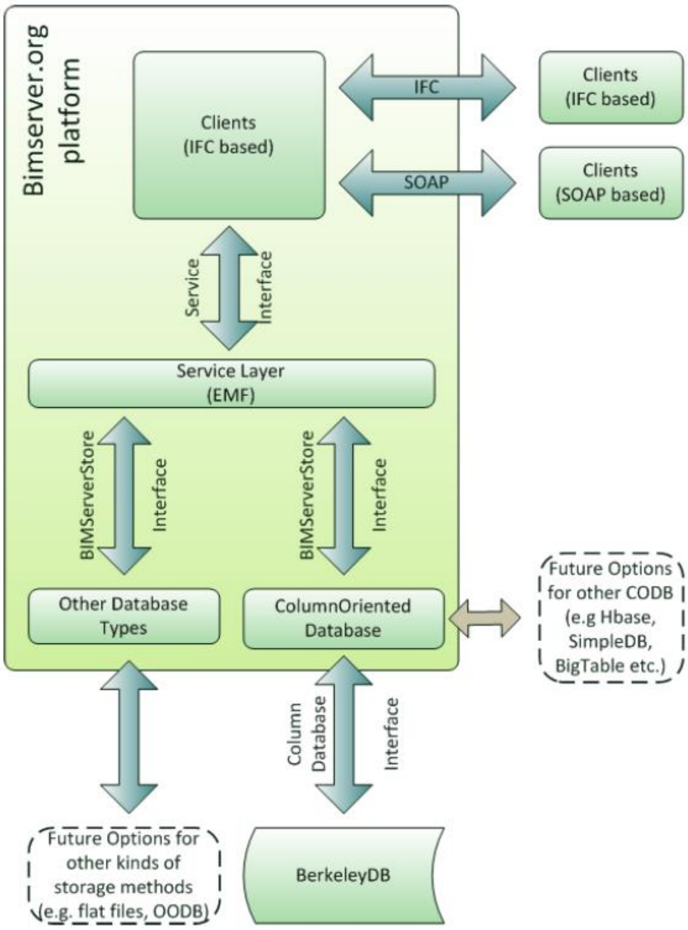


Figure 50. Simplified architecture of the Bimserver.org platform (Beetz et al., 2010)

Starting from this first analysis there are several issues that must be considered:

- The construction sector deals with one-of-a-kid product, in a project based environment that requires the identification of specific rules to design and collaborate. Each project can involve a different group of stakeholders that can change during the development of the project.
- Construction projects are characterised by high dynamicity. Hence, the time required to perform the passages in every collaboration process must be aligned with the natural time of execution of the tasks in the specific sector and/or project.
- The value of data increase according to its volume and the ability of merge different data sources. However, individuals are usually unwilling to share their data and more their information and/or their knowledge. Hence, the process must provide value for the users to push its use in the industry.

Figure 51 represents a multi-level approach that can provide a possible solution for the above-mentioned issues. The first collaboration level can be interpreted as a CDE that according to the different visions described in chapter 3.1.1, must integrate the appropriate rules of collaboration and verification defined at project level. This interpretation guarantee the dynamicity required by the sector addressing the first issue.

Furthermore, this first level could constitute both the driver to feed a company based and/or a national database (and form a basis for a wider platform at a higher level, Figure 55) and a driver to promote the diffusion of the use of digital processes and instruments in the sector. The use of a defined sharing environment is directly in line with the principles reported in the Italian framework document about industry 4.0 (X Commissione, 2016), where are included the lines of action established at European level (EU, 2016d, 2016c, 2016b, 2016a).

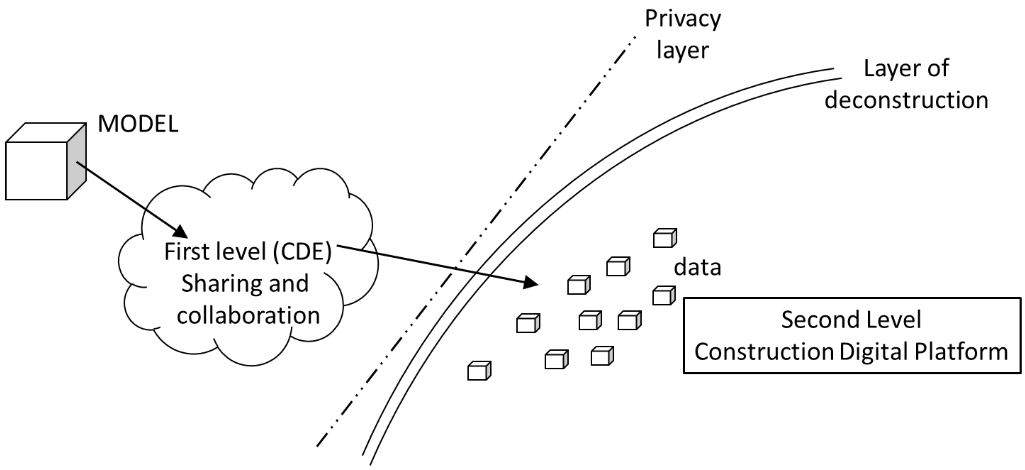


Figure 51. High level structure of the system

While the first level is devoted to information sharing and collaboration, including the temporary group of stakeholders created on the specific project, the second level can embrace a wider range of data including products data and external open data (Figure 52).

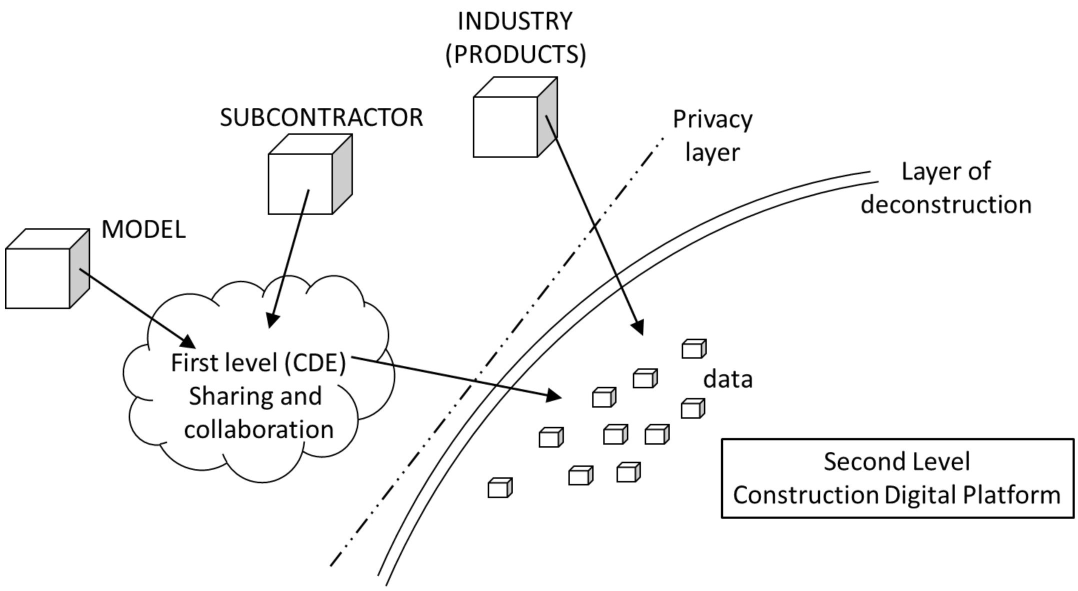


Figure 52. Example of platform schema including subcontractors (e.g. in construction activities) and construction components manufacturers

The platform, i.e. the structured data that it contains, can provide a basis for the development of new services that can improve the functionalities of the CDE defined at the first level. For example, using the data stored in the second level can be defined interactive verification processes of "data quality" acting on the CDE of first level providing a useful instrument able to support the design team. In the vision of a construction company, this aspect is for example applicable as a service for the validation of the information provided by the different actors of the project chain (subcontractors, manufacturers, etc. - Figure 52). Another aspect could be the inclusion of rules provided from standards and regulations that can be implemented in the second level and linked to the first level environment. The result is another kind of design supports as proposed e.g. by Pavan, Odorizzi, *et al.* (2017) and by Dimyadi, Pauwels and Amor (2016).

4.1.2. Multi-project environment

Widening the perspective with the inclusion of multiple projects managed by different project and execution teams (stranger to each other), can be highlighted the real benefits related to the proposed structure. The construction sector is characterised by processes of production and management by project, based on temporary coalitions of different organisations coming together to meet particular targets in a given timeframe (Pellicer, Pellicer and Catalá, 2009) (Chapter 2). Thus it is difficult to capture, store, and explore data and information produced during a project (Gann and Salter, 2000) due to the instability of the structure defined by the interrelations between the involved actors.

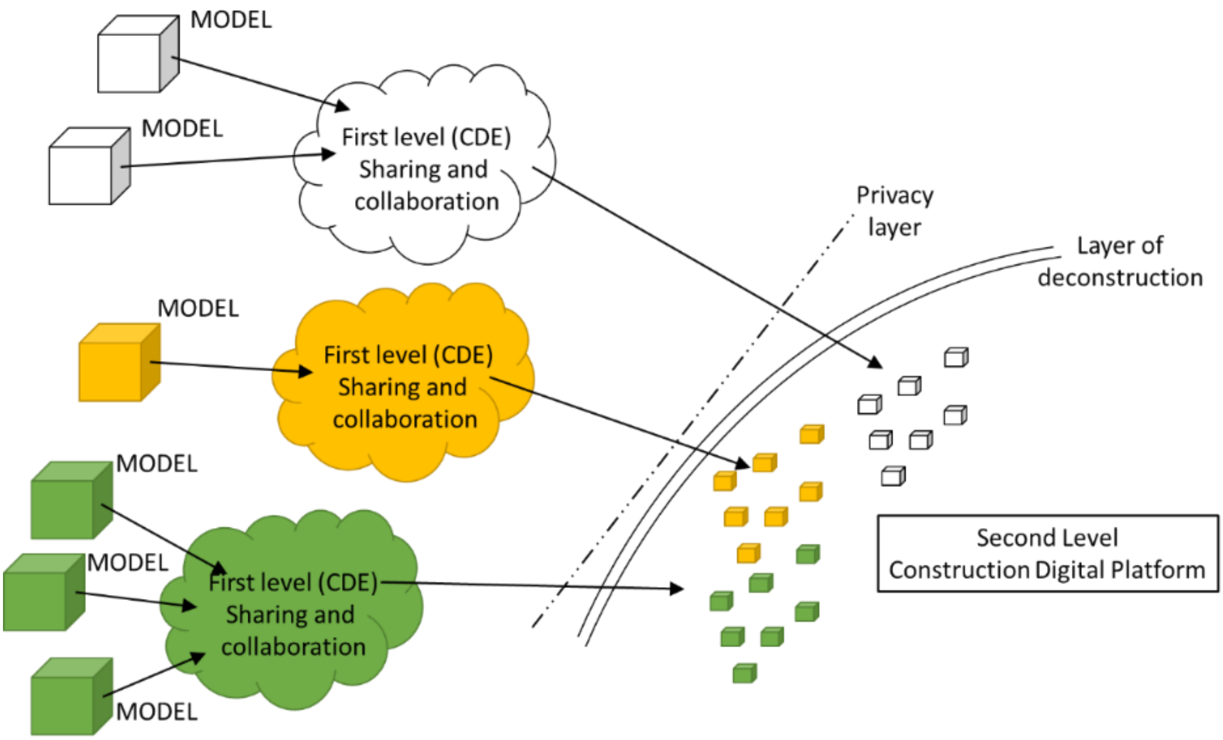


Figure 53. Platform use in a multi-project environment

Figure 53 underlines the framework of a multi-project environment. At the first level (CDE; sharing and collaboration), the sharing environments are separated to each other[35]. Passing the "privacy layer" where are filtered the sensitive information, you meet the "layer of deconstruction" that for each project deconstruct the model[36] and reorganise the contained data in a new database as proposed in the previous section. The second level environment collects all the information provided by different projects, transforming data that traditionally born and die on a single project, in useful atoms that feed a wider basin of shared information. A single project is for its nature isolated to the expertise of the personnel involved in the project itself. Thus, is required a proactivity of the information in order to promote the development of new systems of assisted design based on data (Pavan, Antisari, *et al.*, 2017).

[35] The platform can be de facto unique, but at the first level the different environments are isolated to each other and the access is regulated. Thus, each environment is at the first level an independent silo where can be established different rules according to the specific needs of each project. On the other hand, it is possible to image a distributed architecture, where the data physically separated but linked in a digital ecosystem.

[36] In this context the model is intended in a wider vision including the entire project model constituted by the geometrical model and all the information directly related to it also using external databases (In this vision, the entire CDE is the model itself). The deconstruction activity referred to a single model have to be extended to each model (Chapter 1.4.2).

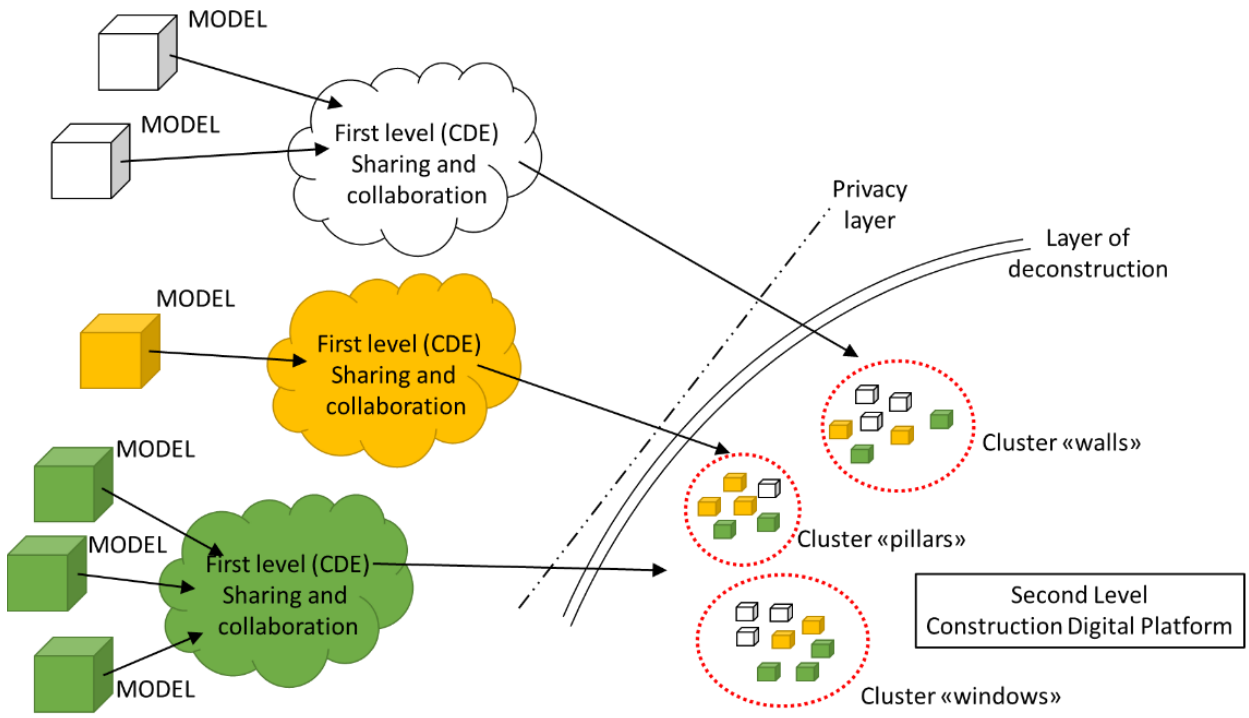


Figure 54. Clusters identification in a multi-project environment

The reorganisation of data allows the identification of clusters of elements provided from the different projects, creating in this way an integration mesh that is effective both in terms of efficiency (Sergey and Lawrence, 1998) and in terms of effective application in a fragmented framework like those of constructions. Clusters can be defined according to the rules embedded in the platform, thus a systematic a priori knowledge needs to be included in the system. During the querying activities that can be performed on the platform, this cluster structure provides an efficient starting point for the research augmenting the performance of the system (e.g. reducing the query time).

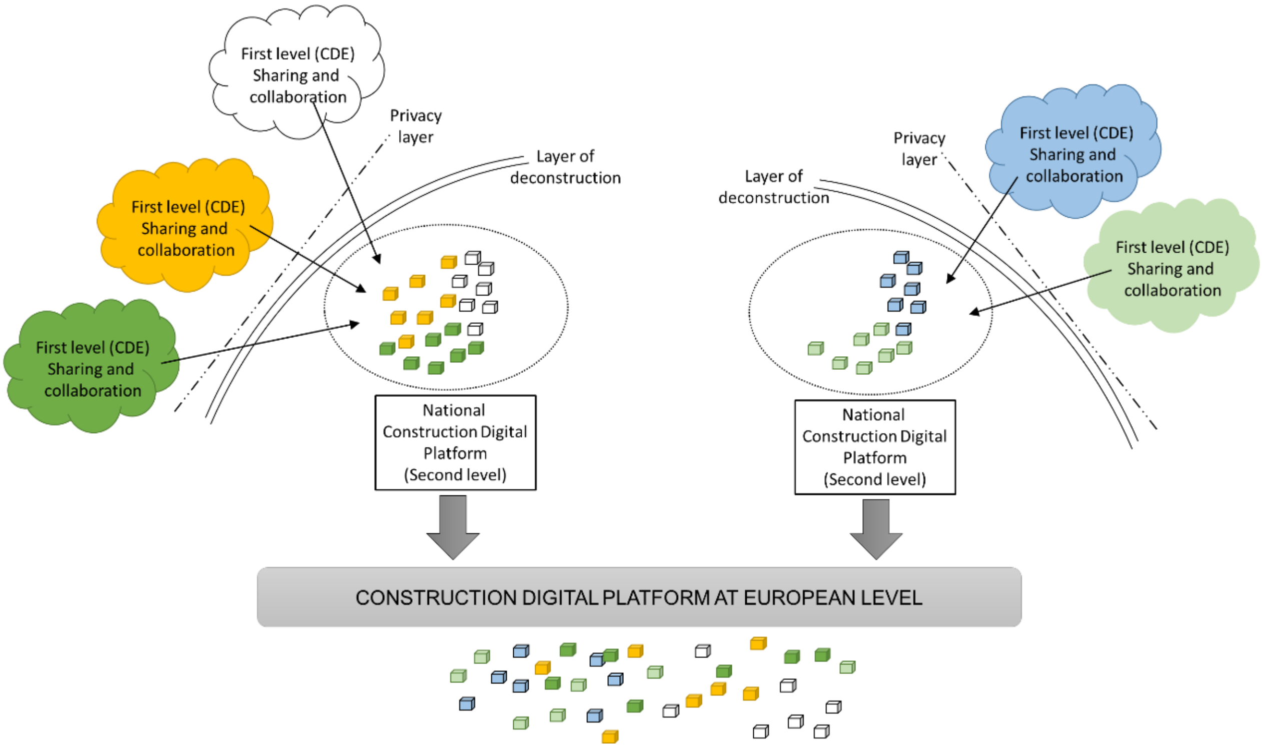


Figure 55. Architecture of the system including a Digital Collaboration Platform at European Level

Following the same principle, the proposed architecture can be extended creating several National Construction Digital Platforms which data can be collected in an inclusive platform at European level (Figure 55). Hence, the system can be interpreted according to a scalable multi-level paradigm that sees the identification of a project level devoted to information sharing and collaboration. A second level that can be read at a national scale, including the rules and the regulation of the specific nation that can include direct services for the stakeholders through the first level and provide domain specific statistical open data. Finally, a third level, that can act at European level as control room. Actually, can be identified more than three levels including for example a company level and/or a regional level.

4.1.3. Direct implications of this structure

The main advantage in adopting a multi-level structure is the disconnection between the velocity required during the execution of projects and the time required to process and use data and information. Moreover, sharing and collaboration need to maintain a link between the instruments used to generate the information and the tools and processes used to share the information. Hence, sharing and collaboration environments maintain the limits embedded in these instruments (e.g. limits of BIM authoring tools chapter 3.3).

A second level where the focus is on data with the objective of use and analyse the data, can break this bond allowing a more dynamic structure. For example, with reference to building information models, the integration of data in the second level can remove the semantic limits embedded in BIM authoring tools bringing to a complete simulation of real world objects towards the concept of *"digital twin"* (Matsokis and Kiritsis, 2010; Tao *et al.*, 2018). This aspect can open a complete new vision on the management of data and information according to different levels of granularity. For example, building information models are limited to the representation of building objects but are not able to represent the building as a whole, the territory where the building is constructed or the city around it, even if these components can contain critical information for the project. On the other hand, the use of Geographic Information System (GIS) can embrace these dimensions. Nowadays, the link between BIM and GIS is an open area of research, and the opportunity to manage data and information in a separated environment can help in creating this link. This framework can be evaluated through a SWOT analysis (Table 5) to identify both the internal characteristics (Strengths and Weaknesses) and the external factors (opportunities and threats).

Table 5. SWOT analysis of the high level architecture

Strengths	**Weaknesses**
Scalability Dynamic organisational structure Knowledge oriented structure Possibility to define localised standards according to the multi-level paradigm Separation between the project level (dynamic and user friendly) and the knowledge based level (structured and articulated)	Need of complex technological components
Opportunities	**Threats**
Further development of open standards Further development of open data sources Increased culture about digitalisation in the domain	Policy limitations Privacy issues Limitation from commercial collaborative environments

4.2. Layers and environments description

Starting from the general structure presented in chapter 4.1, each area is analysed singularly with the goal of defining the framework for a multi-level digital environment.

4.2.1. Requirements

For a proper definition of both layers and environments, it is critical to identify the requirements that they must satisfy. The proposed analysis is based on the combination between literature review and unstructured interviews developed in the collaboration with Assimpredil ANCE. In the literature has been found several works dealing with the identification of the needs expressed by representatives of the construction sector about collaboration environments in different countries at world level. In the construction sector can be identified three main levels, i.e. industry, firm and project (Rezgui, Hopfe and Vorakulpipat, 2010; Sawhney *et al.*, 2014). Each stakeholder in each level shows different requirements. Following the analysis proposed by Sawhney *et al.* (2014), the industry level deals with regulatory requirements, open geo-data, land valuation and acquisition data, market trends, forum, knowledge exchange, trends and statistics. The firm level deals with knowledge management, procurement, ICT integration, and standard adoption. The project level deals with project tracking and monitoring, usage of project-level ICT applications, online meetings, alerts and reminders, modelling and drawing tools, real-time project management, onsite usage of applications, and monitoring. Furthermore, focusing on the design and construction phase, the project level embeds another level that is the discipline specific one. In the construction sector, this subdivision is not completely linear in terms of scale because in one project are involved several firms and one firm can deals with more than one discipline. Nevertheless, it can help in the definition of a schematic representation of the requirements.

From this analysis can be defined three main areas. An area that includes data integration and analytics, knowledge discovery, transfer and management where can be addressed the requirements defined at both the industry and the firm level. A second area that deals with the sphere of collaboration and information sharing between different firms and disciplines where can be addressed the requirements of both the firm (in the context of the project) and the project level. A third area dealing with the collaboration in single disciplines.

Starting from the literature, a critical reading highlights an overlap between requirements that can hardly coexist at the same level. Hence, all the revealed requirements has been reorganised according to the proposed scale (Table 6). Some requirements can be included in two or more classes. However, they can be declined according to the specific class gaining a different meaning according to the context of application.

To facilitate the interpretation of the table, the requirements are aligned on the horizontal axis. Thus, where the same requirement is defined for more than one class, it is repeated in line, otherwise the cell is left empty.

Table 6. Requirements reported from the literature review, reorganised according to a three level structure

Discipline and operation	Project and collaboration	KM and Analytics
Provide help and support facilities	Provide help and support facilities	Provide help and support facilities
Allow different users to customize their interfaces	Allow different users to customize their interfaces	Allow different users to customize their interfaces
Allow users to view and print models online	Allow users to view and print models online	
	Support a central repository for data storage	Support a central repository for data storage
Have a notification system to inform team members about new changes when their data is updated	Have a notification system to inform team members about new changes when their data is updated	

Record changes and transitions when they occur	Record changes and transitions when they occur	
	Allow multiple actors to share information through a common storage system	
Provide real time mechanism for sharing information	Provide real time mechanism for sharing information	
Have built-in communication tools	Have built-in communication tools	
	Have a mechanism for tracking information throughout the whole project	
	Be able to track data during the whole lifecycle of the construction project	
Have an administration user interface with full access rights	Have an administration user interface with full access rights	Have an administration user interface with full access rights
	Have a common environment data area/workspace for sharing and exchanging data	
Define who will produce BIM data what and when	Define who will produce BIM data what and when	
Inform people what to do and when to do it	Inform people what to do and when to do it	
Inform people about the information that they need to provide	Inform people about the information that they need to provide	
Assist with decision making	Assist with decision making	Assist with decision making
	Allow the client to be involved in the early stages of the design	
Define who has access to what and when	Define who has access to what and when	Define who has access to what and when
Inform each actor about her/his roles responsibilities and when they should perform them	Inform each actor about her/his roles responsibilities and when they should perform them	
	Define what the requirements are for each individual stage of the construction project	
	Define what needs to be provided at the end of each stage	
	Define external gates between each stage of the construction project	
Define internal gates among the same actors within the same disciplines	Define internal gates among the same actors within the same disciplines	
		Have a mechanism for preserving a project's information for future reusability with new projects
It should be accessible from anywhere at anytime	It should be accessible from anywhere at anytime	
It should have a plug-in for modelling software such as Autodesk Revit and Google Sketch-up	It should have a plug-in for modelling software such as Autodesk Revit and Google Sketch-up	
It should be hosted on online-shared storage with clear access rights for each actor	It should be hosted on online-shared storage with clear access rights for each actor	

It should give the option of allowing the actors to host their data on their local machines	It should give the option of allowing the actors to host their data on their local machines	
	It should be hosted on a scalable storage media because of the huge amount of information and bid size of models	It should be hosted on a scalable storage media because of the huge amount of information and bid size of models
It should have and provide backup facilities	It should have and provide backup facilities	It should have and provide backup facilities
It should be hosted by a reliable, dedicated, and known IT infrastructure or CSP	It should be hosted by a reliable, dedicated, and known IT infrastructure or CSP	It should be hosted by a reliable, dedicated, and known IT infrastructure or CSP
It should be easy for all team members to use	It should be easy for all team members to use	
It should have a simple user interface	It should have a simple user interface	
It should effectively improve coordination among team members	It should effectively improve coordination among team members	
	It should have clear definitions of actors, their roles, and responsibilities within multiple disciplines through the building's lifecycle	
	Platform development should be based on a standardized overall lifecycle data management policy	
	Platform development should be based on the existing BIM-related standards and protocols	
The platform should increase trust between people by recording changes that have been made (by whom and when)	The platform should increase trust between people by recording changes that have been made (by whom and when)	
It must have a process framework, i.e., process guidelines and protocols	It must have a process framework, i.e., process guidelines and protocols	
	GovernBIM platform must have a technological framework, i.e., BIM tools and API	GovernBIM platform must have a technological framework, i.e., BIM tools and API
	It has to have a legal framework	It has to have a legal framework
	It should clearly define the ownership of BIM documents	It should clearly define the ownership of BIM documents
	It should preserve intellectual property rights for each team member	It should preserve intellectual property rights for each team member
It should support different web browsers requirements	It should support different web browsers requirements	It should support different web browsers requirements
It should be able to support all types of transfer and collaboration tools	It should be able to support all types of transfer and collaboration tools	
It should enforce team members to use the same software version, as agreed upon at the beginning of the contract	It should enforce team members to use the same software version, as agreed upon at the beginning of the contract	
It should maintain a consistency of tools during the collaboration process	It should maintain a consistency of tools during the collaboration process	
It should provide access rights to the stored data based on actors' roles and responsibilities	It should provide access rights to the stored data based on actors' roles and responsibilities	It should provide access rights to the stored data based on actors' roles and responsibilities

It should provide actors with a secure login to the system	It should provide actors with a secure login to the system	It should provide actors with a secure login to the system
It should be hosted on physically secure data centers	It should be hosted on physically secure data centers	It should be hosted on physically secure data centers
It should provide security checks for uploaded/downloaded and transferred models	It should provide security checks for uploaded/downloaded and transferred models	It should provide security checks for uploaded/downloaded and transferred models
It should be affordable to both large companies and small to medium enterprises	It should be affordable to both large companies and small to medium enterprises	It should be affordable to both large companies and small to medium enterprises
The use of the GovernBIM platform should be time and cost-effective	The use of the GovernBIM platform should be time and cost-effective	The use of the GovernBIM platform should be time and cost-effective
There should be an intensive training program for practitioners regarding the GovernBIM platform	There should be an intensive training program for practitioners regarding the GovernBIM platform	
The users of the GovernBIM platform should be able to understand where and how their efforts are contributing toward the entire BIM model	The users of the GovernBIM platform should be able to understand where and how their efforts are contributing toward the entire BIM model	
It should define clear roles and responsibilities for each actor during the construction project	It should define clear roles and responsibilities for each actor during the construction project	
It should not take decisions away from people during the construction project lifecycle	It should not take decisions away from people during the construction project lifecycle	It should not take decisions away from people during the construction project lifecycle
It should not change what an actor does but support his work	It should not change what an actor does but support his work	It should not change what an actor does but support his work
It should provide a comprehensive element of consistency	It should provide a comprehensive element of consistency	It should provide a comprehensive element of consistency
	It has to provide a consistent structure for people during the building lifecycle	
	It needs to be connected to the construction professions and contractors	It needs to be connected to the construction professions and contractors
	Development of the GovernBIM platform should not only focus on Level 2 BIM, but should also go further to Level 3 BIM	Development of the GovernBIM platform should not only focus on Level 2 BIM, but should also go further to Level 3 BIM
The GovernBIM platform development process should take into account actors and data structures, which exists in the BIM execution plan (PEP, BxP, BEP, or BIMM) and responsibility matrix	The GovernBIM platform development process should take into account actors and data structures, which exists in the BIM execution plan (PEP, BxP, BEP, or BIMM) and responsibility matrix	
It should define what to govern in terms of people, information and documents, processes, classifications, and lifecycle	It should define what to govern in terms of people, information and documents, processes, classifications, and lifecycle	It should define what to govern in terms of people, information and documents, processes, classifications, and lifecycle
	It should also take into account all the people involved during a construction project and in particular, record all information received and delivered along the supply chain	It should also take into account all the people involved during a construction project and in particular, record all information received and delivered along the supply chain

		It may act as an intelligent expert system by making use of preserved data, giving it the ability to provide advice on new projects based on experience gained in previous projects
	Document management	
	Project workflow	
	Project directory	
	Central logs and revision control	
	Advanced searching	Advanced searching
Conferencing and white-boarding	Conferencing and white-boarding	
Online threaded discussion	Online threaded discussion	
Schedule and calendar	Schedule and calendar	
	Project camera	
	File conversion	File conversion
Printing service	Printing service	
Website customization	Website customization	Website customization
Offline access	Offline access	
Messaging outside the system	Messaging outside the system	
Wireless integration (mobile apps)	Wireless integration (mobile apps)	
	Archiving of project information	Archiving of project information
	Information service	Information service
	Financial service	Financial service
	E-bidding and procurement	E-bidding and procurement
BIM data generation and documentation	BIM data management and sharing	
	BIM data evaluation	BIM data evaluation
My agenda	My agenda	
My diary	My diary	
Personal task management	Personal task management	
Organizational structure	Organizational structure	
Project calendar	Project calendar	
Collaboration unit	Collaboration unit	
Project task	Project task	
	Schedule planning	
	Project monitoring	
	Group	
Project profile	Project profile	
Add/delete/freeze individual account	Add/delete/freeze individual account	
Organizational structure	Organizational structure	
work breakdown structure	work breakdown structure	
access permission grant	access permission grant	
BIM data import	BIM data import	BIM data import
	project document management	
project/sector access permission grant	project/sector access permission grant	
add/delete/freeze project/sector account	add/delete/freeze project/sector account	

	Data statistical and analysis report	Data statistical and analysis report
Model upload/download	Model upload/download	
	Multiple data model formats	
partial model exchange	partial model exchange	
versioning	versioning	
model merging	model merging	
data locking	data locking	
clash detection	clash detection	
conflict resolution	conflict resolution	
	audit trail	
	data publishing	data publishing
workflow management	workflow management	
model modifications		
2D data modelling		
data querying	data querying	data querying
	reference data linking	reference data linking
		product libraries support
model checking	model checking	
rule-based modelling		
model comparison	model comparison	
change management (notify changes)	change management (notify changes)	
remote model viewing	remote model viewing	
3D navigation	3D navigation	
Mark-up	Mark-up	
collaborative communication	collaborative communication	
Report generation	Report generation	
	FM data support	
Colour customization	Colour customization	
Workflow reporting	Workflow reporting	
Mobile computing support	Mobile computing support	
User profiling	User profiling	User profiling
access control	access control	access control
	data handling	data handling
interface customization	interface customization	interface customization
security	security	security
disaster protection	disaster protection	disaster protection
data archiving	data archiving	data archiving
Model change and versioning management		
ownership and responsibility	ownership and responsibility	
intellectual property	intellectual property	intellectual property
	model reliability	model reliability
model repository	model repository	
nomenclature system editor	nomenclature system editor	
	sub-models and objects with different levels of details	
	public and private model spaces	

	Globally Unique Identified (GUID) for all object data	Globally Unique Identified (GUID) for all object data
	Information Delivery Manuals (IDM)-based specifications	
Secure log-in with access rights	Secure log-in with access rights	Secure log-in with access rights
Hierarchical model administration structure	Hierarchical model administration structure	
	Download/upload model, and Check-in/check out/ check-out with lock	
	version lock and archiving	
model viewing options	model viewing options	
documentation and reports	documentation and reports	
Customizable interface	Customizable interface	Customizable interface
online real-time viewing, printing and markups	online real-time viewing, printing and markups	
On-click object property check and modification		
Design visualization and navigation	Design visualization and navigation	
Team communication and interaction	Team communication and interaction	
Features supporting confidentiality, integrity, and availability	Features supporting confidentiality, integrity, and availability	Features supporting confidentiality, integrity, and availability
System security	System security	System security
User authentication	User authentication	User authentication
Data security	Data security	Data security
Access control	Access control	Access control
	Encryption	Encryption
	Project scoping support	
	Software tool compatibility matrix	
	BIM scoping support	
System configuration manager	System configuration manager	System configuration manager
System configuration layout viewer	System configuration layout viewer	System configuration layout viewer
System status viewer	System status viewer	System status viewer
Help support and training	Help support and training	Help support and training
	Legal and contractual support	Legal and contractual support

Legend
(Alreshidi, Mourshed and Rezgui, 2016)
(Nitithamyong and Skibniewski, 2004) – PM system
(Oh *et al.*, 2015)
(Jiao *et al.*, 2013)
(Shafiq, Matthews and Lockley, 2013)
(Shafiq, Matthews and Lockley, 2012)
(Singh, Gu and Wang, 2011)

In addition to these explicit requirements, Redmond *et al.* (2012) presented a systematic interview to industry experts about the use of cloud computing as integration platform for BIM applications. The results reported a need of effective collaboration in the design and construction stage, while this need is not perceived in the other life cycle phases. Security and legality are of critical importance coherently to the results showed in Table 6. Furthermore, the use of a cloud solution is identified as an improvement for the advancement of BIM applications interoperability. The study highlighted the importance to distinguish the business process at the firm level and the one at the industry level. The definition of an industry standard business process is recognised but it should not influence the business process inside of a company. Finally, one respondent said *"the challenges are that there are so many different exchanges that can assist a new project and I completely support the need to document and identify these changes but I also do not want these standards to infringe on what I am presently implementing and deploying due to "plug and play" in the cloud environment",* underlining the separation between the working and the sharing activities.

Shafiq, Matthews and Lockley (2012) presented a review of existing collaboration platforms providing a summary of industry requirements. Results focused on the reliability of the model exchanged in a collaborative environment, identifying as critical the definition of rules in this environment. Model change and versioning management, ownership and responsibility, intellectual property and model reliability are listed as critical points for effective collaboration.

Many requirements identified for collaborative environments, are aligned with the ones listed for Knowledge Management systems. For example, Dave and Koskela (2009) proposed five main requirements as following listed:

a. Easy to use and to implement.
b. Efficient information retrieval.
c. Improvement in the level of trust between employees.
d. Implementation of Knowledge Management treated as a core system.
e. Implementation of Knowledge Management supported by the top management and integrated within the business strategy.

The first three requirements are aligned with those listed in Table 6. Requirements "d" and "e" show a different perspective, that is the emphasis on the role of the top management in the implementation of the system. This vision is supported by Anumba (2009) where are listed the rules for Next generation knowledge management systems.

- Demonstrable performance.
- Intelligent agent-based component.
- Artificial intelligence integration to capture and reuse knowledge.
- Improve personalisation of the user experience.
- Improve trust and privacy.
- Advanced mechanisms to support communities of practice, storytelling and anecdotes.
- Support the development and management of knowledge chains.
- Contribute to the market value of the sector.
- Be supported by and increased number of people.

The proposed analysis highlights the functional distinction of digital environments according to the phases of the long lifecycle of construction sector products. Hence, it is possible analyse each part of the main system structure according to the requirements structure proposed. It is worth stressing that the repetition of the same requirement at different levels does not imply its equal interpretation. In fact, the same requirements imply different technical, technological and processual approach according to the specific area of action modifying the vision on data including data structure and data quality perspectives.

4.2.2. Collaborative environment

Collaboration can be defined as *"the agreement among specialists to focus their abilities in a particular process to achieve the longer objectives of the project as a whole, as defined by a client"* (Hobbs, 1996). Collaboration involves the stakeholders by sharing both information and processes (Ilich, Becerik and Aultman, 2006). Hence, a collaboration environment such as a Common Data Environment (chapter 3.1.1) must allow the definition of shared and personalised processes as well as rules about information input and output. The final goal of an effective and efficient collaboration on a project is to enhance communication between all the stakeholders involved to reduce errors, omissions, and incoherence (Ilich, Becerik and Aultman, 2006). This vision about collaboration, clarify its focus on the project level. Hence, the first level of the framework must allow the management of the information generated during the project. Furthermore, it should include the rules defined for the specific project that could be different for any project and any stakeholder configuration (project team).

However, it is fundamental identifying a separation between the whole amount of data produced during the project and the information that needs to be shared. The design and construction process is iterative (Cooper *et al.*, 2005) and the final solution is reached through a dynamic evolution according to different factors that can affect the implementation of the project (Caldas, Soibelman and Han, 2002). Hence, the collaborative environment must contain the information approved by the specific design team and avoid sharing information on working stages. This interpretation is supported by the BSI PAS 1192-2:2013 (BSI, 2013) where CDE is defined as *"single source of information for any given project, used to collect, manage and disseminate all relevant* ***approved project documents*** *for multi-disciplinary teams in a managed process"*. Furthermore, the literature review highlights how users are not liking to undergo controls during work in progress; *"they do not want everybody to be analysing their changes in real-time, only when, and if, they have published it"* (Redmond *et al.*, 2012). The collaborative environment, devoted to the collaboration between multi-disciplinary teams, is consequently separated (at least logically) from possible collaborative environments identified for the collaboration on the same discipline and/or model. Figure 56 proposes a different visualisation of the first two areas of a CDE, namely work in progress and sharing. It highlights the distinction between the rules that can be introduced in the first area and the rules that must govern the second area and the link between the two. In fact, in the first area each team is working on its own specific part of the project and can adjust information exchange according to its own structure (e.g. tools, IT structure, and internal protocols) and to the requirements from the client (EIR, BEP, etc.). On the other hand, in the sharing area the rules are project based according to the identification of the better solutions to improve the collaboration between distributed and heterogeneous teams. Hence, both the sharing environment and the connection between this last at the working environments must comply with the requirements established for the project, e.g. EIR and BEP.

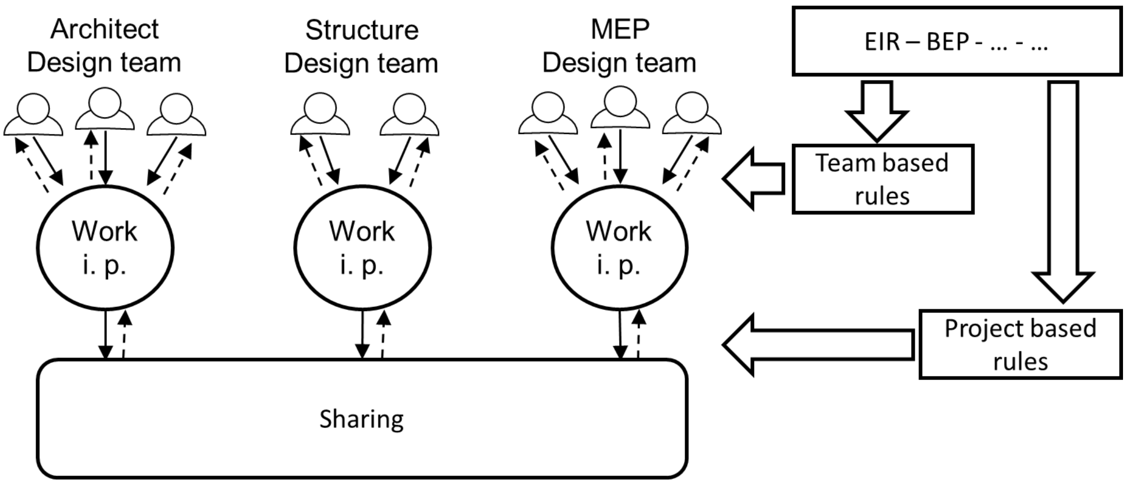


Figure 56. Logical representation of work in progress and sharing sections of the collaborative environment

The distinction between work in progress and sharing area defines a more dynamic vision about the CDE allowing single teams to work with their own tools and to exchange information according to their established procedure that can remain stable during different projects. This dynamicity is reduced moving in the sharing level where an equilibrium between all the stakeholders must be defined according to the temporary structure of the project. Hence, rules based at this level can change for any project, while the rules defined for the operative level (work in progress) may remain stable thanks to their team based (or company based) nature.

It is worth stressing the link between this schema and the theoretical framework presented in Figure 11 where it is clearly identified the inclusion of norms and rules from both the context (project) and the companies involved in the project. Moreover, the theoretical representation underlined how each team can interpret the shared project in different ways according to its own needs and knowledge background.

The framework proposed in Figure 56 can be completed introducing all the areas of a CDE as described in chapter 3.1.1. The graphical representation proposed in Figure 57 clarify the distinction between the two levels, according to the rules structure proposed in chapter 4.2.1.

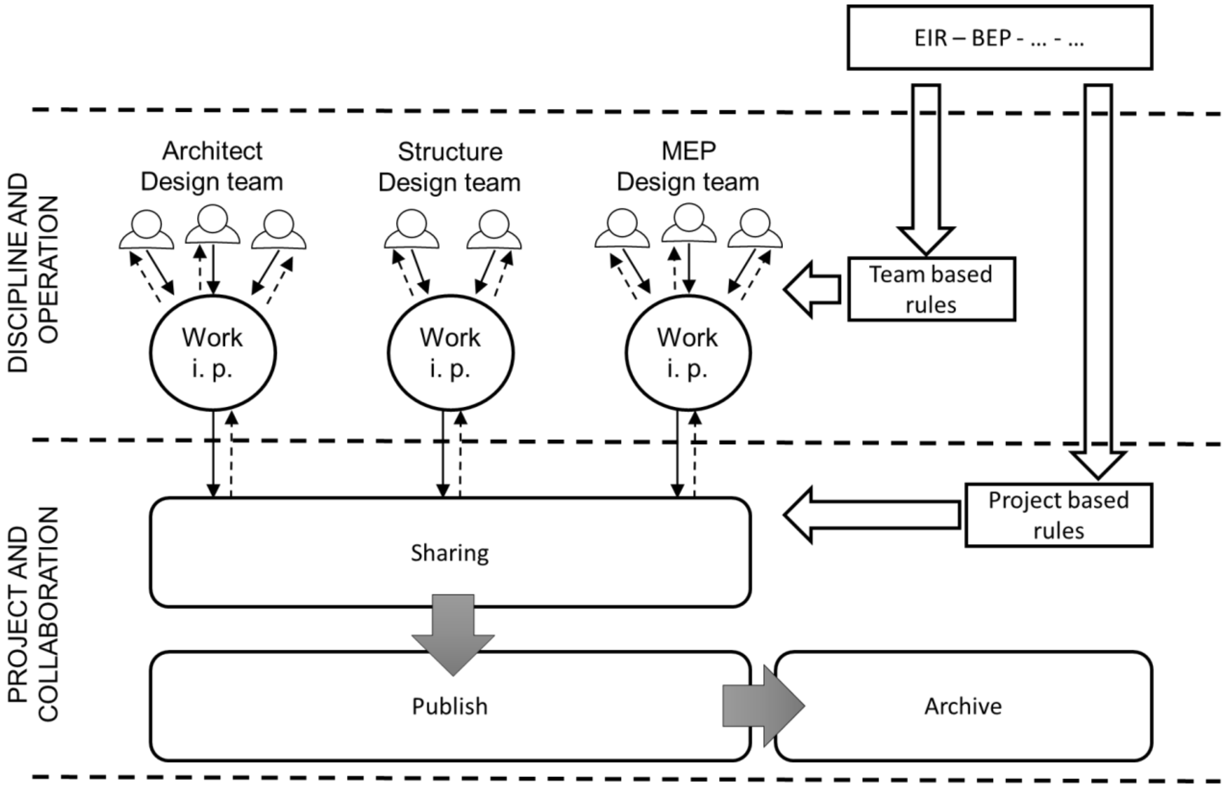


Figure 57. Identification of discipline and operation, and project and collaboration areas in the collaborative environment

4.2.3. Privacy layer

Privacy, reliability and access rights are fundamental requirements in the implementation of digital environments where data, information, and knowledge can be stored and used. Thus, the privacy layer deals with the guarantee of confidentiality of sensible data and information when they are transferred from one level to the next one. Practically the layer act on two points. On the one hand, sensible data and information are identified and stored in a separated environment accessible only from the proprietary of the data and information. On the other hand, all the data are merged following the platform based rules eliminating the possibility of tracing the original source of a specific data or information. Nevertheless, it is maintained the link between sensible information and the one analysed so that owners of data and information can retrieve their own dataset and analyse it separately with deeper analysis.

4.2.4. Layer of deconstruction

This layer deals with the data quality improvement from the sharing and collaboration environment to the platform. As proposed in chapter 3.3, a change in the function perspective of the environment can change the requirements in terms of data quality and data structure. The layer of deconstruction aims to parse the data received from the lower level to obtain a data set that is compliant with the rules and the data quality level established in the platform.

For example, as proposed in the following chapter 5 it deals with the asymmetric semantic representation of reality due to the limits of existing BIM authoring tools as well as to the check of the information focusing on its reuse in a machine-to-machine logic.

While models generated in BIM environments can be translated with data granularity, i.e. data are available singularly in a semantic environment, the extraction of granular data or information from other data sources like pdf documents is more complicated. Indeed, the problem is not embedded in the format, but more in the content of documents that is usually not structured and defined in natural language. In this direction, as already applied in different areas (chapter 2.2.2), it is possible including in the layer of deconstruction specific algorithms able to extract hided topics structure behind the documents and introduce classes and clusters according to this analysis (chapter 5.1).

4.3. Construction digital platform

Caldas, Soibelman and Han (2002) citing Teicholz (1999) stated that information integration has three dimensions, namely horizontal integration, vertical integration, and longitudinal integration. Horizontal integration deals with the integration of multiple disciplines in a construction project. This dimension can be addressed implementing the framework proposed in chapter 4.2.2 accordingly to the introduction of a CDE to manage construction project information.

Vertical integration is focused on multiple stages in the life cycle of a facility. The fragmented character of the construction process poses several problems in this dimension due to the change of the stakeholders during each stage and the change in the requirements that each stage can require. Hence, a CDE-like framework needs to be repeated in each phase identifying how to transmit information from one stage to the other and what information transmit. However, this approach bound the possibility to collect and use the data in a systematic way. In fact, the transmission of data from one specific environment to another impose a filter in data and information reporting the only information that are needed to the stakeholders involved in the receiving stage. This approach is coherent with a criteria of process optimisation (Bazjanac and Kiviniemi, 2007). However, it reduces the possibility of data inclusion and consequently the value that can be extracted from the data (see chapter 3.4).

Finally, longitudinal integration focuses on the integration of data over time and it is usually related to the concept of knowledge capture allowing improved performances and/or better decisions in the future. The use of a traditional framework for the identification of a CDE is not suitable in this case too.

To admit information integration including all the above-mentioned dimensions is proposed a novel framework that integrates the first one proposed in chapter 4.2.2 accordingly to the main vision proposed in chapter 4.1 and the requirements structure defined in chapter 4.2.1. Figure 58 displays a graphical representation of the proposed framework.

It is worth underlining that this framework is not defined according to a specific technology and/or data structure. Even if chapter 4 refers extensively to the use of semantic web technologies and to the existing studies that demonstrated the possibility to translate building information models in data models that can be shared and linked on the web, there are other research that are analysing different "translation" procedures. Moreover, according to the evolution of both digital infrastructures and digital technologies on the data management side, the study of how to generate structured data models from building information models is an open area for future research. Hence, the framework and the related description do not report a defined analytical representation about data structure because it can be applied using different technological solutions maintaining valid the principle, i.e. the identification of a multi-level structure according to the objectives of the specific area of action.

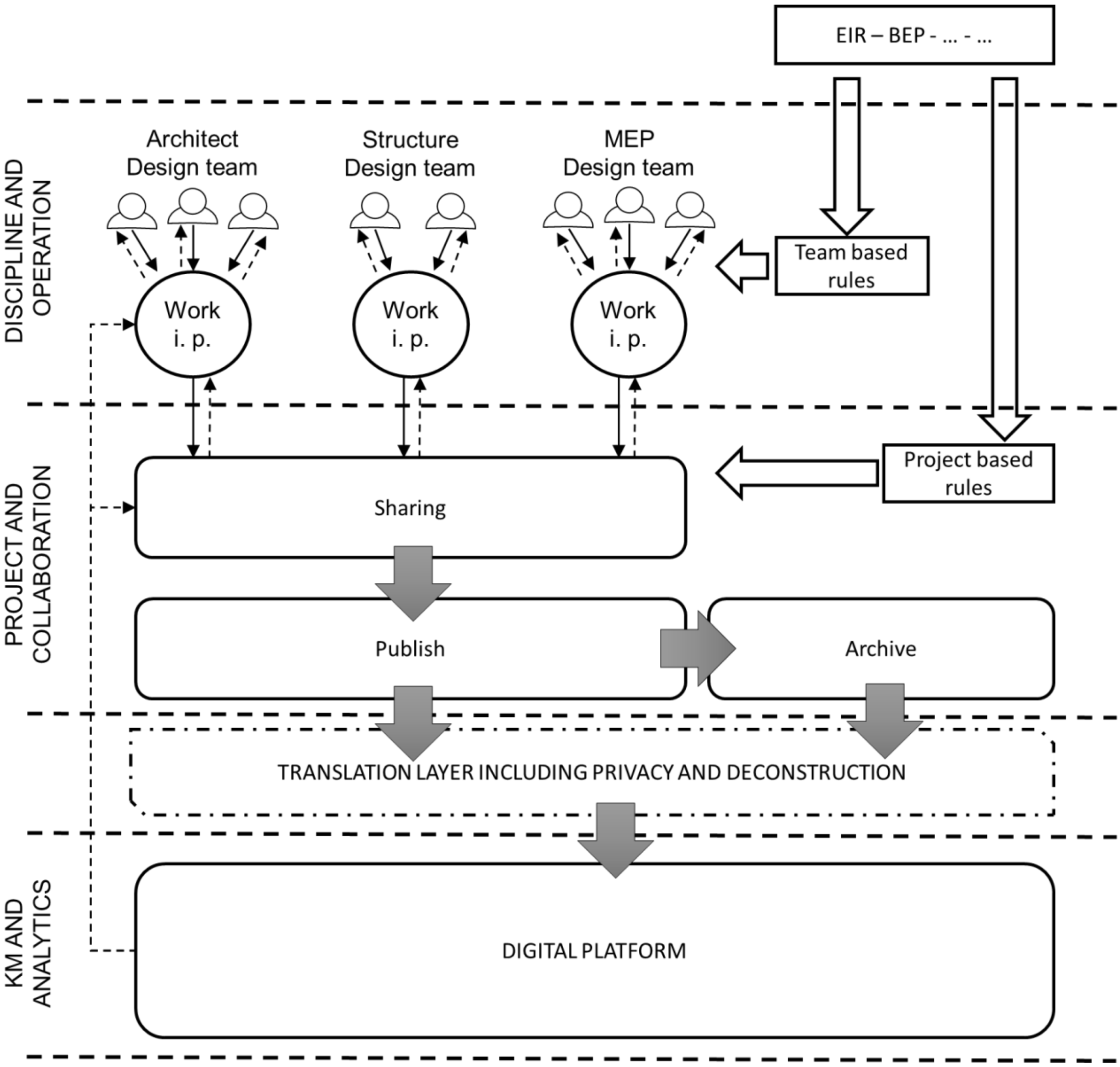


Figure 58. General framework of the system

The first two levels, namely discipline, and operation and project and collaboration remain the same of the first framework. A new level is introduces starting from the last storing area of the project and collaboration level bringing to the Knowledge management and analytics level. Between every spaces identified in the first two level exist defined rules for the transmission of data that are defined at the project level. In the passage between the archive area and the digital platform a structured translation layer is introduced addressing the requirements presented in chapters 4.2.2 and 4.2.3. Hence, the digital platform represents a digital environment where data can be collected and structured according to the needs of the application of data analytics in a period of time that do not need to satisfy the dynamicity requested in the development of projects. Furthermore, the digital platform can structure the data overcoming the limits of the instruments and/or of the format used in the project and collaboration level because its focus is on the use of the data instead of on the exchange of it.

The analysis that can be performed in the digital platform can be used to provide information and services to the lower level acting on both the working in progress area and the sharing area. In fact, following a cloud based logic, data analysis can be executed in a separated environment (the digital platform) providing services in a "as a service" logic to the individuals operating on the lower levels. The inclusion of a quantity of data and information that is not possible using a traditional CDE structure allows the effective application of KDD processes (see chapter 2.2.1) empowering the production of new knowledge starting from the collected data.

4.3.1. Dynamic interpretation

Figure 58 highlights a clear separation between the project and collaboration level and the KM and analytics one. This decision is based on several evidences. The introduction of changes in the data, both in terms of structure and in terms of derived (or inferred) data can open issues in the responsibility associated to the production of the data. During the development of a project, each stakeholder has specific responsibilities and must maintain the control on data and information of his or her competency. The time required for the execution of all the passages included in the translation layer can exceed the time requirements at the project level hindering the effective activation of collaborative processes. The variability of rules and tools that can be registered at the project level can be incompatible to the identification of a general structure that can effectively operate with every tools applying different rules. Finally, the complexity behind the definition and use of an integrated structure that can satisfy all the requirements associated to the different level can slow the process limiting the introduction of new systems. This separated approach is now under discussion in the definition of an evolutionary view in the ontological structure for the construction industry. A three level approach with an increasing semantic richness defined in accordance to the increase in the level is proposed (Bonduel, 2018).

	Level 1	Level 2	Level 3
Query execution time	+++	++	+
Easy/intuitive to discover props	+++	++	+
Reasoning	?	?	?
Alignment to schema.org	?	?	?
Ontology is easy to maintain and extend	?	?	?
Extra information about property	/	✓	✓
Grouping of props	/	✓	✓
Versioning of props	/	/	✓
Units for props	In rdfs:description (string)	QUDT, OM, other	QUDT, OM, other
Literals: data typing + language tags	✓	✓	✓
Complex props	/	✓ (?)	✓ (?)

Figure 59. Different level of complexity in for construction ontology structures – originally taken from (Bonduel, 2018)

However, looking at the evolution that the linked building data is providing, it is possible to identify a dynamic structure where the proposed algorithm are introduced at the collaboration phase separating data and information produced by the stakeholders and data inferred or defined by the machine and focused on the analytics level. Hence, data and information will be included in a unique environment (practically a set of distributed environments linked to each other) in a dynamic structure that can change according to the progressive inclusion of data and information by the machine. However, the definition of an ontological structure able to satisfy all the visions that can be defined in the construction sector is still an open issue. Moreover, this structure would change according to the evolution of technologies and techniques in the domain complicating its development. Nevertheless, other studies are working on the use of different data structures e.g. Ismail and Scherer (2018).

5. Application and experimentation

"It doesn't matter how beautiful your theory is, it doesn't matter how smart you are. If it doesn't agree with experiment, it's wrong."
(Richard Feynman)

The framework presented in chapter 4.3 sees in the "translation layer" a critical component to obtain usable data and/or information in the KM and analytics level. The translation layer must deal with data and information that can be embedded in several file types (e.g. .doc, .dwg, .pdf, .ifc, .rvt). The study and development of systems able to deal with all the file formats that can be generated during a construction process is out of the scope of this thesis. However, it is critical to understand how the system can work with different file types and explore the ways that the proposed approach can open. This chapter presents two possible applications analysing on the one hand, how to automatically organise documents in open formats (.pdf) to optimise the use of the embedded information. On the other hand, combining the proposed framework with the issues explored in chapter 3.3, the chapter explores the use of machine learning techniques to limit data quality issues in building information models, checking and enriching semantically these last.

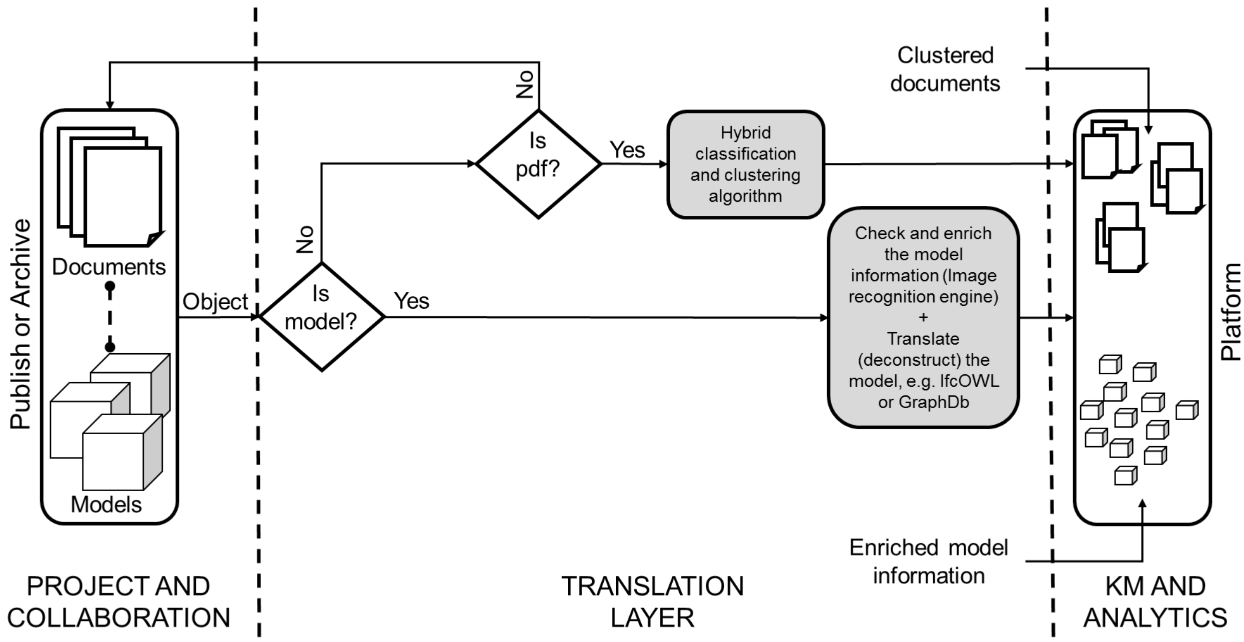


Figure 60. Integration of the applicative case studies in the platform framework

Figure 60 shows the integration between the two applications developed and the platform framework. Starting from the content of the "publish" and "archive" environments at the project and collaboration level, the system enter in the translation layer. This layer represents one of the key components of the framework and can be articulated with several applications and functions. According to the case studies explored, it can include a specific application to organise .pdf documents (chapter 5.1) and a specific application to check and semantically enrich building information models (chapter 5.2) that can be combined with existing application to translate (deconstruct) the model information (e.g. IfcOWL, graph databases, etc.). Hence, the following case studies represent two components of the translation layer exploring the possibilities to deal with documents and building information models. Future research can look at the study and development of other components that can be included in the layer to allow the use of other file types and to allow a better exploration of building information models and documents information content.

5.1. Document management through hybrid classification and clustering approach

In the construction industry, especially during the design and construction processes, it is produced a large spectrum of information. The majority of the information is commonly embedded in text data contained in documents delivered during the process (Caldas and Soibelman, 2003). Even a small project can generate a large amount of digital information such as specifications, computer-aided drawings, and structural analysis reports (Soibelman *et al.*, 2008; Tixier *et al.*, 2016; Zou, Kiviniemi and Jones, 2017). Hence, a large percentage of the construction data is stored on semi-structured and unstructured files (Soibelman and Caldas, 2000).

Unstructured texts remain the largest readily available source of knowledge (Gupta and Lehal, 2009). Hence, text mining is gaining interest in both research and industry as a valuable multidisciplinary area. Focusing on text documents, one of the main activities to promote the accessibility of information is the organisation of the documents to facilitate the individuation of the required information.

In the automatic organisation of documents can be defined two main approaches, i.e. classification and clustering. Classification aims to classify documents into pre-defined classes (Liu, 2006; Manning, Raghavan and Schütze, 2008), while clustering can be defined as the unsupervised classification of patterns into groups called clusters (Jain, Murty and Flynn, 1999).

In the literature can be found several studies devoted to the application of natural language processing (NLP) to organise and use text documents in the construction sector (chapter 5.1.1.3). The majority of these studies is focused on homogeneous data sets limiting the scope of the research to the specific areas of study. However, real world datasets - generated combining different project environments as defined in the proposed framework - are characterised by a high diversity of documents in terms of contents and standards used in their development. Moreover, the background structure that characterise a set of documents can change during the years due to the introduction of new standards, new legal requirements or according to the development of the techniques used in the sector. Hence, it is not possible to know in advance the evolution of this structure. This point highlights one of the main limits in the use of classification algorithms. On the other hand, the use of clustering algorithms is usually more expensive in terms of times and computational resources.

The proposed approach explores the use of a hybrid document organisation process mixing classification and clustering algorithms to prepare an organised and hierarchical documents structure that can be used for the development of specific and dedicated applications.

It is worth mention that this study is focused on text documents in .pdf format. The organisation of other documents that can differ in the content (e.g. drawings or images) or in the format (e.g. .doc, .dwg or .dxf) can be identified as a future line of research.

5.1.1. Background

5.1.1.1. Classification and clustering of text data

Classification and clustering are two main functions of text mining processes (Gupta and Lehal, 2009). Text mining can be defined as the process of synthesising the information by analysing relations, patterns and rules from textual data (Berry Michael, 2004). It is different from what is commonly applied as search in text data. In fact, searching engine are focused on finding something already known. In text mining, the goal is to discover unknown information (Niharika, Latha and Lavanya, 2012).

Classification (or categorisation) is based on the identification of the main themes of a document by placing the document into a pre-defined set of topics (categories) (Sebastiani, 2002; Gupta and Lehal, 2009). Hence, the goal is to classify a set of documents in a pre-defined and fixed number of categories. In some applications, the same document can be included in more than one category, using e.g. multi label (Boutell *et al.*, 2004; Madjarov *et al.*, 2012) or fuzzy classifiers. There are different methods of text classification, namely decision trees, k-nearest neighbour, Bayesian approaches, neural networks, regression-based methods and vector-based methods (Niharika, Latha and Lavanya, 2012). Moreover, classification algorithms can be developed according to different constraints. Even if there are several applications and methods, in text classification can be highlighted two main constant terms. These are 1) the need to handle and organise large quantities of documents in which the textual component is either unique or simple to interpret and 2) the fact that the set of categories is known in advance, and its variation over time is small (Niharika, Latha and Lavanya, 2012).

Clustering is a technique that can automatically organises a dataset containing a substantial number of data objects into a smaller set of coherent groups (Willett, 1988; Jain, Murty and Flynn, 1999; Huang, 2008). It can be used to group similar documents (Iiritano and Ruffolo, 2001). Clustering differs from classification because the documents are clustered dynamically instead of using pre-defined topics and categories (Gupta and Lehal, 2009). One of the main objective of clustering is to provide a structure to a large dataset by organising similar data together to facilitate search and retrieval tasks (Al Qady and Kandil, 2014). Clustering techniques are used in several disciplines including biology, psychiatry, psychology, archaeology, geology, geography, and marketing (Jain and Dubes, 1988) with specific applications related to pattern recognition (Anderberg, 1973), image processing (Jain and Flynn, 1996) and information retrieval (Salton, 1991; Rasmussen, 1992).

The majority of clustering algorithms are based on similarity measures (Gupta and Lehal, 2009). Nevertheless, clustering methods can be divided in two main classes, i.e. hierarchical clustering and flat clustering (Frakes and Baeza-Yates, 1992).

Hierarchical clustering can be further divided in two classes. Agglomerative, i.e. a bottom up approach where each element starts as a single cluster and pairs of clusters are merged as one moves up in the hierarchy and divisive, i.e. a top down approach where all the elements start in one cluster and are then split recursively as one moves down the hierarchy (Al Qady and Kandil, 2014). The final result of a hierarchical cluster is a tree with all the elements on one extreme of the tree and a set of clusters each one containing a single element on the other side. The intervening nodes contain several elements and can contain more or less documents moving up or down the hierarchy (or the tree) (Gupta and Lehal, 2009). Hierarchical clustering is often identified as the best quality clustering approach, however its application can be limited due to its quadratic time complexity (Steinbach, Karypis and Kumar, 2000).

Flat clustering techniques define a single set of clusters on the entire dataset. Usually, clustering algorithms require the identification of the desired number of clusters that can be defined through the introduction of evaluations methods such as coherence scores.

In both classification and clustering processes, text documents need to be pre-processed cleaning the text and representing the documents in a measurable form that can be used to evaluate the similarity between documents (feature weighting or language model). There are several schemes to document representation such as binary, term frequency (tf), inverse document frequency (idf), and term frequency-inverse document frequency (tf-idf) (Mirończuk and Protasiewicz, 2018).

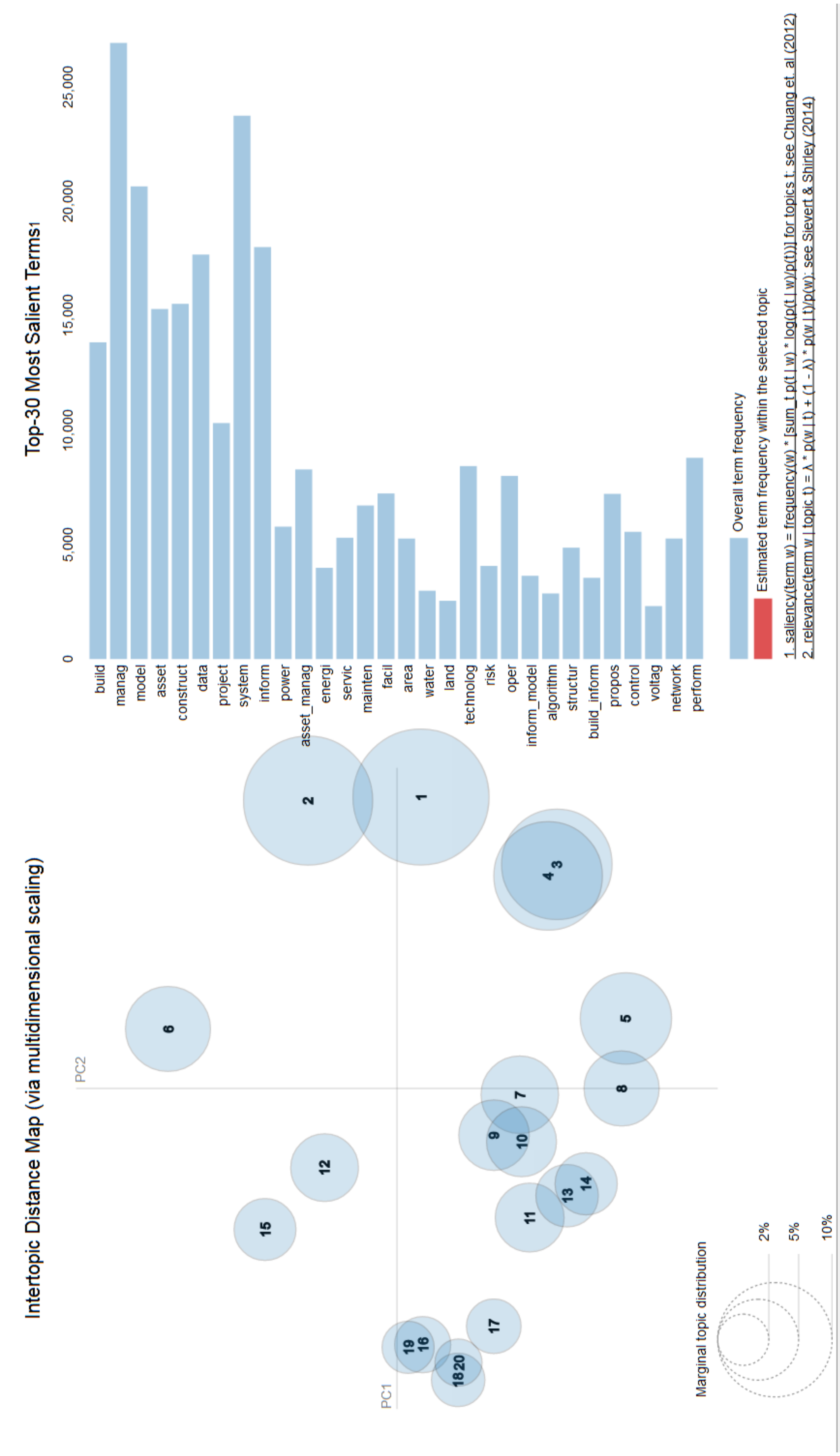


Figure 61. Example of a flat clustering representation obtained using Latent Dirichlet Allocation (Blei and Lafferty, 2009) applied on a corpus of 22900 research articles

5.1.1.2. Natural language processing (NLP)

The application of classification and/or clustering algorithms requires the analysis of textual data through the application of NLP techniques. In fact, the application of the above mentioned schemes to document representation can produce poor results if applied directly on a text document. For example, the term frequency scheme represents each word as a feature of the document that is represented as a vector with non-negative value for each feature (Huang, 2008). The weight of each features is calculated according to the frequency of the related terms in the document. However, in natural language several words appear frequently in the text but are not relevant for its evaluation. This is the case, e.g. of stop words such as "a", "the", "and", and "so". On the other hand, several terms can appear in different forms even if they refer to the same meaning. This is the case, e.g. of verbs. Hence, in both classification and clustering processes can be identified several steps to prepare the text documents through NLP techniques:

- **Tokenisation:** this is a process of chopping a document up into pieces (known as 'tokens'), usually correspondent to each word. It may include the removing of punctuation and of special characters.
- **Converting words in lowercase:** this is a process that converts all the words into lowercases to improve the identification of similar words.
- **Lemmatisation:** it consists in removing words endings returning the base or dictionary form of a words, which is known as the lemma (Manning, Raghavan and Schütze, 2008).
- **Stemming:** this is an alternative process to lemmatisation and it consists in removing words suffix to generate word stems (Gupta and Lehal, 2009). It differs from lemmatisation because it does not require an extended vocabulary.
- **Removing stop words:** stop words are extremely common words that have little impact on the meaning (thus in the evaluation) of a document. There are several lists of stop words available for different languages that can be used to effectively clean a text document.
- **Filtering:** filtering processes are not standard and can vary according to the specific document set. Can be used ad hoc dictionaries developed according to the area of study to remove non-common words or generic filters used to remove e.g. words with less than 3 digits or with more than 10 digits identified as strange in a text.

5.1.1.3. NLP, classification and clustering in the construction sector

During the construction process, information can be generated in several different forms and embedded in a variety of different documents. The increasing volume of digital information imposes the exploration of means to automatically manage and organise all this information to improve its accessibility and to allow its effective use. Classification and clustering represent two fundamental techniques related to these aspects that are not limited to text documents.

In the literature can be found several studies based on the use of classification and clustering for data management in the construction sector. Saitta *et al.* (2008) developed a clustering application based on K-means to narrow down the number of candidate structural models and improve the performance in identifying the one that best represents the measurements of in site sensors. Cheng and Teizer (2012) applied clustering to point cloud data to improve the visibility of tower crane operators. Ng, Toukourou and Soibelman (2006) proposed a clustering application to group facility condition assessment reports into similar clusters with the aim of understand possible relationship between reported deficiencies and facility types. Raz *et al.* (2004) explored the use of clustering to develop truck weight-in-motion traffic data with the aim of identify data anomalies. Moreover, clustering techniques were applied in image processing to defect detection (Lee and Chang, 2005; Guo, Soibelman and Garrett, 2009) and to manage construction site images (Brilakis, Soibelman and Shinagwa, 2005).

Focusing on text documents, Yu and Hsu (2013) developed a content-based CAD document retrieval system to manage CAD drawings according to their text content. Wood (2000) proposed a method to extract concepts from design documentation. Tixier *et al.* (2016) proposed a prototype to extract precursors and outcomes from injury reports. Caldas and Soibelman (Caldas, Soibelman and Han, 2002; Caldas and Soibelman, 2003) proposed an automatic hierarchical classification of construction project documents to improve information organisation and access. Al Qady and Kandil (2014) presented a method to automatic clustering construction project documents based on textual similarity. Zou, Kiviniemi and Jones (2017) developed a system to retrieve similar cases related to injury reports to optimise project risk management. Garrett, Fenves and Stasiak (1996) explored the use of NLP to classify regulation sections. Brüggemann, Holz and Molenthin (2000) presented a document markup methodology based on arbitrarily structured metadata. Scherer and Reul (2000) explored the use of clustering techniques to retrieve project knowledge from heterogeneous AEC/FM documents.

5.1.1.4. Motivation and aims of the experimentation

The literature review highlights the variety of applications developed in the research field in the area of NLP, classification and clustering in the construction sector. However, there are few studies devoted to organising and analysing heterogeneous documents. Moreover, the majority of these studies are based on clustering techniques using a flat, hard clustering approach (Al Qady and Kandil, 2014). Caldas and Soibelman (2003) argued that the use of a flat and exclusive clustering procedure can limit the approach due to the complexity of the documents generated in the construction sector that can contain information related to different arguments (clusters and/or classes). Other studies are based on classification techniques that require the inclusion of a specific knowledge into the algorithm according to the document set. Moreover, classification can admit only small variations of the dataset composition because changing the imposed classification requires the complete redefinition of the process. The literature review underlines the development of specific applications, e.g. to improve documents retrieval or to extract specific information from documents. However, these studies are usually based on a single source of documents (single database) usually homogeneous with reference to the requirements of the algorithm of classification and/or document research. Unfortunately, real datasets are composed by a mix of not homogeneous documents and the application of solutions based on homogenous documents set can produce a drastic loss in the efficiency and reliability of the algorithm. Hence, the proposed approach tries to solve this problem identifying a hybrid classification and clustering process to pave the way for the development of specific applications that can select the specific set of documents. Starting from the framework presented in chapter 4.3 - Figure 58 - this experimentation would allow the background organisation of all the documents collected from the lower levels of the platform and the consequent development of applications based on the organised information treated by the higher level of the platform.

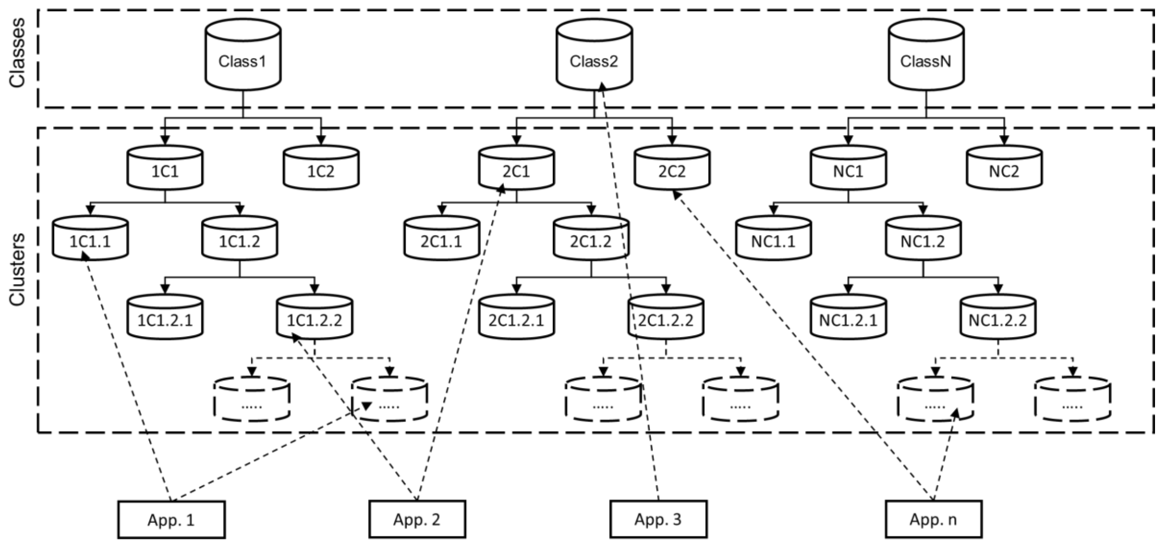


Figure 62. Logical representation of the proposed approach

This approach is well known in the field of data mining where it is identified with the name of *segmentation* (Jain, Murty and Flynn, 1999). In large databases, clustering methods can be used to segment the database in homogeneous groups so that other applications can work on clusters instead of working with the entire database improving the performances of the final applications. In the case of construction documents, the introduction of a previous phase of classification is fundamental to optimise the performance of the system and the results obtained with the clustering process. In fact, previous experimentations developed by the author during this case study demonstrated poor performance of clustering algorithms applied directly on the entire document sets. In document clustering applications, the inclusion of information referred to different disciplines and/or areas of interest can produce poor clustering results in the case of exclusive clustering processes (i.e. clustering processes where documents are assigned only to a specific cluster). Moreover, the time of execution is greatly improved limiting one of the critical issues of hierarchical clustering that is the required high efforts of computable resources. Nevertheless, it is worth mention that the time of execution does not represent a critical point in the proposed approach thanks to the platform framework. In fact, the classification and clustering applications will be embedded in the higher level of the platform that does not need to respect the velocity imposed by the construction process that is addressed in the lower levels.

5.1.2. Framework, methodology and system development

The framework of this experimentation can be divided in two main areas, namely classification and clustering. The classification part, as the first step of the procedure, comprehends all the activities required to prepare the text documents. The first step in the classification activity is the definition of the classes that will be used to classify the documents. Following the logical structure of the proposed approach, the classes will be defined according to general principles that can be identified as "stable" in the construction sector. This point is critical to guarantee the required dynamicity of the system and to identify classes that can contain a sufficient number of documents for the future clustering activity.

The entire process was translated into a prototype developed using Python programming language. Python is an object-oriented programming language and one of the most used with a continuous development of additional features. The main reasons to use Python in the development of the prototype are:

- Fast in development and test activities: Python can be executed directly without the need of compiling an executable format.
- Dedicated packages for the development pre-processing, classification and clustering activities (e.g. the Scikit-learn package (Scikit-learn, 2018), or the Treetagger module (Otto, 2018)).
- Easy reading of the code for its dissemination in future research works.

5.1.2.1. Definition of the classes

To determine the possible definition of classes a first analysis of the dataset and of the context where the dataset has been developed and for whom and for what it could be used in the future is required. The dataset used in the experimentation is composed by project documents derived from public tendering and offering documents from construction companies that participated to the tendering procedure. The documents have been collected in the Italian context, hence a first understanding of the documents that can be found is provided by the Italian law for public contracts, Decree of the President of the Republic October 5, 2010, n. 207 (Stato Italiano, 2010), that establishes the minimum documents for tending procedures during their different stages. In synthesis, the main documents are following listed:

- General report of the work.
- Technical reports and specialist reports.

- City planning documents for the work.
- Drawings.
- Environmental impact study.
- Calculations about structures and plants.
- Description and performances of technical elements.
- Design of interferences.
- Expropriations.
- Price description and list.
- Estimative metric computation.
- General economic framework.
- Contractual schema.
- Tender specifications.
- Safety and security plan.

In the literature, Caldas, Soibelman and Han (2002) proposed 13 classes during the classification of minutes related to a construction project. The classes were defined studying the structure of the document set. The classes were general, schedule, Demolition-civil, landscape-site, structures, building_skin, roofing_waterproof, interior finishes, conveyance, plumbing, fire protection, HVAC, electrical. These two classifications are too detailed with reference to the scope of the classification process identified in the proposed approach.

Analysing the organisational structure of some engineering companies, it is possible to define a more general subdivision where the above mentioned documents can be included. Engineering and construction companies are usually based on a project oriented organisation (Shirazi, B., Langford, D. A., & Rowlinson, 1996; Ilin *et al.*, 2016). Moreover, according to the dimension of the company can be identified several departments related to specific areas of action, e.g. infrastructures, buildings, and energy. The project-oriented nature is intersected with the main sectors of the organisation creating a matrix structure. Due to the high variability of this last organisational structure, it is difficult to define a precise and clear subdivision. Nevertheless, it is possible to identify some general areas as following listed:

- **Administration**: this department can control all the bureaucracy related to the project. In some cases, the same department manage the quality, while in other organisations there is a dedicated department according to the dimension of the company.
- **Security and Safety**: this department includes all the aspects of security and safety.
- **Design**: this department collects the aspects related to architectural design and technology (in this last area, it shares many components with the engineering one).
- **Engineering**: this department includes all the engineering aspects, e.g. structures, HVAC, electrical, and fire safety.
- **Economic and management**: this department deals with the economic and scheduling themes.

Usually, in engineering and construction organisations, the engineering department does not exist as a whole but it is divided according to the specialisation (e.g. structure, HVAC, electrical, plumbing, fire protection). This is also in accordance with the classification proposed by , Caldas, Soibelman and Han (2002). However, the a priori identification of these specific areas can limit the algorithm imposing specific disciplines, while the evolution of the construction sector can introduce new specialisation areas (e.g. internet connections related to the Internet of Things or 4G and 5G installations for buildings). Hence, this subdivision can be managed through the clustering algorithm that allows the evolution of the cluster's identification according to the evolutions of the dataset. Following this evaluation, the classification algorithm will be developed including 5 main classes, namely administration, security&safety, design, engineering, and economic&management.

5.1.2.2. Classification process

Starting from the defined classes, the first step in the classification process is the definition of the training and test documents set. Figure 63 shows the process applied for this activity.

The documents, collected from different projects, are manually divided in the five main classes filtering the documents that cannot be used in the analysis (e.g. scanned documents) and inserting each document in the specific belonging class. The majority of the documents can be assigned to a specific class thanks to the generality of the defined classes. However, some documents contain so many arguments (topics) that cannot be identified in one specific class. Hence, was introduced a new class named "Others" that collect all the documents where a specific class cannot be assigned.

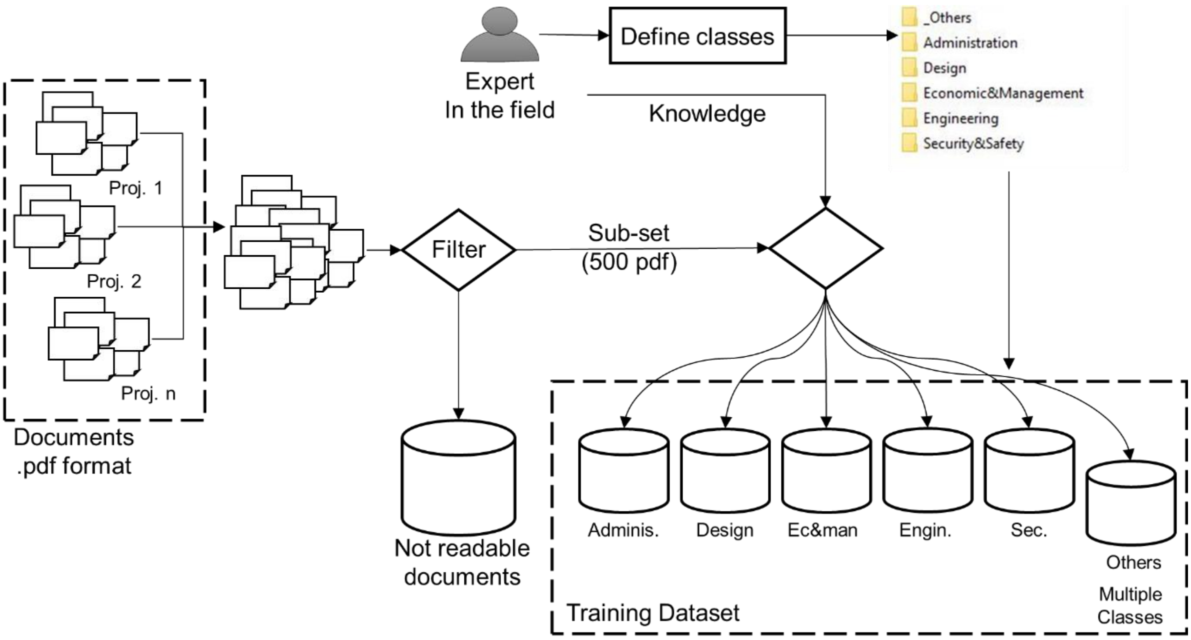


Figure 63. Preparation of the training and testing document set

To guarantee the repeatability of the experimentation, the manual classification of the documents must follow some general rules.

- **Administration:** all the documents related to forms, tendering instructions, and contracts are assigned to this class.
- **Design:** all the documents related to the description of the work, the performance specifications of the work (that do not refer to a specific engineering discipline), the urban and architectural consideration are assigned to this class.
- **Economic&management:** all the documents that have as objective the economic evaluation of the good, such as price lists, quantity take off, estimative metric calculation, as well as all the documents related to the management of the project such as schedules, Gantt charts, etc. are assigned to this class.
- **Engineering:** all the documents dedicated to engineering arguments such as structural, HVAC, electrical, geotechnical, thermal, etc. are assigned to this class.
- **Security&safety:** all the documents containing the specifications related to the security and safety for the project such as the plan for the security and coordination during the construction phase, the evaluation of the risks, etc. are assigned to this class.

Defined the training dataset the classification algorithm can be trained and tested. As explained in chapter 5.1.1.2, the text data needs to be prepared before the classification algorithm can be applied. Hence, a pre-processing phase is embedded in the classification process. First, text documents need to be transformed into a format ready to be read by the algorithm, thus all the document files were converted from Portable Document format (PDF) to text file format (keeping the original information into the document). Then, tokenisation, converting words in lowercase, lemmatisation, and removing stop words can be applied to obtain a file with one word (token) per character that can be used in the transformation and evaluation processes. To remove stop words it is required a dictionary of standard stop words that for Italian language can be found at https://github.com/stopwrods-iso/stopwords-it. Figure 64 and Figure 65 report the code developed for the text data pre-processing phase.

```
# set up Treetagger
tagger = treetaggerwrapper.TreeTagger(TAGLANG='it',
                                      TAGPARFILE="tree-tagger-MacOSX-3.2/italian-par-linux-3.2-utf8.bin",
                                      TAGDIR="tree-tagger-MacOSX-3.2")

# read stopwords list
with open("stopwords-it.txt") as f:
    stopwords = [line.strip() for line in f.readlines()]

def preprocess(text):
    try:
        tags = tagger.tag_text(text)
    except:
        tags = []

    # preprocessing: filtering
    # tokenization
    tokens = filter(lambda t: t.strip()!='', [tags2dict(tag)['word'] for tag in tags])
    # filter out single characters
    tokens = filter(lambda t: len(t) > 1, tokens)
    # filter out stopwords
    tokens = filter(lambda t: not t in stopwords, tokens)

    return tokens
```

Figure 64. Text data pre-process using Treetagger (Otto, 2018)

```
def tags2dict(tags):
        """Converts the output of Treetagger into a Python structure.
        Input: a string with three tab-separated fields for word, POS-tag and lemma.
        Output: a dictionary {word, POS-tag, lemma}.
        """

        # when Treetagger finds a URL, it does not tag it and insteac returns a "repdns" tag
        if tags.startswith('<repurl'):
                word = tags.replace("<repurl text=\"", "").replace("\" />", "")
                # we give it a costom POS-tag: "URL"
                return {'word': word,
                        'pos': "URL",
                    'lemma': word}
        if tags.startswith('<repdns'):
                word = tags.replace("<repdns text=\"", "").replace("\" />", "")
                # we give it a costom POS-tag: "URL"
                return {'word': word,
                        'pos': "URL",
                    'lemma': word}
        if tags.startswith('<repip'):
                word = tags.replace("<repip text=\"", "").replace("\" />", "")
                # we give it a costom POS-tag: "URL"
                return {'word': word,
                        'pos': "IP",
                    'lemma': word}
        if tags.startswith('<repemail'):
                word = tags.replace("<repemail text=\"", "").replace("\" />", "")
                # we give it a costom POS-tag: "URL"
                return {'word': word,
                        'pos': "EMAIL",
                    'lemma': word}
        if tags.startswith('<'):
                word = tags
                # we give it a costom POS-tag: "URL"
                return {'word': word,
                        'pos': "TAG",
                    'lemma': word}

        try:
                word, pos, lemma = tags.split('\t')
        except:
                log.error("Treetagger returned an error tag: {0}".format(tags))

        return {'word': word,
                'pos': pos,
                'lemma': lemma}
```

Figure 65. Treetagger configuration

The pre-processed documents need to be transformed and analysed with a defined metric and a defined algorithm. According to the literature, Support Vector Machine (SVM) is the best solution in text categorisation (Joachims, 1998). Moreover, the tf-idf feature weighting guarantees a simple application with a good ability to handle the identification of relevant terms in a large dataset. The tf-idf is a measure of relative frequency of words.

For a given word, its term frequency (tf) is the number of times that word occurs in the dataset, and its document frequency is the number of documents where that word occurs at least once. The mathematical expression for features based on tf-idf function is reported below:

$$w_i = f_i \cdot \log \frac{n}{df_i}$$

Where w_i refers to the weight of the *i* term in the document; f_i refers to the frequency of terms, n is the number of documents and df_i is the frequency of documents. Using this function, it is possible to limit the number of features selected for each document filtering all the features (i.e. words in the document) according to a selected threshold value of the weights.

Words with high tf-idf are more specific and carry more information with respect to the document where they occur, therefore weighting the document vectors with tf-idf provides an information that refers both to the content of a specific document and to the relation of this document to the dataset where it is inserted.

From these premises, the training process can be described as follow:

- The pre-processed documents are read together with their labels (determined according to the folder, i.e. the directory). Each pre-processed document is a list of tokens.
- The list of tokens is transformed into a numeric vector by a vectorisation process according to the following points:
    - Documents are read and dictionary of all the tokens is created, counting how many times each token occurs in the documents.
    - Each document is represented by a sparse vector of N dimensions, where N is the total number of words in the dataset. Each component of the vector corresponds to a word in the dictionary and the value is the number of occurrences of that word in the document.
    - The values of the vectors are weighted according to the ft-idf of each word.
- The classifier model is trained evaluating vectors value and associated labelled documents. At this stage a cross validation approach is used to select the model. Only the training dataset is used in this step. A subset of 90% of the dataset is used to train the model, the remaining 10% is used to assess the prediction performance.
- The trained model is saved and ready to use in the classification of unseen documents (prediction).

```python
# read categories from directory names
categories = [directory for directory in os.listdir(options.preprocessed_dir) if not directory.startswith("_")]

# read text files
log.info("reading text files from directory {0}".format(options.preprocessed_dir))
texts = []
y_data = []
for category in categories:
    txt_dir = os.path.join(options.preprocessed_dir, category)
    for item in os.listdir(txt_dir):
        txt_file = os.path.join(txt_dir, item)
        with open(txt_file) as f:
            tokens = [line.strip() for line in f.readlines()]
        texts.append(tokens)
        y_data.append(category)
log.info("read {0} text files from {1} categories: {2}".format(len(texts), len(categories), ",".join(categories)))

# pair texts with category labels and randomize the order
combined = list(zip(texts, y_data))
random.shuffle(combined)
texts[:], y_data[:] = zip(*combined)

# dummy function to use a custom preprocessor
def dummy(x):
    return x

# make feature vectors
log.info("creating TF-IDF feature vectors")
tfidf = TfidfVectorizer(analyzer='word',
                        tokenizer=dummy,
                        preprocessor=dummy,
                        token_pattern=None)
tfidf.fit(texts)

X_train_tfidf = tfidf.transform(texts)
log.info("created {0} TF-IDF feature vectors (dimensions: {1})".format(X_train_tfidf.shape[0],
                                                                        X_train_tfidf.shape[1]))
# train the classifier
log.info("training")
clf = LinearSVC(multi_class="crammer_singer",
                max_iter=100)
clf.fit(X_train_tfidf, y_data)

# saving the model
log.info("saving the model to file {0}".format(options.model_file))
with open(options.model_file, "wb") as p:
    pickle.dump(clf, p)
    pickle.dump(tfidf, p)
# test the classifier
log.info("testing")
X_test_tfidf = tfidf.transform(X_test_texts)
pred = clf.predict(X_test_tfidf)
```

Figure 66. Python algorithm applying tf-idf model and training the classifier model

Figure 66 reports the Python code representing the above listed passages. It is worth mention that the classification model was based on Support Vector Machine (SVM) algorithm. According to the literature review explored in chapter 5.1.1, SVM demonstrated the highest performance in terms of accuracy when applied on labelled text documents as in the case of text classification processes. Hence, the choice of this specific algorithm for the development of the Python prototype.

One can evaluate the ability of the algorithm in classifying unseen documents according to several classification metrics. These evaluation metrics are based on the comparison between the ground true labels obtained from the manual classification of documents and the labels predicted by the classifier model.

Accuracy is the rate of correct predictions with respect to all the predictions.

$$accuracy(y, \hat{y}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} 1(\hat{y}_i = y_i)$$

Precision is the rate of current predictions over all the predictions made. Hence, lower values of precision indicate high values of prediction mistakes in the classification.

$$Precision = \frac{true\ positives}{true\ positives + false\ positives}$$

Recall is the rate of corrected predictions over all the true values. A low value of recall indicates an high missing rate.

$$Recall = \frac{true\ positive}{true\ positives + false\ negatives}$$

F1-score is the harmonic mean of precision and recall, and it is computed for each class. The F1-score metric can be divided in micro-averaged F1-score and macro-averaged F1-score. The first one represents the aggregate of contributions of all classes to compute the average metric by considering the rates of each individual recall and precision separately. The second one represents the average of the F1-scores of all classes.

$$F1 - score = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

Figure 67 reports the Python script where the evaluation metrics above described are implemented.

```
# print evaluation metrics
cm = confusion_matrix(y_test, pred)
print ("Confusion matrix:")
print ("predicted:\t"+"\t".join(categories))
print ("true")
for i, category in enumerate(categories):
    print(category+"\t"+"\t".join([str(v) for v in cm[i]]))
print ("accuracy: {0}".format(clf.score(X_test_tfidf, y_test)))
print ("precision: {0}".format(precision_score(y_test, pred, average='macro')))
print ("recall: {0}".format(recall_score(y_test, pred, average='macro')))
print ("macro F1-score: {0}".format(f1_score(y_test, pred, average='macro')))
print ("micro F1-score: {0}".format(f1_score(y_test, pred, average='micro')))
```

Figure 67. Python script applying the evaluation metrics on the classification algorithm

Once the algorithm is trained and evaluated, it can be used to classify the entire dataset as shown in Figure 68.

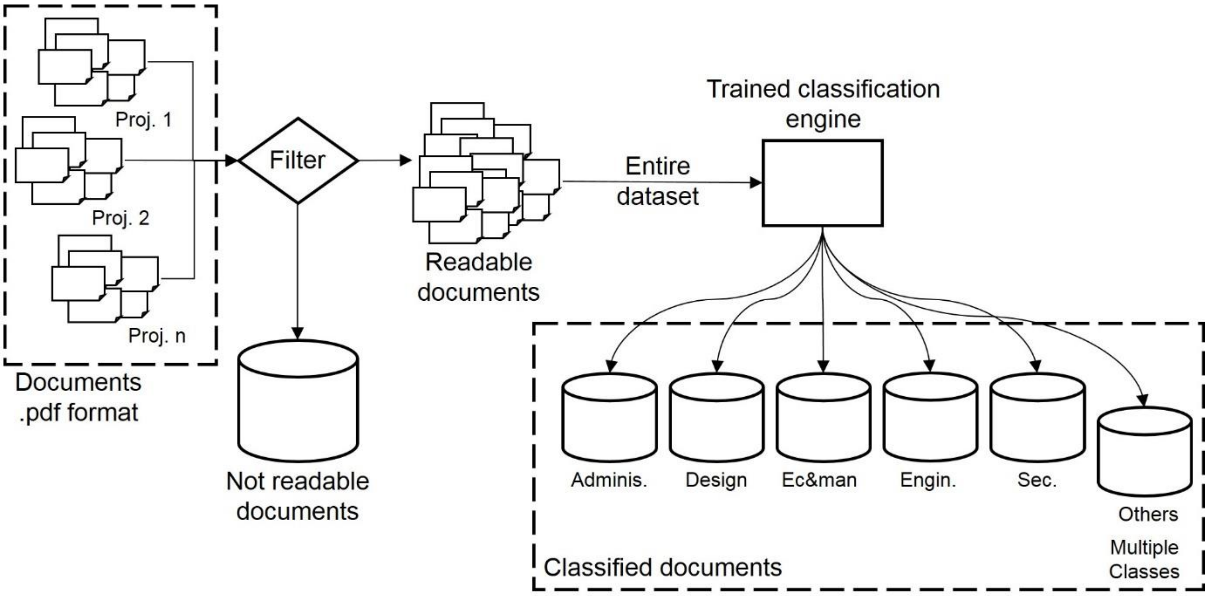


Figure 68. Application of the trained and tested classification engine

Figure 69 reports the Python script that predict the classification of unseen documents according to the trained classifier model.

```
# load the model
log.info("loading the model from file {0}".format(options.model_file))
with open(options.model_file, "rb") as p:
    clf = pickle.load(p)
    tfidf = pickle.load(p)

def predict(tokens):
    log.info("creating TF-IDF feature vector")
    #tfidf.fit([tokens])
    X_test_tfidf = tfidf.transform([tokens])

    log.info("predicting")
    pred = clf.predict(X_test_tfidf)

    return pred[0]

# read text files
log.info("reading text files from directory {0}".format(options.input_dir))
texts = []
y_data = []

for item in os.listdir(options.input_dir):
    if os.path.isfile(os.path.join(options.input_dir, item)):
        txt_file = os.path.join(options.input_dir, item)
        with open(txt_file) as f:
            text = f.read().decode("utf-8")
        tokens = preprocess(text)
        label = predict(tokens)
        label_dir = os.path.join(options.output_dir, label)
        if not os.path.isdir(label_dir):
            os.mkdir(label_dir)
        for ext in ['.pdf', '.PDF']:
            try:
                copyfile(os.path.join(options.pdf_dir, item.replace(".txt", ext)
                         os.path.join(label_dir, item.replace(".txt", ext)))
            except:
                pass
```

Figure 69. Python script applying the prediction process based on the trained classified model

5.1.2.3. Clustering process

The clustering process starts from the results obtained from the classification process. Hence, the starting point is a set of folders each one containing documents belonging to a specific class, reusing the same pre-processing procedure of classification process. According to the objectives and motivations of this research, hierarchical clustering algorithms represent the optimal solutions (in terms of expected results). However, due to the limited number of clustering applications on text documents in the construction sector, the results are compared with one of the most diffuse clustering algorithm, i.e. K-means.

The result of a hierarchical clustering algorithm is a *dendogram* that can be broken at different levels to yield different clustering of the data. Hierarchical clustering algorithms can use both a bottom up or a top down approach. In a bottom up approach, each observation starts in its own cluster and clusters are then merged together. On the contrary, in a top down approach each observation starts in its own cluster and clusters are then divided in two derived clusters. With specific reference on the bottom up approach, the linkage criteria determines the metrics used for the merge strategy:

- Ward linkage minimises the sum of squared differences within all clusters.
- Maximum or complete linkage minimises the maximum distance between observations of pairs of clusters.
- Average linkage minimises the averages of the distances between all observations of pairs of clusters.
- Single linkage minimises the distance between the closest observations of pairs of clusters.

Figure 70 reports a schematic example of dendogram that can be obtained applying the algorithm to the Administration class, while Figure 73 and Figure 74 represent the dendogram exported from the application of the prototype.

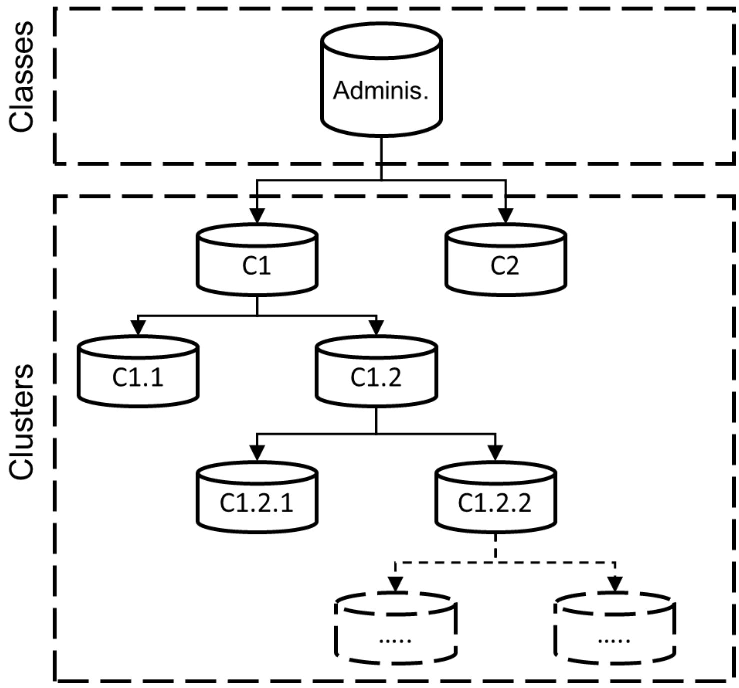


Figure 70. Graphical representation of a dendogram that can be obtained after the clustering activity

The K-means algorithm produce a flat clustering result according to a partition method. In particular, the algorithm starts from the central set of a cluster randomly selected. During each iteration, each sample point is assigned to the nearest cluster according to the similarity evaluation. Then, the centre of the cluster is recalculated.

The centre of each cluster can be calculated according to the following expression.

$$\mu_k = \frac{1}{N_k}\sum_{q=1}^{N_k} x_q$$

Where N_k is the number of samples belonging to cluster k, and μ_k refers to the centre of cluster k. Usually the application of K-means algorithms requires the definition a-priori of the number of clusters. This approach can hinder the dynamicity of the clustering approach. Nevertheless, it is possible to automate the clustering process implementing a specific method (called Elbow Method) able to select the optimal number of clusters using the K-means algorithm. The Elbow Method evaluate the quality of the clustering results increasing progressively the number of clusters used. The optimal number of clusters is reached when (and if) the quality drops dramatically. Figure 71 reports the algorithm developed to define the clustering process including K-means, the Elbow Method and the hierarchical clustering.

```
# read categories from directory names
categories = [directory for directory in os.listdir(options.preprocessed_dir) if not directory.startswith("_")]
#categories = ["Engineering"]
# load the model
log.info("loading the model from file {0}".format(options.model_file))
with open(options.model_file, "rb") as p:
    clf = pickle.load(p)
    tfidf = pickle.load(p)


# read text files
log.info("reading text files from directory {0}".format(options.preprocessed_dir))
texts = []
y_data = []

for category in categories:
    txt_dir = os.path.join(options.preprocessed_dir, category)
    vectors = []
    original_files = []
    for item in os.listdir(txt_dir):
        txt_file = os.path.join(txt_dir, item)
        original_files.append(item)
        with open(txt_file) as f:
            tokens = [line.strip() for line in f.readlines()]
        vectors.append(tfidf.transform([tokens]).toarray()[0])
    vectors = np.array(vectors)

    # determining the optimal number of clusters with the elbow method
    if (options.elbow and options.algorithm == "kmeans"):
        wcss = []
        cluster_min = eval(options.clusters_min)
        cluster_max = min([eval(options.clusters_max), len(vectors)-1])
        for n_clusters in range(cluster_min, cluster_max+1):
            log.info("[elbow] Testing {0} documents in category {1}, n. clusters = {2}".format(len(vectors),
                                                                                                  category, n_clusters))
            kmeans = KMeans(n_clusters=n_clusters, random_state=0)
            clusters = kmeans.fit(vectors)
            wcss.append(kmeans.inertia_)
            cluster_labels = kmeans.fit_predict(vectors)
            #print (n_clusters, silhouette_score(vectors, cluster_labels))
        distances = []
        for i in range(cluster_min, cluster_max+1):
            distances.append(distance_to_line(i,wcss[i-cluster_min],
                                              cluster_min,wcss[0],
                                              cluster_max,wcss[cluster_max-cluster_min]))
        n_clusters = distances.index(max(distances))+cluster_min
    else:
        n_clusters = min([eval(options.num_clusters), len(vectors)-1])
 # clustering
 if options.algorithm == "kmeans":
     clusterer = KMeans(n_clusters=n_clusters, random_state=0)
     log.info("Clustering {0} documents in category {1} with {2} clustering (n. clusters = {3})".format(len(vectors),
                                                                                                   category,
                                                                                                   options.algorithm,
                                                                                                   n_clusters))

 elif options.algorithm == "agglomerative":
     clusterer = AgglomerativeClustering(n_clusters=n_clusters)
     log.info("Clustering {0} documents in category {1} with {2} clustering (n. clusters = {3})".format(len(vectors),
                                                                                                   category,
                                                                                                   options.algorithm,
                                                                                                   n_clusters))

 else:
     log.error("clustering algorithm unknown: {1}".format(options.algorithm))
 clusters = clusterer.fit(vectors)
```

Figure 71. Python script applying clustering algorithms: hierarchical clustering and K-means

The quality of clustering results can be evaluated using the Dunn Index (DI). The DI is an attempt to find a good intra-group and inter-group relationship. Its mathematical representation is as follows:

$$DI_m = \frac{\underset{1 \leq i < j \leq m}{min} \cdot \delta(C_i, C_j)}{\underset{1 \leq i < j \leq m}{max} \cdot \Delta_k}$$

Where m is the number of clusters and DI_m is the distance formula that depends on m. $\delta(C_i, C_j)$ indicates the metric that defines the intercluster distance between C_i and C_j. Δ_k represents the average distance of group elements to the centre of group k.

It is worth noting that even in the case of unsupervised algorithms - like the ones related to hierarchical clustering - it is critical to introduce domain knowledge in the process. All clustering algorithms will produce clusters with any type of data set, regardless of whether the data contains clusters or not (Jain, Murty and Flynn, 1999). Thus, in the case of hierarchical clustering the algorithm does not require the identification of the number of clusters as a starting input.

However, it is required to understand where to cut the dendogram to obtain a reasonable picture. In the context of this research, this need can be moved from the algorithm itself to the developers that will use the structure to provide services and applications. Hence, in this case the domain knowledge will be introduced after the execution of the clustering path and according to the specific requirements defined for the development of a service and/or an application.

5.1.3. System experimentation and results

In order to test the process and the Python prototype described in section 5.1.2, the Python scripts were applied on a document dataset collected from the industry thanks to the collaboration with a construction company and an engineering company.

5.1.3.1. Illustrative example

Starting from the entire dataset collected from the industry, a subset of documents was prepared to train and test the classification algorithm and to tune the clustering process. 859 documents were manually classified according to the 5+1 classes defined in chapter 5.1.2.1. Around 60% of the entire dataset was used to train the model (529 documents as following listed), while the remaining 40% was used for testing processes.

Table 7. Composition of the training dataset

Tranining Dataset	
Class	**Number**
Administration	113
Design	29
Economic&Management	168
Engineering	132
Security&Safety	52
Others	45

Training and testing the prototype on the above described dataset it is possible to evaluate the performance of the classification process. Each classification metrics is evaluated over five iterations. Table 8 shows the results of the evaluation metrics underlining the performance of the prototype.

Accuracy, precision and recall reaches respectively 0.936, 0.948 and 0.906. Focusing on macro F1-score it can be noted an accuracy above 0.95 in three cases. These aspects demonstrate how the automatic classification framework is able to resemble human performance.

Table 8. Evaluation metrics obtained over five iterations of all classes

	Accuracy	Precision	Recall	Macro F1	Micro F1
Iteration 1	0.89	0.93	0.85	0.88	0.89
Iteration 2	0.98	0.99	0.93	0.95	0.98
Iteration 3	0.96	0.95	0.97	0.96	0.96
Iteration 4	0.87	0.90	0.79	0.82	0.87
Iteration 5	0.98	0.97	0.99	0.98	0.98
Average	0.936	0.948	0.906	0.918	0.936

Finally, a confusion matrix was developed to comprehend the behaviour of the algorithm in predicting the documents. Table 9 reports the confusion matrix obtained using a randomly selected subset of the testing dataset comparing the predicted labels (rows of the matrix) and the ground truth labels (column of the matrix).

Table 9. Confusion matrix resulted from the training and test process. The first row identify the number of documents used to train the algorithm and the number of document use to evaluate the prediction capability

Train/Predict	**113/17**	**29/2**	**168/13**	**132/11**	**52/4**
Predicted/true	**Administrative**	**Design**	**Economic_Mng**	**Engineering**	**Security_safety**
Administrative	14	0	0	0	0
Design	0	2	0	1	0
Economic_Mng	1	0	13	1	0
Engineering	1	0	0	9	0
Security_safety	1	0	0	0	4

Starting from the results of the classification process it is possible to evaluate the performance of the clustering activity. A Dunn Index was used to evaluate the quality of the clustering results.

Figure 72 shows the results obtained for different numbers of clusters applying the algorithms selected on the administrative folder. Larger the value of the Dunn Index, better is the cluster result. It is clear how the K-means clustering process tends to lost in quality using more than 6 clusters, while the use of agglomerative hierarchical clustering with complete linkage provide the better quality results in the clustering process.

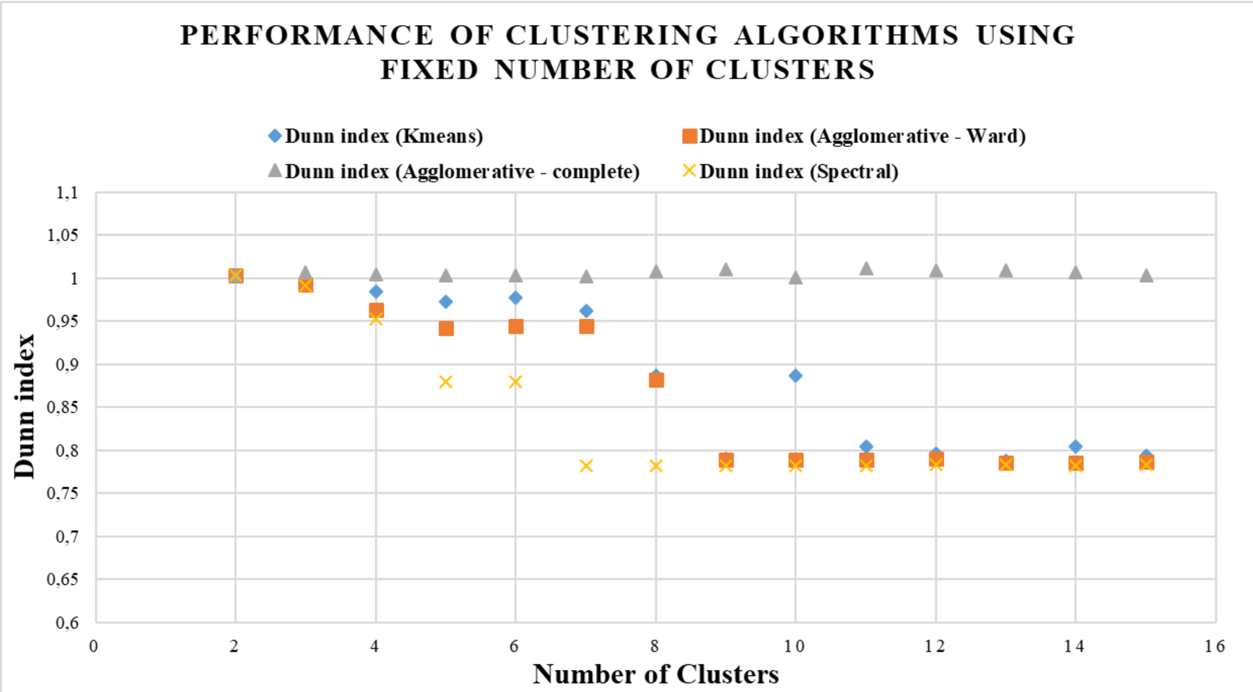


Figure 72. Dunn Index calculated on different clustering applications on the administrative folder

Figure 73 and Figure 74 show the dendogram obtained applying the hierarchical clustering to the administrative folder and to the engineering folder. This graphical representation makes clear the structure of the tree defined through the clustering algorithm. On the top of the tree, there is the entire folder of documents while on the bottom of the three there are the single documents contained in the starting folder.

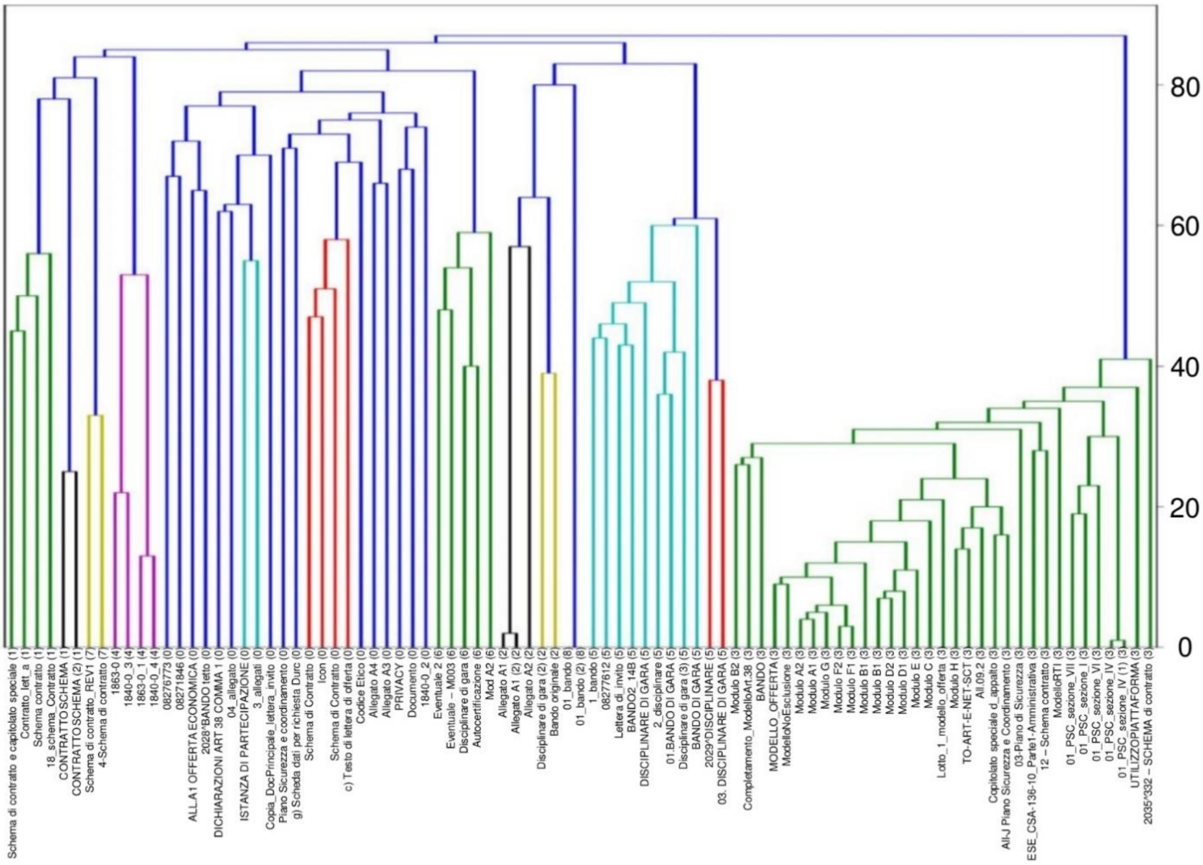


Figure 73. Dendogram obtained from the hierarchical clustering of the administrative folder

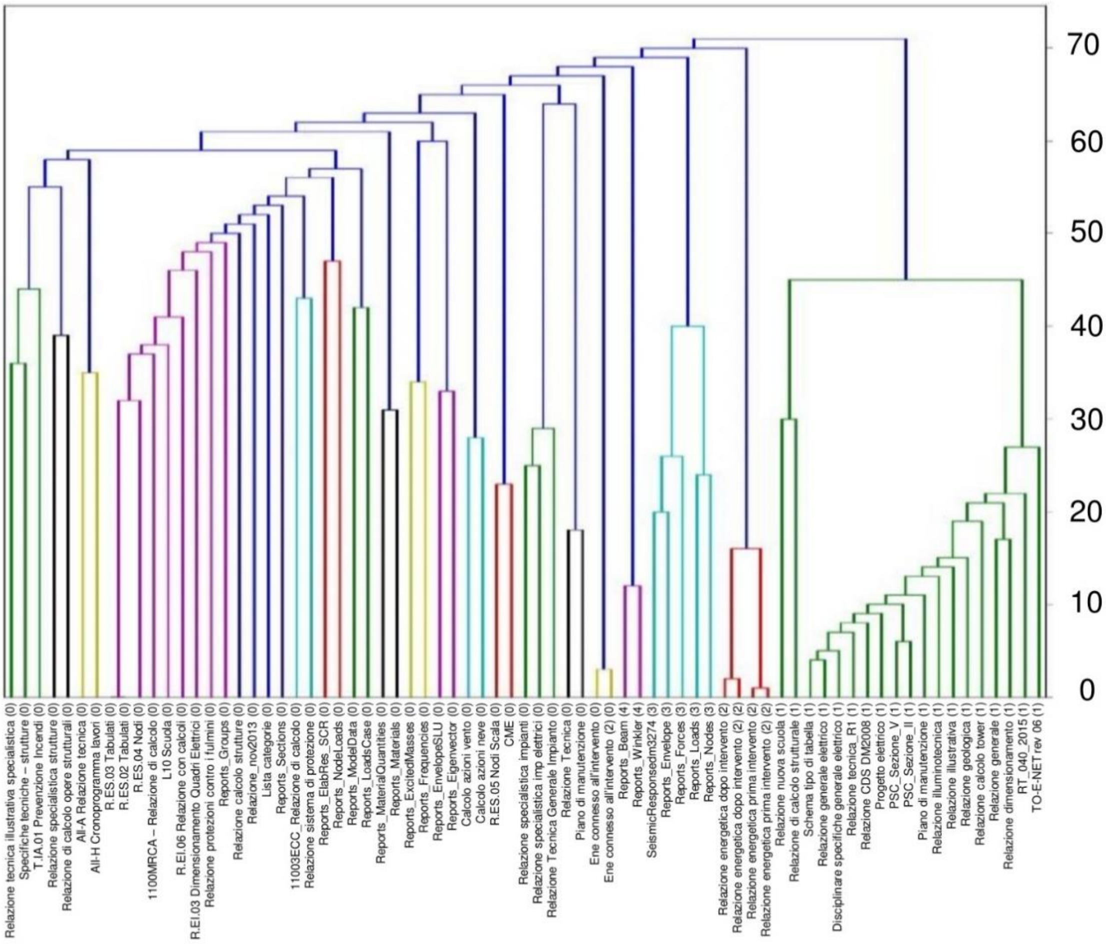


Figure 74. Dendogram obtained from the hierarchical clustering of the engineering folder

5.1.4. Discussion

The results presented in chapter 5.1.3.1 demonstrated the ability of the algorithm in classifying the documents sets provided by disparate project environments in the five main classes established in chapter 5.1.2.1. However, the generalisation of the obtained results are limited due to two main factors. The pre-processing phases such as lemmatisation and filtering are language based. Hence, the algorithm can process documents in Italian language but it is unable to deal with documents in other languages. It is possible to deal with this technological aspect implementing in the algorithm multiple dictionaries. However, approaching this issue solely on the technological side may produce poor results. In fact, the definition of the classes and the ability of the algorithm in recognising the similarity between documents and thus classifying these documents depends on the context of application. Hence, changing language usually involves changing the context of study where different factors such as regulatory frameworks, construction sector cultures, and best practices can modify the overall structure of document contents altering the performance of the classification and clustering processes.

The results obtained from the clustering tests revealed a performance limitation in the proposed approach. Combining two machine learning algorithms - the second one starting from the results of the first one - generates a derived error that affect the clustering results. In fact, the clustering algorithm starts from a documents set that may contain documents classified in the wrong class because of the classification process performance. Hence, the performance of the clustering process are reduced due to the quality of the starting documents set that according to the classification performance can contain 5% to 10% of documents in the wrong class.

Moreover, the clustering algorithm demonstrated a peculiar aspect that is the organisation of documents according to the input project structure. This is caused by the uniqueness of each construction project that introduces specific characteristics to be recognised by the algorithm. Hence, at certain levels of the clustering hierarchy, the algorithm returns to the input data structure, losing its ability to aggregate documents according to their topic. Future research may consider the inclusion of semantic text analysis to limit the influence of project document structure on the background topics. These factors open the way to further research in order to understand how to generalise the obtained results.

Integrating the presented case study in the framework proposed in chapter 4 (i.e. defining the hybrid approach as a component of the "translation layer"), it is possible to limit the issues related to the generalisation of the case study. In fact, changing the perspective from a view directed to the development of a unique comprehensive solution (Figure 54) to the development of distributed solutions each one dedicated to the specific context organised on a multi-level structure (Figure 55) could allow the integration of multiple languages, contexts, and needs.

5.2. Image recognition for information check and semantic improvement

According to the limits related to BIM authoring tools in the representation of real world information explored in chapter 3.3, this chapter proposes a machine learning application to address the highlighted issues checking and enriching the informative content of building information models. The intuition behind the proposed experimentation is that graphical means for information exchange remain a fundamental point in the construction sector due to the complexity of the designed products. Hence, as individuals can read the design and correctly classify objects and information according to their context based knowledge, it could be possible to embed this knowledge in the machine and automate this task. This experimentation is related to the need of expose information in a computable form when this information needs to be used in analytics applications. Moreover, the same approach could be used to check the relation between graphical and non-graphical information improving the information quality of building information models before the development of data analysis applications.

5.2.1. Background

5.2.1.1. Challenges in representing semantics through building information models

Building information models represent one of the main component of BIM (Jernigan, 2007). A building information model can be interpreted focusing on technical issues related to its development (narrow view) or embracing interrelated functional, informational, technical and organisation/legal issues (broader view) (Volk, Stengel and Schultmann, 2014). When semantic is examined, both these views must be considered. A focus on technical issues is crucial because the ability of BIM authoring tools in representing objects can affect the possibility associated to the semantic representation of these objects. Nevertheless, the analysis of the issues that can arise from an incorrect semantic representation must take into account the interrelation with functional and informational aspects and consider the future uses of the model and of the information that it contains or with which it is related.

Two of the main components in the semantic representation of objects in building information models are the identification of object's category and the relationship between objects (Tang *et al.*, 2010). Nevertheless, a digital model where a component has no defined categories and no relations does not guarantee the unicity of its understanding (Tang *et al.*, 2010) due to the lack of its information. Furthermore, the characteristic information defining the geometry and the identification of the materials and properties that constitute the object are fundamental. The assignment of the correct category to a defined object depends both on the technical capability of the software and on the judgement of individuals that operate the software. In some software, the number of categories that can be assigned to an object is defined. Standard BIM categories include, for example, wall, roof, slab, beam, and column. Every object that is not in the embedded list of the BIM authoring tool must be assigned to a generic category and then it must be classified using a personalised attribute to guarantee its semantic identification (chapter 3.2.1). However, the use of classification system does not solve the issue related to the interpretation of graphical representation and can produce limitations in the semantic representation (Caldas, Soibelman and Gasser, 2005). Furthermore, manual classification methods can cause some reliance issues (Caldas and Soibelman, 2003).

The absence of a defined category induces the absence of defined rules and relationship that are defined starting from the specific category. Some BIM authoring tools allow the manual definition of embedded categories limiting the need of personalised attributes for the definition of object classification (Tang *et al.*, 2010). However, for these categories the relations cannot be defined in a personalised manner. Furthermore, the possibility of defining ad hoc categories increases the risk of human errors in the assignment of the category. Due to these limitations and due to the need to overcome the limits imposed by authoring environments and rigid data structures, nowadays the research field about semantic web and linked data is of primary importance as described in chapter 3.1.3.

The technical limitations of some software due both to the limited number of categories and to the rigid rules that regulate specific categories can push individuals that operate with the software in finding alternative ways to develop their design. The erroneous use of objects from individuals operating with BIM authoring tools can be seen in several projects. For example, in the development of building information models with Autodesk Revit the use of curtain system category to simulate frames in plasterboard walls, the use of wall category to simulate skirting, the use of roof category to simulate non-vertical walls are only some examples. These lasts are not necessarily BIM problems, some authoring tools like BricsCAD and FreeCAD are more flexible and they allow to assign different classes and semantic information to every kind of objects. However, we must consider the possibility that models contain the above-mentioned issues without limiting the solution to a specific authoring tool.

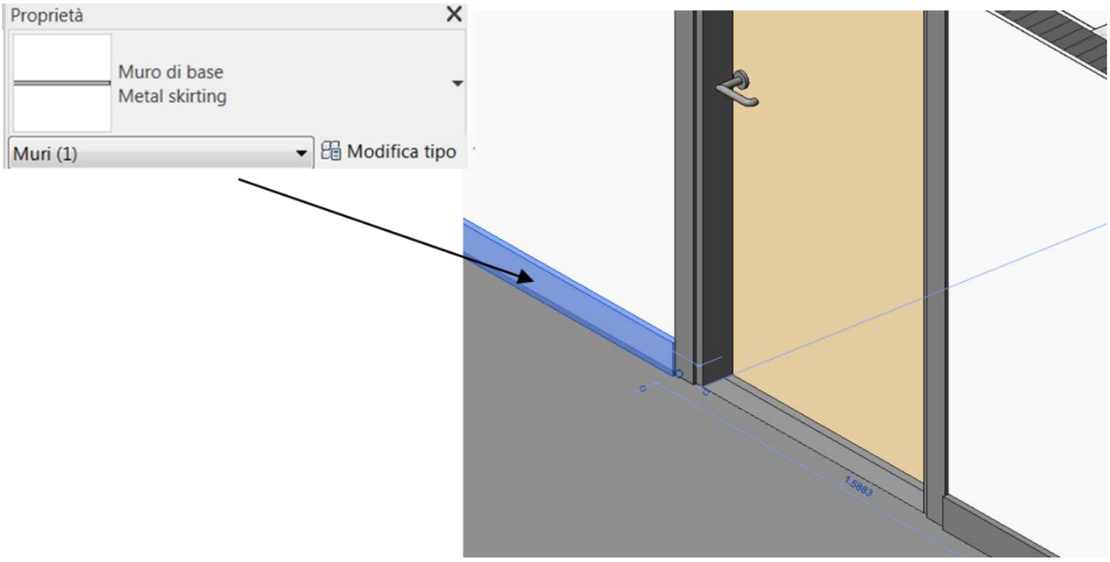


Figure 75. Example of metal skirting modelled as a standard wall

The difficulties in representing the semantics of buildings is highlighted in the case of existing buildings, in particular the historical ones. The development of BIM applications in the maintenance of historical buildings has demonstrated an increasing interest in last years (Volk, Stengel and Schultmann, 2014; Logothetis, Delinasiou and Stylianidis, 2015; Megahed, 2015). The experimentations developed on the *"Albergo dei Poveri"* in Genova (Musso and Franco, 2014), the Basilica of *"S. Maria di Collemmaggio"* in l'Acquila (Oreni *et al.*, 2013), the "Masegra" castle in Sondrio (Barazzetti *et al.*, 2015) and the Dome in Milano (Fassi *et al.*, 2015) are only some examples. These studies underline the difficulties in the semantic representation of historical elements in building information models. For example, the representation of complex elements like the spire of the Milan Dome, required the integration of three information sources, namely a 3D model, an external database and a photographic catalogue (Fassi *et al.*, 2013). However, the geometrical model does not embed object classification and rule-based relations between objects and all the information is managed through external means. Other research points out that commercial BIM software is not ready to be used for existing (in particular historical) buildings (Fregonese *et al.*, 2015). This is related both to the complex geometry of some objects and to the difficulties (and sometimes to the impossibility) to provide a correct semantic representation of objects in a BIM environment. Anyway, the case of historical buildings demonstrated the need of specific approaches pushing the development of dedicated Level of Development (LOD) (UNI, 2017a) and dedicated methods in the management of aggregated sources of information (Della Torre, Mirarchi and Pavan, 2017). However, the need to overcome the limits imposed by commercial BIM authoring tools remains an open issue. It is worth underlining that the use of different data structures outside of BIM authoring environment can remove these issues. This is one of the critical advantages in moving from the idea of models to the ideas of collaborative environments and beyond to the concept of platforms exposed in this thesis. However, the creation of models using commercial BIM authoring tools remains a fundamental passage and there is the need to manage the translation passages between these tools and other (more dynamic) environments. This experimentation aims to verify the possibility to partially automate these passages embedding the application in the *"translation layer"* (Figure 58).

5.2.1.2. Object and image recognition in the construction sector

The use of object recognition in relation to BIM is not new. In fact, the advent of digital technologies and techniques for the survey of existing buildings, such as laser scanning techniques (Mahdjoubi, Moobela and Laing, 2013) introduces the need to convert efficiently the generated point clouds in building information models. The manual modelling of a facility starting from a point cloud scan is a time consuming activity that often hinders the use of these technologies and techniques in the practice (Tang *et al.*, 2010). Hence, the need to improve the performances in the development of building information models starting from point cloud models imposed the exploration of new means to recognise objects in the point cloud and translate these into building information model objects. To create the building information model, geometrical and topological information of building elements has to be gathered and completed with semantic property/attribute information (Volk, Stengel and Schultmann, 2014). Thus, in the conversion process from point cloud to building information model can be identified two main tasks, i.e. the geometrical modelling task and the semantic modelling task. This last includes the identification of object's category, materials, properties and relationships between components (e.g. intersection and connection) (Tang *et al.*, 2010).

In the literature can be found several studies devoted to automate this process using different object recognition approaches. Three main approaches can be identified, i.e. 1) data-driven approaches based on the extraction of building information from captured and processed data, 2) model-driven approaches based on the definition of a predefined structure according to topologic relations and/or constraints used to match with captured data through knowledge or contextual information, and 3) other approaches including tagging and manual identification (Volk, Stengel and Schultmann, 2014).

Some experiments proposed combined approaches including 1) the classification of objects based on context information starting from a point cloud without semantic information (Xiong and Huber, 2010; Huber *et al.*, 2011), 2) the complete recognition of surface objects like walls in occluded environments (Oliver and Huber, 2010), and 3) the use of statistical approaches to handle uncertainty in the reconstruction of building interiors (Furukawa *et al.*, 2009). Furthermore, image-based approaches have been explored to verify as built documentations with particular reference to geometrical information (Klein, Li and Becerik-Gerber, 2012) and to progress monitoring and infrastructure asset management activities (Bhatla *et al.*, 2012).

From the literature can be observed that the majority of the applications of objects and images recognition in the construction sector are dedicated to the automatic construction of as built models starting from point clouds and/or camera images and the verification of as built documentation. However, even in this last case, the focus is on geometrical information and does not consider the information that can be extracted from images. Two main arguments can be underlined:

- Image and object recognition are extensively explored in computer science disciplines, however their application in the construction sector is limited and focused on conversion activities between point clouds and building information models. No efforts have been revealed in the literature in the exploration of alternative image and object recognition approaches using the advanced techniques proposed in the computer science field.
- The automated development of building information models starting from point clouds and/or images is bounded by the semantic representation limits presented in Section 5.2.1.1.

5.2.1.3. Motivation and aims of this research

As presented in Section 5.2.1.2 some applications of image and object recognition in the construction sector have shown the possibility of identifying building objects and automatically define and/or verify a building information model (Dickinson *et al.*, 2009; Volk, Stengel and Schultmann, 2014). However, the inclusion of semantic information remains a critical passage and only few studies have demonstrated the capability to obtain objects' classifications (Xiong and Huber, 2010). Furthermore, semantic information is still limited to standard information provided in BIM authoring tools and no experimentations have been found in the identification of non-standard semantic information and/or the inclusion of classes for non-standard objects. To demonstrate the possibility to deduce non-geometrical information starting from real world images and embed it in similar objects defined in building information models, this paper proposes an approach that uses deep learning image recognition algorithms (see Section 5.2.2.2) to recognise specific features in building information model objects. This objective is focused on the inclusion and/or the verification of non-geometrical information related to the geometrical data that could be missing in available commercial BIM authoring tools due to human errors and/or to the inability of the software in representing it. The use of an image based recognition process is justified by the limits demonstrated by the existing experimentation in semantic enrichments explored in section 3.2.2. Data quality remains a critical limit in these studies while the use of images can overcome these aspects.

5.2.2. Framework and methodology

The overall framework and methodology are described in this section. In particular, the logical process of the proposed image recognition checking system is presented in Section 5.2.2.1. It is based on an image recognition engine, presented in Section 5.2.2.2, used to classify images extracted as renders from objects contained in IFC models as proposed in Section 5.2.2.3. Finally, the results of the recognition process can be used both to check and/or to enrich the IFC model as described in Section 5.2.2.4. This last application can be directly related to the platform framework as proposed in the introduction of chapter 5. In fact, instead of reporting the corrected and/or enriched information into the IFC models, the same process can be used to report the information in a different environment using a different data structure to feed the platform. The process proposed in this experimentation was developed reporting the information in an existing IFC model to demonstrate in general terms its applicability. Nevertheless, changing the "writing phase" the information generated thanks to the application of machine learning algorithms can be used to export the information to other environments. Paving the way for future research, the same principle can be used to explore other applications of machine learning algorithm to optimise the translation and/or the verification of existing building models according to defined data structures and/or rules.

5.2.2.1. Logical process

Figure 76 shows the logical process of the proposed system. The process is logically divided in two main parts, namely project and analytics. This distinction highlights the different requirements in terms of data quality in the two applications. Thus, while in the project scenario the focus is on the identification of possible incoherence between geometrical and non-geometrical information, in the analytics scenario the focus comprises both the latter and the need of transforming all the characteristic geometrical features of objects in computable information to promote the application of data analysis processes.

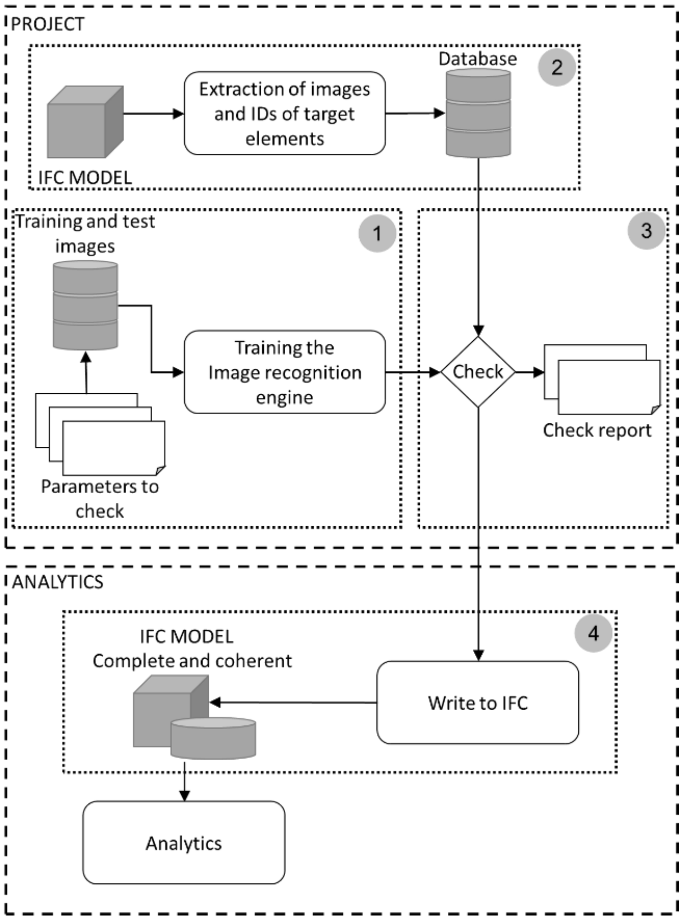


Figure 76. Framework of the process

Firstly, the parameters that need to be checked are defined. This is a manual and knowledge intensive activity because requirements can change according to the LOD of the objects, to the context (project and/or analytics), and to the specific software used to develop the original model. These parameters are used to collect and/or select in an existing dataset the real world images required to train the image recognition engine, 1 in Figure 76 (here based on a Convolutional Neural Network, ConvNet). According to the defined parameters, the images could need a manual labelling activity due to the difficulties in finding existing labelled image datasets (e.g. it is easy to find an image dataset that includes labelled doors and windows, while it is difficult to find an image dataset where windows with different number of openings are labelled). Nevertheless, this activity is required only in the first use of specific parameters because the defined image dataset can then be reused in the future applications. Once the image recognition engine has been trained and tested it can be used to recognise and label unseen images that can be extracted from a building information model. This activity consists in the extraction from a building information model (e.g. in the IFC format[37]) of images related to the objects according to the defined parameters. These images can then be stored in an external database where they can be related to the Globally Unique Identifier (GUID) of the originator object to maintain a direct link with the building information model, 2 in Figure 76. Finally, the images can be labelled using the image recognition engine[38]. The obtained results can be used both to verify the original model (3 in Figure 76) including the identification of possible incoherence between graphical and non-graphical information (for example erroneous

[37] This experimentation was based on the IFC schema to promote its generalisation. However, the same approach can be used with all the other model in their native form acting directly on the BIM authoring tool.

[38] It is worth specifying that all these passages, i.e. image extraction, database storing, image labelling are completely automated through an algorithm. No human intervention is needed.

classification of objects) and to improve the completeness of the non-geometrical information in the model reporting the information in the IFC model (4 in Figure 76).

5.2.2.2. Image recognition engine

The use of machine learning techniques for the development of algorithms able to classify large set of images has been studied long since in the field of Artificial Intelligence. Many current approaches on image recognition are based on machine learning (Krizhevsky, Sutskever and Geoffrey E., 2012). Among these techniques, ConvNets demonstrates high performances in image and object recognition tasks (Lawrence *et al.*, 1997; Karpathy *et al.*, 2014; Liang and Hu, 2015). ConvNets are deep-learning methods, i.e. *"representation-learning methods with multiple levels of representation, obtained by composing simple but non-linear modules that each transform the representation at one level (starting with the raw input) into a representation at a higher, slightly more abstract level"* (Lecun, Bengio and Hinton, 2015). Figure 77 shows a graphical representation of a ConvNet reported from Krizhevsky, Sutskever and Geoffrey E. (2012).

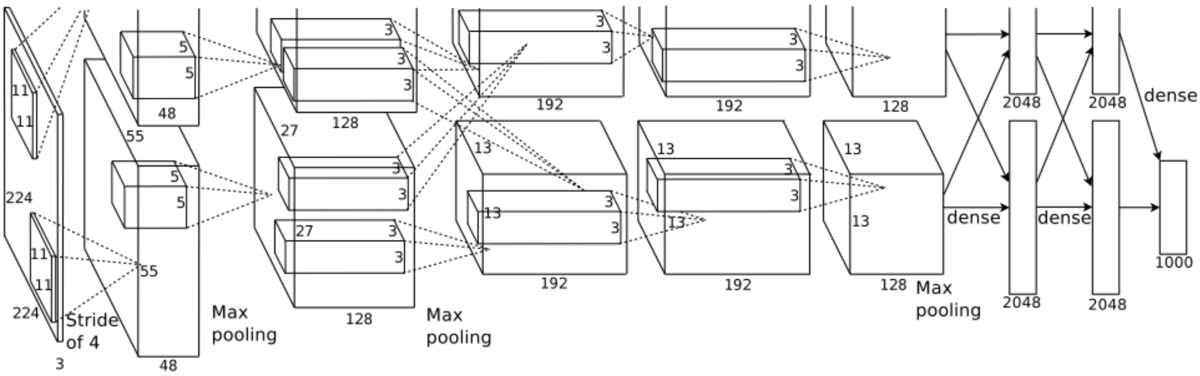


Figure 77. Graphical representation of a ConvNet from Krizhevsky, Sutskever and Geoffrey E. (2012)

ConvNets are characterised by four main layer types, namely convolutional layer, pooling layer, normalisation layer and fully connected layer (dense). In the convolutional layer, a set of filters is applied to the input image to identify the features (or weights) used by the algorithm to evaluate the image (convolution process). The pooling layer is used to reduce the image dimension in particular it is usually based on the extraction of the maximum value from the output images of the convolutional layer according to the pooling windows dimension. The normalisation layer, applying e.g. the Rectified linear unit (ReLu) acts normalising the outputs of the pooling layer removing negative value. Imposing y the output of the function and x the inputs from the pooling layer, the ReLu activation function can be represented in the following form:

$$\begin{cases} y = 0 \; if \; x < 0 \\ y = x \; if \; x > 0 \end{cases}$$

Finally, the fully connected layer is a neural network that takes as inputs the results of the process defined through several convolutional and pooling layers.

ConvNets demonstrated excellent performance in the recognition of both similar objects with a great difference in the images (for example two dogs in two complete different positions) and different objects that have similar features (for example a Samoyed dog and a white wolf) (Lecun, Bengio and Hinton, 2015). Hence, ConvNets can be used in the classification of objects but also in the identification of peculiar parameters such as the number of openings in windows or the type of a specific door.

Commonly, machine learning methods (deep or not) are based on a supervised learning procedure. This means that a training set of data or examples (e.g. a set of images) is labelled and processed by the machine to train itself. After the training procedure, the performance of the system is measured on a different set of examples called test set. There are several open data sets of images that can be used for training and testing tasks, e.g. the Google open images dataset (Google, 2018). However, available images are real world images representing real world objects, such as animals, furniture, or cars, and cannot be used to recognise elements in technical 2D representations. Nevertheless, in building information models the graphical representation of objects is based on 3D geometries simulating real world objects with different approximations according to the LOD of the objects. Hence, as starting point for the research, it is possible to assume using real world images to train and test the image recognition engine and use it to recognise and label images of virtual objects extracted from a building information model.

5.2.2.3. Image/data extraction from building information models

The IFC structure allows the exploration of all objects that compose an IFC model through its GUID. Hence, every object can be selected and its render can be extracted in an external environment and exported as image (e.g. in .jpg format) (Ifcopenshell). This process requires only a correct export passage from the native format to the IFC one. In fact, Ifcopenshell is a package that can be embedded in the program and can automatically read the IFC file and render the desired object. The IFC schema includes relations between objects allowing the extraction of images that can include not only a single object but also the object inserted in its context (e.g. a window in the hosting wall). The quantity of related objects can be defined in the extraction procedure imposing the level of the relations of the query. For example, considering direct relations the algorithm will retrieve the selected object and the objects related to it, while considering relation on a second level the algorithm will retrieve the selected object, the objects related to it and the objects related to them. This study focuses on queries maximum of first level.

5.2.2.4. Check and enrich the IFC

Once the image recognition engine has been trained and the images have been extracted from the building information model, it is possible to start the recognition process. The image recognition engine parses the images extracted from the building information model and according to the parameters imposed in the initial phases it provides the outputs (e.g. the category or a specific feature of the object) storing these lasts in a table. This information can then be compared to the information contained in the building information model to check the coherence between geometrical and non-geometrical information. Furthermore, if the information provided by the image recognition engine is not included in the model, it can be introduced in the model in its IFC format enriching its non-geometrical content (Mirarchi *et al.*, 2017) paving the way for future data uses. It needs to be highlighted that the definition of the requirements and consequently of the parameters to check and/or to introduce is still a manual activity. In fact, the same information can be introduced with different names and/or different meaning in the model due to the variety of standards and requirements that can be used and defined in each project. Hence, the definition of the rules remains a critical point that needs to be performed by expert users with domain-based and project-based knowledge.

5.2.3. System development and implementation

5.2.3.1. Prototype development

In order to test the process described in Section 5.2.2.1 a prototype was developed using Python programming language. Python is an object-oriented programming language and one of the most used with a continuous development of additional features. The main reasons to use Python in the development of the prototype are:

- Fast in development and test activities: Python can be executed directly without the need of compiling an executable format.
- Dedicated packages for the development of ConvNet (e.g. the Keras package (Keras, 2018)).
- Easy reading of the code for its dissemination in future research works.

The training images used in the prototype have been selected and labelled manually from available web image repositories. However, for large image recognition applications automated processes for research and fit of images can be executed through the use of dedicated image repositories, such as the Google open images dataset (Google, 2018).

5.2.3.2. Illustrative example

To explore the effective applicability of the proposed process an illustrative example was developed. It was designed to recognise a specific feature in windows modelled as objects in a building information model, i.e. the number of openings distinguishing between windows with two openings and windows with three openings. The process applied in the experimentation can be described as follows:

- Before starting the training of the image recognition engine, the image sets need to be defined and organised according to the identified requirements. In the experiment the requirement was the number of openings in windows. Hence, a training set of images representing windows with two openings and windows with three openings was collected and the images were labelled. A total of 90 images for each category was collected for the training set. Furthermore, as discussed in Section 5.2.2.2, the training activity requires the identification of a testing set of images. A total of 25 images for each category were collected for the test set. The two sets of images were defined with the requirement that no images contained in the training set were included in the testing set.
- The images included in the training set and in the testing set, were used to train the ConvNet according to the imposed requirements. The training and testing activities reveal the performance of the ConvNet and its ability to learn according to specific parameters. Hence, several tests were performed to define the combination of parameters that guaranteed the best performance of the system. The parameters and the learning curves resulting from these activities are presented in Section 5.2.3.3.
- Concluded the training and testing activities the image recognition engine can be used to recognise the images extracted from the building information models. As described in Section 5.2.2.3 images can be automatically generated from a building information model and stored in an external database. The images of the virtual objects can refer to a specific object alone or comprising the context in which the object is used. In this case study, both single windows and windows in their hosted wall were included in the validation activities. Moreover, the images of the virtual objects can represent the object with different angle views and choosing between black and white or colour images. Thus, during the experiment different combinations of images were tested to identify the effects of these changes on the image recognition performance.

5.2.3.3. ConvNet training and validation, results evaluation and definition of parameters

The algorithm used to define the ConvNet is based on several parameters that need to be calibrated to reach a correct learning curve and to maximise the performance of the image recognition activities. The dimensions of the filters and of the pooling side were maintained constant and respectively equal to 3x3 and 2x2. The number of iterations over the dataset provided (called number of epochs), the number of steps before declaring an iteration finished (called steps for epoch), and the number of steps to validate before stopping (called validation steps) were empirically tuned to reach the best quality results.

```
from keras.models import Sequential
from keras.layers import Conv2D
from keras.layers import MaxPooling2D
from keras.layers import Flatten
from keras.layers import Dense

classifier = Sequential()

classifier.add(Conv2D(32, (3,3), input_shape = (64, 64, 3),
                      activation = 'relu'))

classifier.add(MaxPooling2D(pool_size = (2,2)))

classifier.add(Flatten())

classifier.add(Dense(units = 128, activation = 'relu'))

classifier.add(Dense(units = 1, activation = 'sigmoid'))

classifier.compile(optimizer = 'adam', loss = 'binary_crossentropy', metrics = ['accuracy'])
```

Figure 78. Definition of the ConvNet and of the input images parameters

The steps for epoch and the validation steps can be evaluated according to the number of images in training (or validating) dataset and the batch size, i.e. the number of training examples used in one iteration. Using a batch size equal to 30, the steps for epoch have been defined equal to 4 while the validation steps have been defined equal to 1. Due to the low number of images contained in the training and validation dataset, the ConvNet showed an early overfitting (i.e. the ConvNet stopped learning and lost in performance) with a number of epoch greater than 55. The ConvNet has been tested using different combinations of parameters. The algorithm showed the highest performance around 50 and 55 epochs and with the steps for epoch and validation steps described above.

```
from keras.preprocessing.image import ImageDataGenerator
from keras import callbacks


csv_logger = callbacks.CSVLogger('Test.csv')


train_datagen = ImageDataGenerator(rescale = 1./255, shear_range = 0.2, zoom_range = 0.2, horizontal_flip = True)
test_datagen = ImageDataGenerator(rescale = 1./255)
training_set = train_datagen.flow_from_directory('dataset_fin/training_set',
                                                 target_size = (64, 64),
                                                 batch_size = 30,
                                                 class_mode = 'binary')
test_set = test_datagen.flow_from_directory('dataset_fin/test_set',
                                            target_size = (64, 64),
                                            batch_size = 30,
                                            class_mode = 'binary')
classifier.fit_generator(training_set, steps_per_epoch = 4,
                         epochs = 55,
                         validation_data = test_set,
                         validation_steps = 1, callbacks=[csv_logger])
```

Figure 79. Application of the ConvNet image recognition to the training and to the test sets

Figure 80 and Figure 81 show the learning curve identifying the accuracy and the loss functions (i.e. the function used to describe the error in the recognition process) obtained on the training dataset.

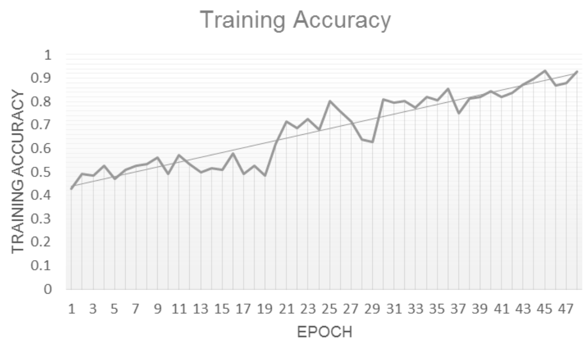


Figure 80. Training accuracy obtained in a test with 50 epoch, 4 steps per epoch and 1 validation step

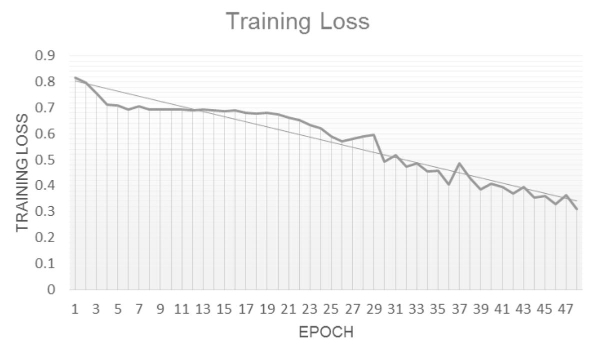


Figure 81. Training loss obtained in a test with 50 epoch, 4 steps per epoch and 1 validation step

Figure 82 and Figure 83 show accuracy and loss functions obtained on the validation set. In both the training and the validation curves, the ConvNet shows an increasing learning curve that tends to remain around 84% of accuracy at the end of the learning and validation procedure.

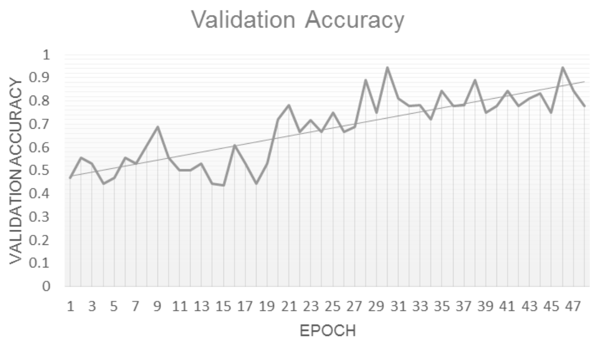


Figure 82. Validation accuracy obtained in a test with 50 epoch, 4 steps per epoch and 1 validation step

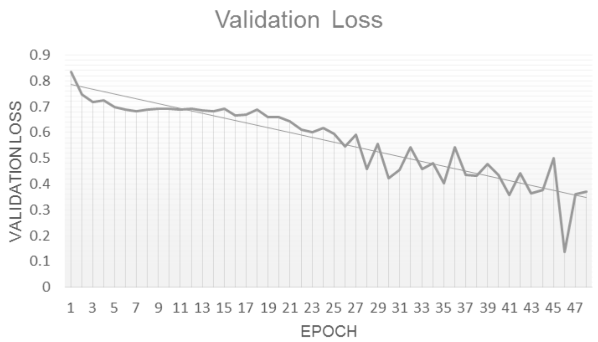


Figure 83. Validation loss obtained in a test with 50 epoch, 4 steps per epoch and 1 validation step

The performance displayed in the graphs refers to the recognition of real world images while the scope of the experiment was the recognition of simulated objects extracted from building information models. Hence, a second level of validation was defined where the trained ConvNet was used to recognise images extracted from building information models.

```
import numpy as np
from keras.preprocessing import image
L = []

for x in range(50):
    y = str(x)
    test_image = image.load_img('dataset_fin/single_prediction_fixed/' + y + '.jpg',
                                target_size = (64, 64))
    test_image = image.img_to_array(test_image)
    test_image = np.expand_dims(test_image, axis = 0)
    result = classifier.predict(test_image)
    training_set.class_indices
    if result[0][0] == 1:
        prediction = '3'
    else:
        prediction = '2'
    L.append(prediction)
    print ('your window has ' + prediction + ' openings')

file = open ('Results.txt','w')

for item in L:
    file.write("%s\n" % item)

file.close()
```

Figure 84. Application of the trained ConvNet to recognise the number of openings in the images

5.2.3.4. Validation and evaluation on building information model objects images

According to the process described in chapter 5.2.2, the image recognition engine can be used to recognise images extracted from building information models or features related to these images. The first passage of the algorithm consists in the automatic extraction of the images of the objects contained in the building information model. This specific prototype is limited to a feature related to the "IfcWindow" object type. Figure 85 reports the Python script used to progressively individuate all the ifcwindow elements in an ifc model and extract its render in a separated system window.

```
occ_display = ifcopenshell.geom.utils.initialize_display()

def render_entities(ent):
    for entity in ent:
        print 'RENDERING OF: ', ifcopenshell.entity_instance.id(entity)
        if entity.Representation:
            shape = ifcopenshell.geom.create_shape(settings, entity).geometry
            display_shape = ifcopenshell.geom.utils.display_shape(shape)
            if entity.is_a('IfcPlate'):
                ifcopenshell.geom.utils.set_shape_transparency(dispaly_shape, 0.8)
        raw_input()
        occ_display.EraseAll()

# Open the IFC file using IfcOpenShell
ifc_file = ifcopenshell.open(file_name)

objectType = 'IfcWindow'
entities = []
while (objectType != 'X'):
    try:
        entities = ifc_file.by_type(objectType)
        if (entities.__len__() == 0):
            print 'No entities of this type present in file'
        else:
            s = 'There are ' + repr(entities.__len__()) + ' entities of this type in this file'
            print s
            render_entities(entities)
        objectType = raw_input('insert a new IFC type or eXit: ')
    except:
        print 'no such type in IFC'
        objectType = raw_input('retry or eXit: ')
```

Figure 85. Automatic extraction of the images of IfcWindow objects contained in an ifc model

To validate the accuracy of the ConvNet in the recognition of images extracted from building information model objects a first dataset of 50 images was defined. As mentioned in the introduction of Section 5.2.3.2 the images can be defined according to different parameters including the angulation, the inclusion of the context and/or the inclusion of colours. Hence, this first dataset was composed mixing different characteristics for the extraction of the images to identify possible relations with the performance of the image recognition engine.

Figure 86 shows some examples of images related to a window in the building information model. In particular, image 1 represents a two openings windows with material textures including the hosting wall; image 2 shows the same window without context, with a frontal view in black and white; image 3 identifies the same window with a different angulation in black and white including the hosting wall.

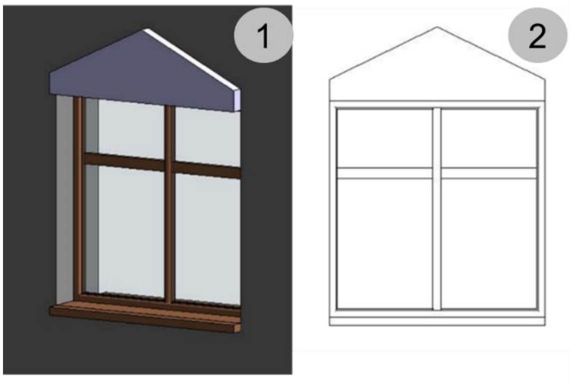


Figure 86. Example of images extracted from building information model windows

This first dataset was used to test the ability of the image recognition engine in the identification of the defined feature, i.e. the distinction of windows with two openings and windows with three openings. Results showed an overall percentage of correct recognition around 70%. Analysing the results, it was possible to identify some relations between the characteristics of images and the performance of the image recognition engine. In particular, the algorithm performed better in the recognition of isolated windows (i.e. images that do not include the wall hosting the windows, e.g. 2 in Figure 86), while the other parameters including colours and orientation did not affect the performance. This result is aligned with the characteristic structure of the ConvNet that "explode" the images and is not affected by colours or positions of the object, while wall can hide some parts of the windows with a consequent deterioration of the performance.

To validate this result, a second dataset of 50 images was defined including only images with isolated windows. In this case, the test was repeated 100 times reaching a mean accuracy of 82% with peaks of 90%.

5.2.4. Discussion

As described in Section 5.2.1.2 the use of image recognition has been proposed in several applications in the construction industry with specific reference to the reconstruction of geometries starting from point clouds. Some articles highlighted the accuracy obtained in the application of these techniques. For example, the experiment proposed in Huber *et al.* (Huber *et al.*, 2011) shows an accuracy of 93% in detecting and modelling wall openings, while Xiong and Huber (Xiong and Huber, 2010) compare Regularized Logistic Regression (RLR) and Conditional Random Field (CRF) in the recognition of objects based on context and obtained respectively 84% and 90% accuracy. Hence, the accuracy obtained in the described experiment is aligned with the earlier experiments in the construction sector that can be found in the literature. However, the developed prototype was trained with a small image dataset and an increment in the number of images used to train the model can increase the performance of the image recognition process according to the results obtained in other fields (Hadsell, Chopra and LeCun, 2006; Sun *et al.*, 2014).

Some studies propose the use of image recognition in the identification of object categories in existing buildings (Xiong and Huber, 2010; Volk, Stengel and Schultmann, 2014). However, the recognition procedure is limited to standard elements such as walls, windows, and doors and does not include elements that cannot be defined in BIM authoring tools. Furthermore, the non-geometrical information associated to the geometrical data is not included in the recognition process in those studies. This research proposes a novel framework that allows the recognition of both object's classification without limitations imposed by commercial BIM authoring tools and features recognition based on the geometrical representation of virtual objects. Furthermore, it demonstrates the ability of ConvNet in recognising specific features of virtual objects starting from real world images collected from the web and using them as training and validation sets.

A brief consideration about LOD is necessary. There is not a one agreed definition of LOD (Bolpagni and Ciribini, 2016), however it is commonly accepted that LOD can be interpreted as the combination between the geometrical and non-geometrical information related to objects in building information models (BSI, 2013; NBS, 2015; BIMForum, 2016; UNI, 2017a). LOD can refers to both Level of Development and Level of Detail. The former integrates the evolution of the project and thus comprises the changes in the information content according to the changing needs of the project stages. The latter is more related to the specific use of an information model, thus multiple level of details can coexist in a specific phase of the project according to the uses of the model. In this case study the focus is on LOD intended as level of development considering the entire set of information related to an object to check and enrich its content. Nevertheless, the ability to recognise specific geometric features in images extracted from building information models depends on the LOD of the object with specific reference to its geometrical components. The level of information required in the recognition, checking and enriching process cannot overcome the one contained in the object identified through its LOD. If the information is not represented neither in the geometrical nor in the non-geometrical form it cannot be inferred by the image recognition process.

5.2.5. Conclusions

This chapter proposed a framework for the verification of the coherence between geometrical and non-geometrical information in building information models and for the verification of the correct semantic representation of the objects that the model contains paving the way for future data uses. The proposed framework uses a machine learning based image recognition engine that demonstrated good performance in the recognition of images related to virtual objects starting from a training set of real world images.

The proposed system - an image recognition engine based on ConvNet - was trained according to predefined rules and requirements. After the training, images related to objects contained in building information models were extracted and stored in an external database where each image was associated to the GUID of the original object. Finally, the trained image recognition engine was used to recognise the defined features in the extracted images and use the inferred information to check and/or enrich the model. The effective interpretation of images by the machine can reduce the gap between the information transmitted graphically and the computable information, overcoming the limited semantic structure of existing BIM authoring tools. Furthermore, the correct interpretation of graphical information can be used in checking activities handling the accuracy and quality of data, i.e. the coherence between the geometrical and non-geometrical content of building information model objects. The system was tested by developing a prototype based on Python.

The research results show the capability of the algorithm in expose graphical information in a computable manner with recognition performance comparable with other studies in the field of image recognition. Nevertheless, the evaluation of the time required to develop the entire process, including the identification and definition of the parameters to check can be identified as a future line of research. However, some issues need to be addressed in the future research:

- The proposed solution is based on the recognition of singular elements identifying classes and/or other information that can be inferred from a graphical representation of the object in the model. However, for some elements this recognition is difficult due to the similarity between objects and the difficult interpretation of the scale in the extracted images (e.g. distinction between walls and skirting). Following the human comprehension and the studies under development in the field of automatic reconstruction of BIM from point clouds (Xiong and Huber, 2010), future developments can focus on the integration of context comprehension algorithms creating a double check between the image recognition of single objects and its identification in the context of the model.
- The definition of the recognition and checking rules is a time consuming and knowledge intensive activity. Hence, future research can focus on investigating the reuse of rules and requirements to improve the usability of the system in practical applications.

- In the same way, collecting and labelling images is still a manual process. Even if the images need to be collected only once, changes in the requirements can introduce changes in the related labels. Hence, the exploration of automatic means to collect labelled images using e.g. image web scraping algorithms can improve the quality and efficiency of the process.
- Finally, the proposed system uses ConvNet as image recognition engine according to its demonstrated performance in performing this task. However, other image recognition algorithms can be explored to compare the performance.

6. Implications

"Knowledge is anything that increases your ability to predict the outcomes."
(Daryl Morey, from (Michael Lewis, 2017)

The study and the related results proposed in this research provide two main implications. On the one hand, the extensive analysis of the existing means to collaborate in relation with the study of the social aspects on collaboration in the construction sector, provides the instruments to develop a structured interpretation on the evolution of collaborative environments in the construction sector. This interpretation is shaped by the framework proposed in chapter 4 that identify a possible future approach to collaboration in the construction sector. On the other hand, the proposed framework can be interpreted as an instrument for the development of platforms in the construction sector. Because of the existence of different projects developed at national level, it is worth studying the impact and the integration of the proposed framework with the existing projects and instruments highlighting the added value of the proposed work.

6.1. Evolutions of collaboration environments

We presented in chapter 2.1 an analysis related to the concept of knowledge focused on the construction sector and on the impacts of the knowledge management processes. This analysis highlighted the impact of previous knowledge in the interpretation of new facts and how the context and the available information can shape the way in which knowledge is perceived. The same issue is identified in the development of research activities where the introduction of researchers' knowledge drives the interpretations of phenomenon and thus the expression of the results. According to Chua (1986) can be identified three research epistemologies, namely positivist, interpretive and critical studies. Positivist is directed to study phenomenon with structured instrumentation including quantifiable measure of variables, defined hypothesis tested during the study and the identification of formal propositions. Critical studies start from a status quo and try to analyse its structural contradictions identifying the differences between the existing social practices. On the other hand, interpretative approach sees the inclusions of social aspects as a fundamental point to understand phenomenon and provide an interpretation that can follow the view of the operators in the field. Because of the great impact of social aspects in the construction sector – already analysed in chapter 2.1 – the interpretative approach seems to be the most appropriate to provide a holistic interpretation of the possible evolution of collaborative environments. However, the proposed analysis is shaped by the knowledge gained by the author during the development of this thesis and is well expressed by the path described in the evolution of the different chapters with particular reference to chapters 2, 3 and 4. Nevertheless, Orlikowski and Baroudi (1991) argued that the interpretive schemes of researchers always intervene shaping the information gained and its representation in the research path.

The proposed representation of the evolution of collaborative environments in the construction sector is reported in Figure 87. On the horizontal axis is identified the evolution of ICT technologies. This representation does not show the state of the art of these technologies, but it underlines their integration in the construction sector with specific reference to the development of collaborative environments. On the vertical axis, it is reported the perception on the content treated by the collaborative environments. The four stages identified on the vertical axis represent the evolutionary milestones and will be used to detail the proposed schema. These stages include two aspects in the content perception evolution. On the one hand, there is a cultural change in the operators of the construction sector. In fact, the evolution of the technology must follow a change in the social and cultural approach to the technology itself. This corresponds to a change in the users' approach to collaborative environments. On the other hand, thanks to the technological evolution, machines are nowadays able to interpret the data, to read the documents and extract valuable information, and to transform file instance in computable and valuable sources of data. Hence, the evolution of the content perception can be read in terms of change in the capability of the machines in interpret the contents parsed in collaborative environments and through related and/or integrated platforms.

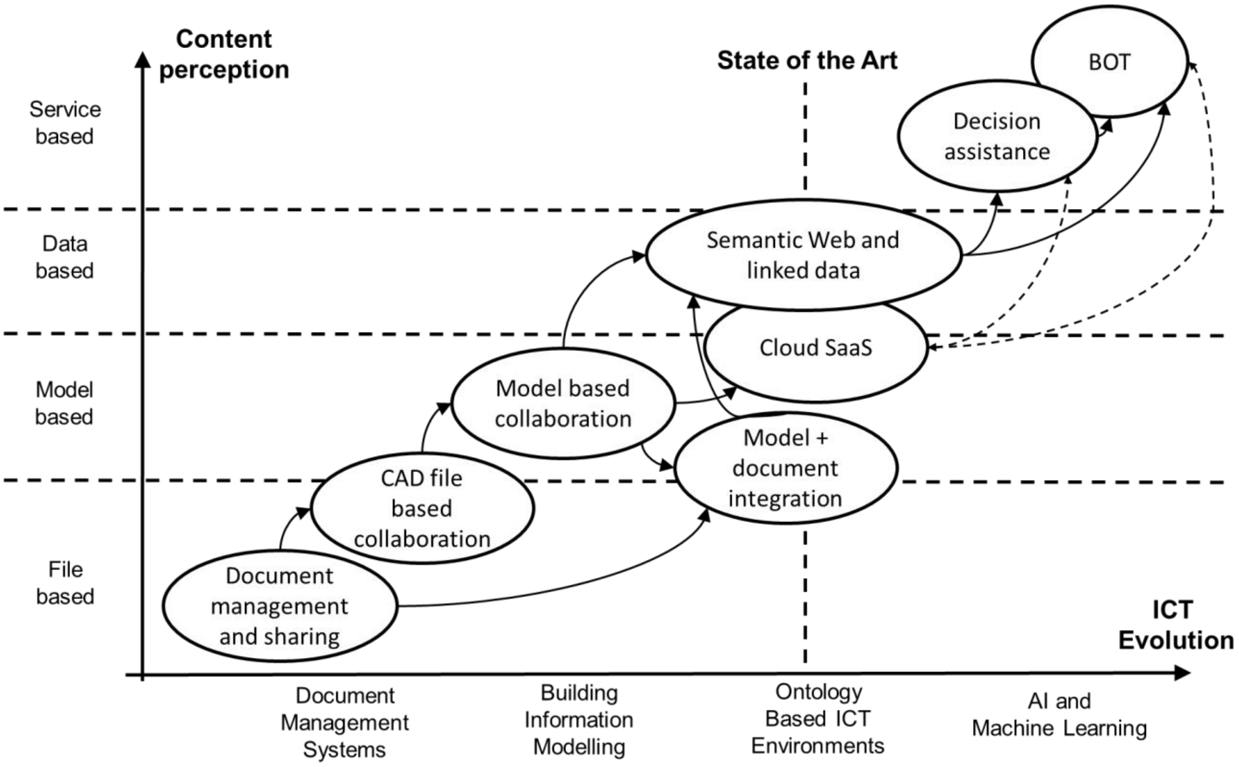


Figure 87. Evolution of collaborative environments

It is worth noting that the evolution of the technological solutions presented in the graph can follow different path and many times the pushes from the industry and the research fields produced the concurrent development and exploration of different solutions with an overlap between one to each other. Hence, even if the proposed description follows a linear evolution, it must be interpreted in a dynamic environment where the different areas are mixed and can coexist multiple interpretation on different levels.

6.1.1. File based perception

The file based perception level refers to the interpretation of documents and/or file shared and managed using collaborative systems as impenetrable entities. Documents are identified as information and knowledge carriers. However, this information is embedded in the document and its use is reserved to single subjects that must open a document and read its specific content. Many times, due to the traditional forms used in the transmission of information (e.g. text, images or drawings), the time required to retrieve the desired information imposes extra costs to the collaboration process. With the advancement of IT, document management systems start to introduce the possibility to add metadata to the documents to facilitate the identification of the main information related to the documents itself. However, the introduction of metadata remains a manual process subject to human errors.

The advent of CAD instruments in the design offices introduces new potentialities in the development of collaborative means. CAD files can be linked one to each other and can contain and embed standardised information with the consequent definition of shared processes for the identification of collaborative actions. The progressive use and direct exchange of proprietary formats and de facto standards such as DXF (drawing/data exchange format) as well as the introduction of drawing layering and the related standards, impact substantially on the ability of collaborate in the construction sector. Nevertheless, the advent of CAD technologies did not change the perception on file and document contents that remains obscure to the machine. Drawings can be shared according to defined standards, but the graphical information is represented in CAD environments as vectors then interpreted by technicians as parts of a design. In the same way, not graphic information is reported on drawings manually and cannot be linked to other documents. The advent of object oriented models starts changing this perception introducing the ability to include geometric and not geometric information in the same environment – than usable by a dedicated collaboration environment – paving the way for a change in the perception of file contents and a consequent evolution in the development of collaborative environments.

6.1.2. Model based perception

The capability of object-oriented environments to include geometrical and not geometrical information as well as relations between objects pave the way to the introduction of model based collaborative environments. After the introduction of the firsts object-oriented solutions, it was clear the need to define more coordinated standards to push the development of collaborative processes. From the STEP application protocol, the International Alliance for Interoperability defined the IFC (see chapter 3.1.2). Starting from this standardised base, it was possible defining collaborative environments able to explore the models and change the perception from an impenetrable document to a navigable model. Due to the crucial importance of collaborative processes, in recent years the commercial software houses changed the focus of their development efforts pushing the introduction of dedicated collaborative environments based on proprietary formats and able to read open formats like IFC.

Introducing these collaborative environments in the industry it becomes clear the need to integrate document management systems and model based collaborative systems. Hence, the development of hybrid systems able to explore building information models and relate the objects contained in these models with external documents. Nevertheless, this evolution did not change the perception on documents that are external to information models. In fact, while the information contained in information models are usable directly by the machine, the related documents remain black boxes impenetrable by the machine. Moreover, the link between documents and objects remains a manual and rigid task based on human interpretation.

In parallel with the development of these integrated systems, the research community starts exploring the possibility to develop processes based on the software as a service paradigm. This approach is based on the installation of all the required instruments on remote machines providing the use of these tools as services accessible via the web. However, this approach introduces some drawbacks such as the visualisation of the model and its property (also in terms of responsibility in the introduction of information) that must remain in the hand of the specific subject in charge to work on the specific component of the model.

The difficulties demonstrated by the effective application of a model-based perception in the industry such as the limits imposed by both proprietary formats and IFC in sharing efficiently the information (and thus in providing effective interoperability between systems), defined the need to explore different approaches. The introduction of linked data based on semantic web technologies starts the way towards a second change in the perception on the content, moving from a model based perception to a data based perception[39].

[39] *It might seem strange to place data after models. Nevertheless, the vertical axis represents the content perception in the collaboration environment. Of course, data and information should be organised according to models also at the "data based perception" level. However, the perception of users and the focus of the environment should be on the single data instead of a massive model (reducing the misinterpretation of information models by the industry).*

6.1.3. Data based perception

The developments in the research on the field of ontologies, linked data and semantic web in the construction field signal the start of a critical change in the interpretation of models and in the possibilities related to their use (see chapter 3.1.3). The development of environments based on these technologies allows the effective access to single pieces of information instead of providing access to information through the exploration of a massive model (or several massive models). Moreover, thanks to the definition of standardised and shared ontologies, both humans and machines can understand and use the information uploaded in the collaborative environment empowering the interoperability between systems and the possibility related to the development of collaborative processes. Thanks to the possibility of act on specific information, for each information can be defined specific privacy criteria and each information can be shared between different systems and used in a linked environment. This can change the interpretation of collaborative environments and platforms moving the perception from a single place where information models can be shared and stored to a distributed ecosystem where the information is dislocated in different places and linked. Nevertheless, the use of semantic web technologies does not represent the only line of research and other studies are exploring different data structures to explore the information content of models and expose this information in shared environments and/or analysis environments.

This area of research is nowadays one of the most promising ones and the increasing efforts spent by the research community in this direction is pushing its incremental development.

6.1.4. Service based perception

Once the open issues related to the use of collaborative environments that the research community and the industry are facing will be solved (at least in part), the focus of collaborative environments should be on providing added value to the final users in the form of services. Following this direction, the perception will change from the effective capability of the system to deals with models and/or single data to its ability in providing usable services for the final users. This dimension is already under development for example by TNO (2018) that proposed the introduction of BIM BOTs to support the integration of the different stakeholders involved in the process. The platform framework proposed in this research highlights some possible future developments based on the integration of data provided by multiple environments empowering the possibilities offered by the introduction of analytics based on machine learning and/or artificial intelligence applications. Hence, the possibility to introduce services able to support designers and technicians in day-to-day decision making processes though knowledge-based analysis (limiting the existing issues underlined e.g. in chapter 2).

6.2. Relation between existing platforms and the proposed approach

The literature review presented in chapters 3 and 4.2, underlines the diffused efforts of both industry and research in the development of solutions devoted to improve communication, collaboration and interoperability. These solutions can assume different forms. However, the idea of integrated digital platforms is gaining interest as demonstrated by several experimentations and projects that can be found in the European context. For example, the Drumbeat research project, concluded in December 2017, reports one of the first experimentation in the development of means for the effective link of information through web environments. The project is based on the use of semantic technologies and the principle of linked building data already presented in chapter 3.1.3.

In addition to these experimentations, at national level can be identified in the EU context three main projects devoted to the development of a digital platform for the construction sector. The analysis of these projects can constitute a good basis to understand the possible implications of the proposed framework highlighting the evidences matured thanks to the development of the related case studies. The national project mentioned are synthesised in Table 10 and briefly described in the following.

Table 10. Synthesis of the main national project in the development of digital platforms

Date	Name	Nationality	Description
2017	Kroqi	France	CDE Server, project oriented; Basic integrated services and open API to allow the development of integrated and structure services/applications from private and public bodies.
2015	BIMToolkit NBL Library	United Kingdom	BIM Library; Standard definition of geometric an informative attributes (LOD-LOI); valuation Toolkit of objects LOD and LOI level in a model.
2013	INNOVance	Italy	BIM Library for manufacturer objects, standard attributes; BIM Server for models and projects, tenant management; BIM-GIS platform for buildings, infrastructures and site.

6.2.1. Kroqui – France

In France the *Kroqi* platform developed by CSTB constitutes a virtuous example for the development and use of a dynamic digital platform in the construction industry. The platform is developed on a core basis that provides fundamental services and a central orchestration ecosystem. In this context the terms platform refers to a collaborative environment focused on the project level.

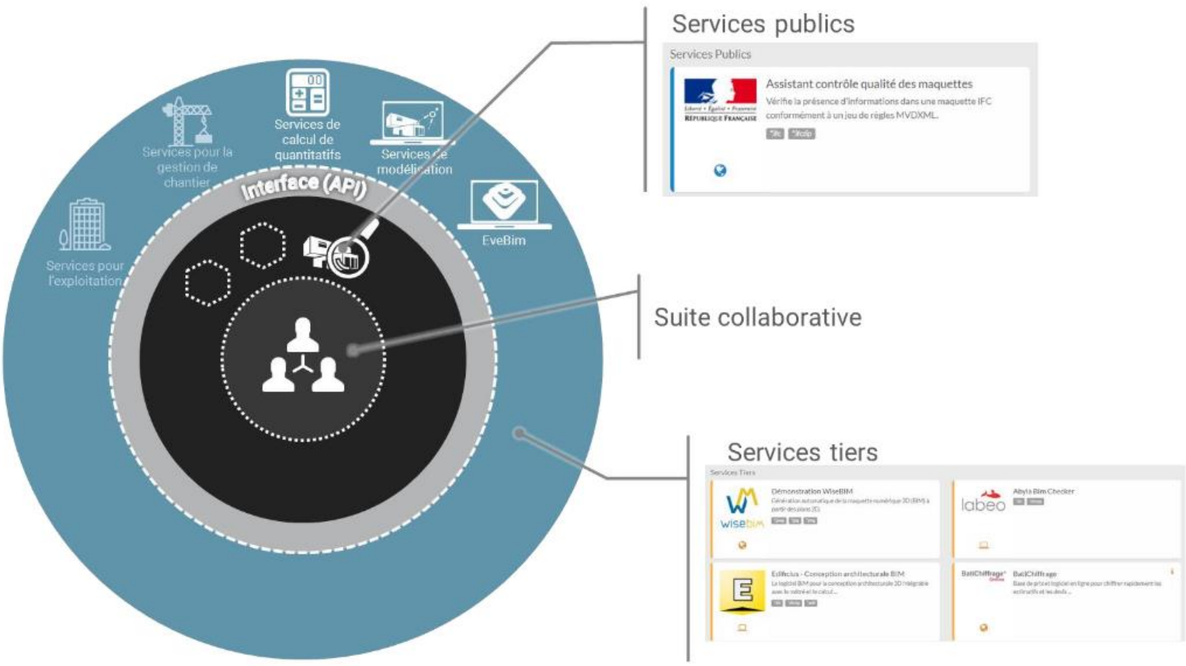


Figure 88. Synthetic schema of the “Kroqi” ecosystem (République Francaise, 2018)

This core structure can be related to external services through dedicated API (application programming interface) allowing the integration of both external services and external collaborative platforms (Figure 88). The platform born with basic instruments dedicated to the management of communication and collaboration processes such as a private chat, a basic file management system, a schedule management interface and other basic instruments. These instruments can then be combined with external services and/or applications that can communicate with the platform trough the API interface.

The dynamic structure of the platform allows a continuous development of external services. In this way, external ICT partners can integrate their products according to a service-oriented paradigm in the platform.

6.2.1.1. Frameworks comparison

It is possible to represent the core schema of the Kroqi platform as proposed in Figure 89 on the left. A frontend component represents the interface used by the stakeholders to interact with the platform. This fronted is integrated with two main layers. On the one hand, there is a layer used to manage the storage function of the platform integrating existing commercial solutions and/or a doc developed solutions with the basic services of the platform. On the other hand, there is a service layer used to provide dedicated services that can communicate with the platform through the dedicated APIs. The logical representation of the Kroqui platform can be compared to the well-known three tiers architecture (Figure 89 right). Nevertheless, the data layer has a different configuration in the Kroqui platform. Usually, in a three tiers architecture the data layer comprises database and/or data storage system such as MySQL, Oracle, MongoDB, etc. This means that the data layer is has a well-defined structure that communicates with the application layer through defined rules. In the Kroqui platform the data layer is more dynamic and can change according to needs of the project moving from databases to data storage services such as Google Drive, Dropbox and/or dedicated services developed according to the Kroqui's APIs.

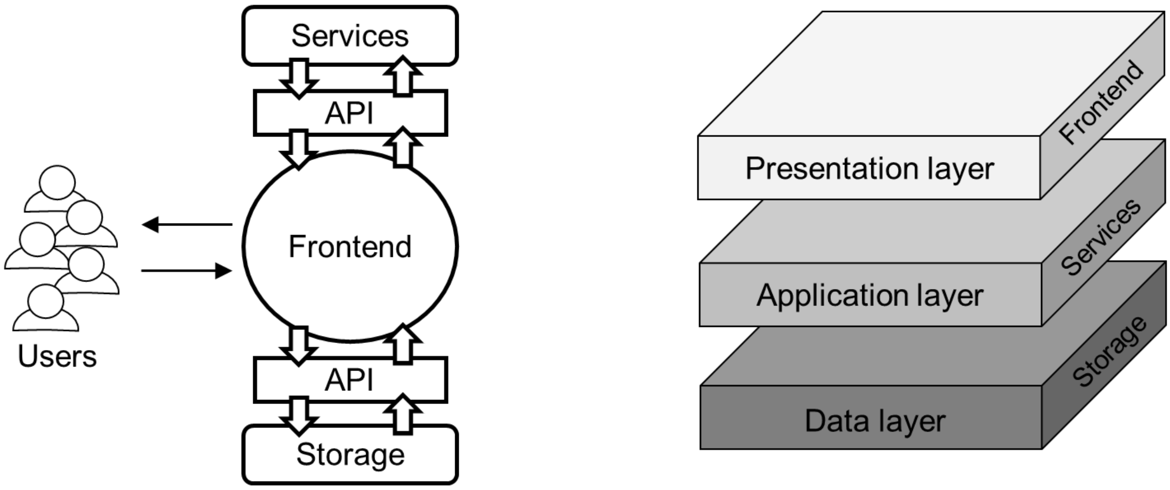


Figure 89. Logical representation of the "Kroqi" platform (left) and compared representation of the three tiers architecture

The Kroqui platform born with the objective of supporting the industry in day-to-day activities. As explained in chapter 4, this means that it must follow and guarantee times of executions aligned with the needs of the sector and provide services that are usually dedicated to a single project without a vision on the integration of different projects. The proposed framework can be integrated with this vision adding a backend environment that can collect the information provided and managed using the existing services of the platform – as demonstrated and explained in the previous chapters – allowing the development of services based on domain/sector information and needs instead of project/company needs (Figure 90). Hence, the framework proposed in chapter 4 does not intend to overcome and/or substitute the existing structures. On the contrary it should be interpreted as a support to coordinate the development of future environments and to optimise the behaviours of the existing ones. Integrating this vision with the three tiers architecture, the backend layer described in Figure 90 represents a structured data layer that can follow specific rules and models because it is not linked to a specific project environment.

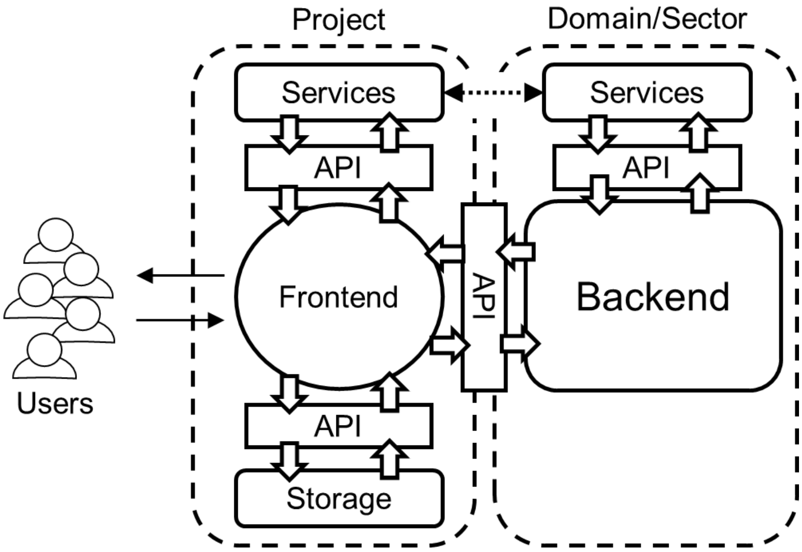


Figure 90. Integration between the Kroqi platform framework and the proposed framework

The different velocities between a collaboration platform developed for a project dimension and a platform focused on the aggregation of multiple sources and to knowledge management and analytics applications its only one of the points that validate the need of a multilevel approach. The experiments presented in chapter 5 underlined another and deeper dimension. In fact, even if the experimentations demonstrated the capability to organise and elaborate documents and information provided by information models, they highlighted the need to embed knowledge domain in the process to guarantee the demonstrated performance. However, domain knowledge is difficult to generalise. For example, the experimentation presenting a hybrid approach to organise project documents exposed in chapter 5.1 was developed according to algorithms able to work with text document in Italian language, and that were developed following the Italian practices in the construction sector. The application of the same algorithms trying to organise documents from different countries would probably result in poor performances. Hence, the need to define a different level and different interfaces dedicated to the peculiarity of a nation and/or a specific area.

6.2.2. BIM Toolkit and National BIM Library – UK

The BIM toolkit is a national platform developed in UK by NBS with the objective of supporting the private and public market in meet the requirements imposed by the Level 2 BIM (BSI, 2013). The platform comprises several services (Figure 91) including a dedicated verification tool that can check that the submitted data meets the requirements imposed through the platform standards. Moreover, according to Designing Buildings Ltd. (2018) it can be used to:

- *Define information requirements aligned to specific project stages.*
- *Assemble the project team, define deliverables and assign roles and responsibilities.*
- *Manage the delivery of information.*
- *Access free-to-use BIM objects and manufacturer's technical literature.*
- *Follow a reference library of definition templates describing the typical level of definition for different stages of a project consistent with the unified classification system Uniclass 2015. Classification mappings are in place for mapping to NRM1 and NBS Create and may be expanded to include systems such as CESMM.*
- *Provide digital information to specifiers.*

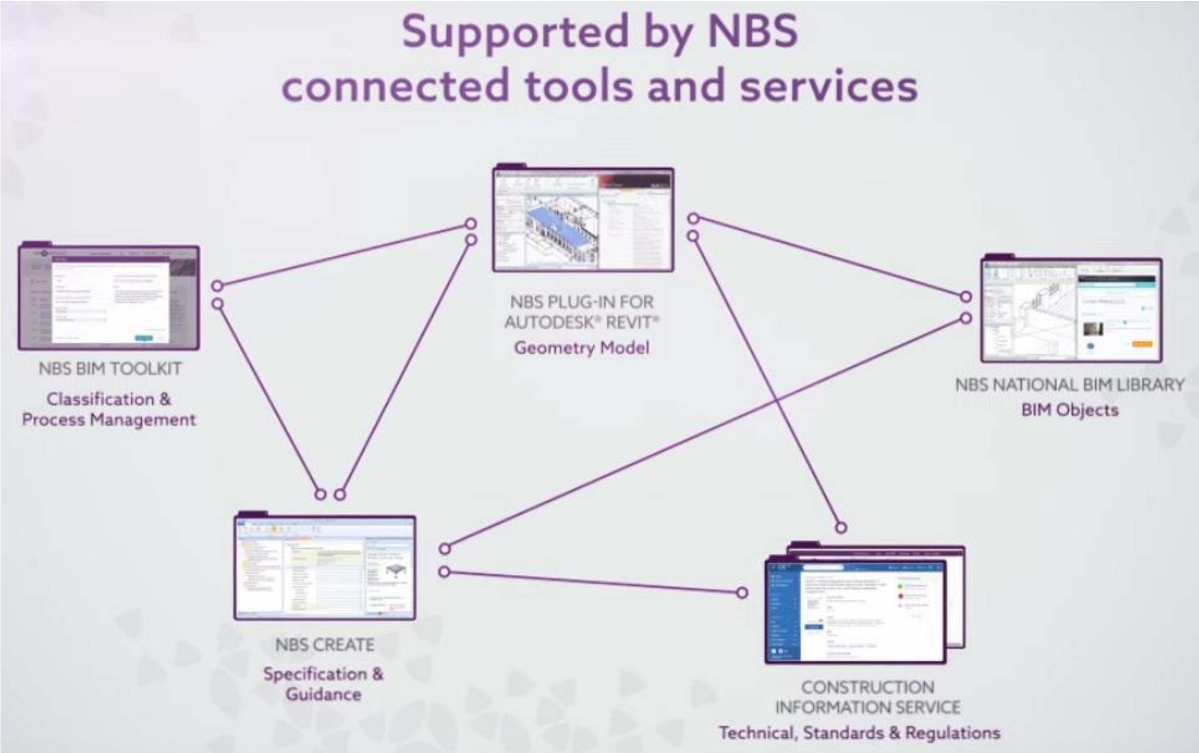


Figure 91. NBS BIM Toolkit connected tools and services (NBS, 2018)

6.2.2.1. Frameworks comparison

The NBS BIM toolkit provides services that are based on a set of structured and standardised information. It does not provide collaboration means in terms of places but in terms of standards that can help in the definition of collaboration processes. Relating this structure to the proposed framework, the NBS BIM toolkit can be interpreted as a source of rules and knowledge that can be embedded in the platform level enriching contents on both side from the BIM toolkit to the platform and from the platform to the BIM toolkit. In this case, the front end is represented by the services provided by the NBS BIM toolkit even if there is not a specific collaborative environment that should be defined by the project team. The possible link between commercial collaborative environments and organised platforms according to the proposed framework can close the link matching the information provided by the toolkit, by the collaborative environment and by the platform.

6.2.3. INNOVance – Italy

In Italy, the INNOVance research project developed between 2007 and 2014 broadly explored the development of a national digital platform for the construction industry. The collaboration between several stakeholders through the Italian standardisation organisation and the involvement of different teams in the INNOVance project resulted in the definition of criteria for the unification of terminology, organisation, collection and exchange of information for the AEC sector.

Two main results have been achieved: an unambiguous classification system and models for performance-based computational digital technical datasheets. Both the classification system and the models for technical datasheets have been proposed for different technical solutions adopted in the building process.

The criteria identified for the definition of an unambiguous classification system allow to identify families of objects considering different aspects. Namely, these aspects are: category, typology, reference standard or function, main performance, geometry, dimensions and physical-chemical characteristics. By selecting a specific choice for each field, a unique code is created for each object.

Furthermore, models have been defined for collecting, organising and archiving technical information about construction products and technological solutions. Particularly, standard criteria have been identified to describe construction products, in identificative, qualitative and quantitative terms. The structure has been defined in accordance with harmonised standards for CE marked products or in agreement with other relevant reference standards (if available and/or applicable) for non-CE marked products. Once defined the models for construction products, a comparable structure has been developed also for technological solutions, providing datasheets for layers and technological systems.

The models for the collection of technical information have been organised into informative blocks of homogeneous data. The structure collects information into classes and provides guidance on how to fill datasheets, standardising the process of description and characterisation of those ones.

The information blocks, differentiated for CE and non-CE marked products, are related to the following points: identifying manufacturer information; identifying product information; technical information; information about packaging, movement, storage in factory and transport; commercial information; additional technical information; supplementary documents; attachments; information on data reliability.

The informative blocks named "supplementary documents" can contain additional information about laying/installation, maintenance and disposal of products and about the main components of these ones, respectively collected in: dossier - guidebook about laying/installation, maintenance and disposal; datasheet of the product components.

Beyond the definition and development of the classification system and computational technical datasheets, a web-portal has been created for the fruition of information. The web-portal was composed of four main sections for:

- the collection of information related to technological systems, MEP objects and construction works;
- the fruition of information related to technological systems, MEP objects and construction works;
- the collection of BIM models linked to the standardised structure of the database;
- the collaboration among several stakeholders in a BIM Server.

Particularly, a link among information collected within the standardised structure and BIM models have been established.

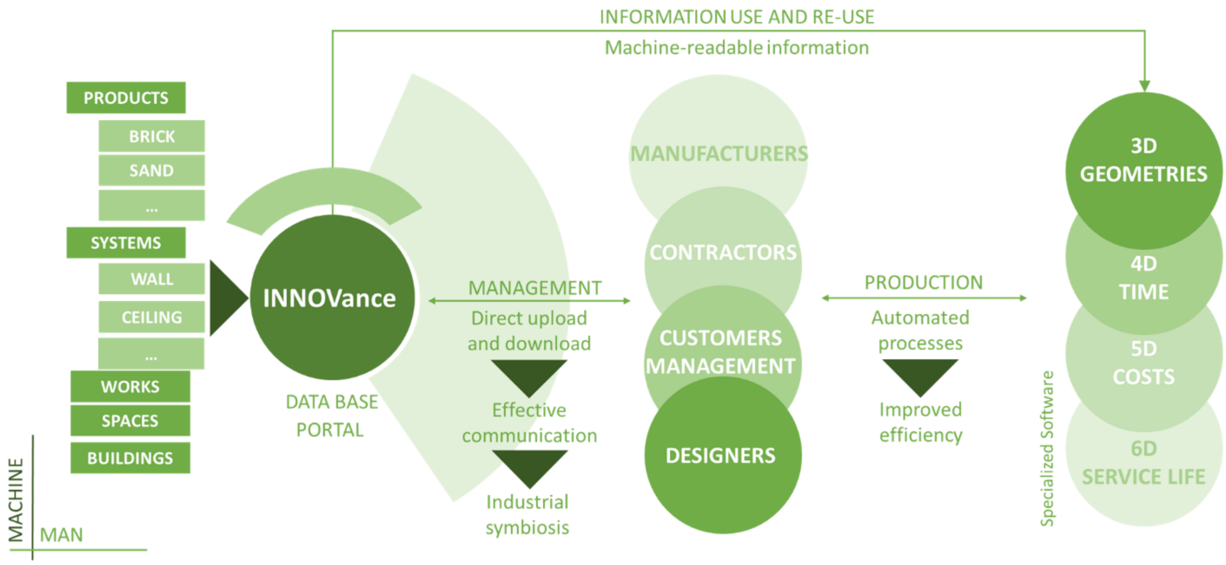


Figure 92. Overall schema of the INNOVance project

INNOVance, as an experimental research project, does not consist in an available commercial platform usable in the sector. Nevertheless, the knowledge produced during the 7 years period of development both in terms of technological and social issues, can contribute in the identification of possible future directions in the development and/or coordination of platforms for the construction sector.

6.2.3.1. Frameworks comparison

In contrast to the Kroqi platform that adopts a bottom up approach providing a dynamic environment that allows the sector (including the IT sector) to develop the required solutions addressing the needs of the market, the INNOVance project is mainly based on a top down approach defining precise rules and information structure. Nevertheless, the information structure defined through the INNOVance project paves the way for the development of specialised services developed by stakeholders even if the direct users are bounded by the rigid structure defined by the platform. According to the double dimensions of the INNOVance platform it can be related to the proposed framework from both the perspectives above described for the Kroqi platform and for the NBS BIM Toolkit. In fact, the BIM server, defined as a collaborative environment to upload and share building information models can be associated to the principles exposed for the Kroqi platforms that see the INNOVance platform as the front end for the user and the high level of the framework as the beck end. On the other side, the BIM library representing a rich source of information and knowledge can be directly embedded in the high level of the framework (as proposed for the NBS BIM toolkit) providing direct exchange between the two levels.

6.3. Relation with the European context – The DigiPLACE project vision

6.3.1. Introduction

The European construction represents 9% of EU GDP and employs 18 million people.

This sector is at the centre of a tough but also promising set of economic, environmental and societal challenges. Governments and owners must therefore deal with climate change, resource efficiency, greater demands on social care, urbanisation and immigration, ageing infrastructure, as well as constrained budgets as key aspects of their policy and projects.

Construction is a driver for economic growth and home to 3 million companies, most of which are SMEs. An innovative, competitive and growing construction sector is a crucial component for tackling these challenges.

Similarly to other sectors before, construction is now seeing its own “digital Revolution”, having previously benefitted from only modest productivity improvements. Digital tools and Building Information Modelling (“BIM”) is being progressively adopted by different parts of the value chain as strategic processes and tools to deliver cost savings, productivity and operations efficiencies, improved infrastructure quality and better environmental performance.

As identified by the works conducted by the EU BIM Task Group and explained in the published handbook, one key factor to enable digitalisation of the construction is to build a collaborative framework. Recommendations outline the need for vendor-neutral data exchange, object-oriented organisation of information, common data environment, and the assignment of responsibilities for data and information management.

Europe has therefore the opportunity to collectively encourage the development of collaborative information sharing standards practices for use in domestic or international markets. This ensures open competition in the supply chain and the open sharing of information across software platforms.

HORIZON 2020 WORK PROGRAMME 2018-2020 aims to scale and foster the emergence of user driven innovative solutions, products and services cutting across sectorial silos. By reaching out beyond EU Programmes towards MSs and Regions, it will increase political visibility and critical mass while offering better alignment of research and innovation agendas and develop synergies. Through this programme, it enables all sectors and application areas to adapt, transform and benefit from digitisation and enabling technologies, notably by allowing also smaller and newer players to capture value. It also allows developing industrial strategies, including new business models, and leverage this major transformation to increase the competitiveness of EU industries and create new markets. Last, it can remove barriers for innovation enabled by digitisation, by addressing issues such as up- and cross-skilling, harmonising regulatory frameworks and standardisation. Overall, the final Horizon 2020 work programme has the potential to make a real and sustainable difference to the quality of life in the EU, as well as the EU's position in the world, towards implementation of the Sustainable Development

Through the Digitising European Industry initiative, the European commission will contribute to the launch of a set of initiatives supporting the building of the digital platforms of the future in several application areas. This action managed by DG Connect aims to establish next generation digital platforms and re-build the underlying digital supply chain on which all economic sectors are dependent. The initiative should enable all sectors and application areas, including construction, to adapt, transform and benefit from digitisation, notably by allowing also smaller players to capture value.

In this context, DG GROW supports the vision of the construction value chain working together with the European Commission towards the common goal of developing a digital platform for the construction sector.

This initiative can create a digital ecosystem dedicated to construction connected with national initiatives start-ups and software vendors towards an open and neutral construction marketplace for the whole lifecycle of a building.

In addition, since one of the goals of an European platform would be to promote BIM initiatives all along the whole chain, and for different type of projects, examples that cover different areas should be included (buildings, highways, railways, airports, seaports, etc.), from both public and private initiatives. This allows giving a wide perspective of how different companies can use this approach to get tangible benefits. It promotes the capabilities of BIM to the AEC companies that are embracing BIM methodologies, this in turn improves their image to potential clients and owners of public assets; and it is useful for attracting talent in the form of new potential employees.

As identified by the European Commissions and some of its member states individually, Digital Platform gathering these requirements is a key factor in the on-going 4th industrial revolution.

6.3.2. Industry 4.0

The paradigm of the 4^{th} industrial revolution (industry 4.0) focuses on the data management and on the interconnection between machines, objects, persons and processes.

The keywords on this undergoing revolution are:

1. Information in real environments: AR (Augmented Reality).
2. Data management: Big Data and A.I. (Artificial Intelligence).
3. Digital collaboration.
4. Smart objects: IoT (Internet of Things).
5. Additive manufacturing: 3D printing.

Figure 93. Imagining construction's digital future – McKinsey & Company

Compared to the past industrial revolutions (18th century: mechanical loom vs iron and steel; 19th century: production line vs reinforced concrete; 20th century: automation manufacturing vs precast concrete) the construction sector is nowadays completely involved in the transition defined by the industry 4.0 paradigm, in parallel with any other industrial sector and/or services sector.

The new approach to production and use of products involves objects, subjects and processes, all integrated through the generation of common information that continually evolves.

This information needs to be collected (in a structured way), processed (singularly and in a statistical form), and redistributed (in an open and transparent form).

One example that involves the quality of buildings and infrastructure and the competitiveness of the sector is related to the impact of the regulation n. 305/2011 that imposes to all the manufacturer the declaration of technical characteristics for construction products before their marketing.

This product information must be made available in an open and standard form (Reg. 1025/2018, c. 3), guaranteeing the data transparency for the public and private sectors.

In the construction sector, the main informative means are historically the drawing and the documents related to it. In the first technological passage, from the drawing table to the CAD (Computer Aided Design) technologies, there has been a small change on these principles/processes. The use of CAD instruments (vector software), in the same way of the drawing table, involves the management of vectors and text used to represent concepts. Through this approach, the interpretation happens between technicians using analogical (static) instruments that are avulse from the process.

6.3.3. BIM and object oriented programming

The introduction of BIM (Building Information Modelling/Management) in the construction sector and the introduction of object programming in the modelling software for architecture and engineering changes this picture. The new paradigm does not focus on vectors, terms and files but it deals with the data management: geometric and non-geometric data (alphanumeric and multimedia), through the direct involvement of the machine. The computer becomes a “dynamic” instrument of interaction able to reprocess the information and to support all the subjects involved (technician and non-technician).

In the transition from CAD (i.e. means to “represent“ through terms and vectors; typical instrument of the design phase) to BIM (i.e. means to “simulate” through geometric and informative data; incremental instrument dedicated to all the phases of the process, from strategy to operation) it is critical to create a dedicated digital environment where all the information that are generated in the process can be collected and managed: the Common Data Environment.

In a first interpretation, this task was entrusted to the 3D parametric model generated using the BIM authoring tools (e.g. Revit, Allplan, Archicad, Microstation, and Edificius). This model can be identified as *“graphic model”* where are included the geometric objects and the information collected inside of the modelling tool.

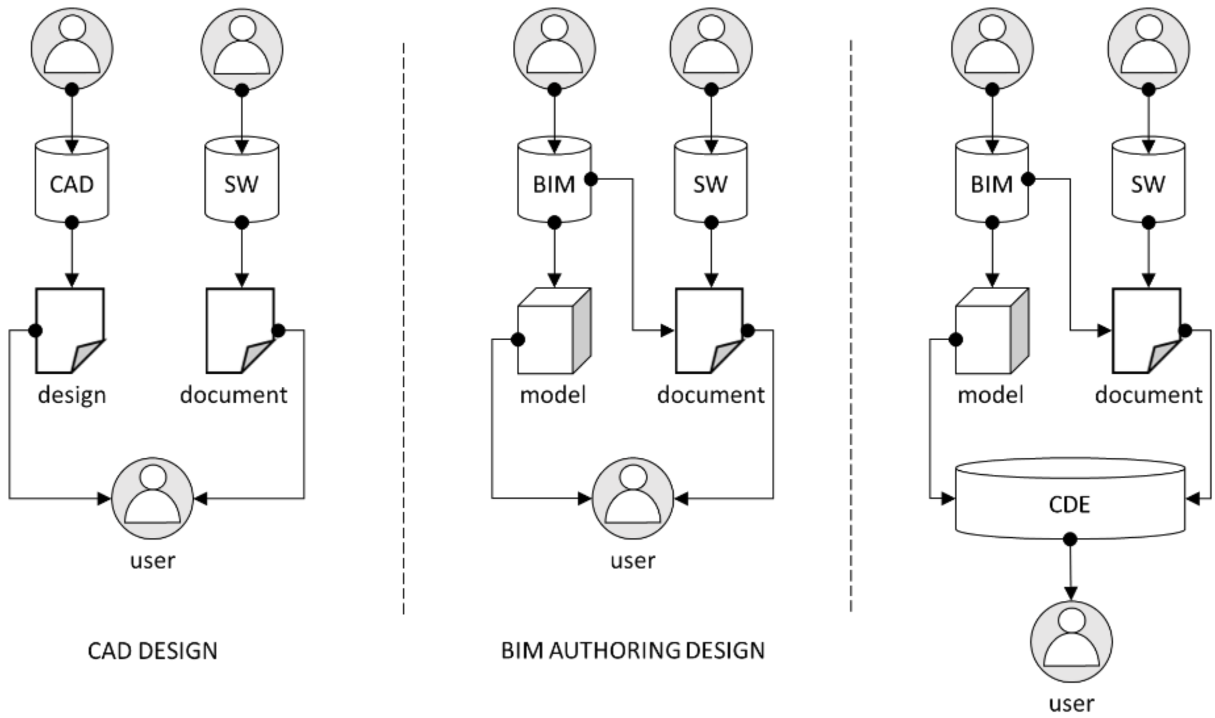


Figure 94. Traditional CAD Design; actual BIM Authoring Design; BIM (method) process

However, this approach does not provide critical advantages compared to the traditional process schema based on analogical informative means. The simple introduction of graphical models generated using BIM authoring tools (object-oriented programming) in the design, compared to the paper-based processes of CAD (vector programming) do not introduce an effective involvement of the entire chain with a consequent inclusion of all the process phases: strategy production and operation. The graphical model on its own does not include (and cannot contemplate) the information of the entire chain that still remains bound to the traditional documents related to the drawings.

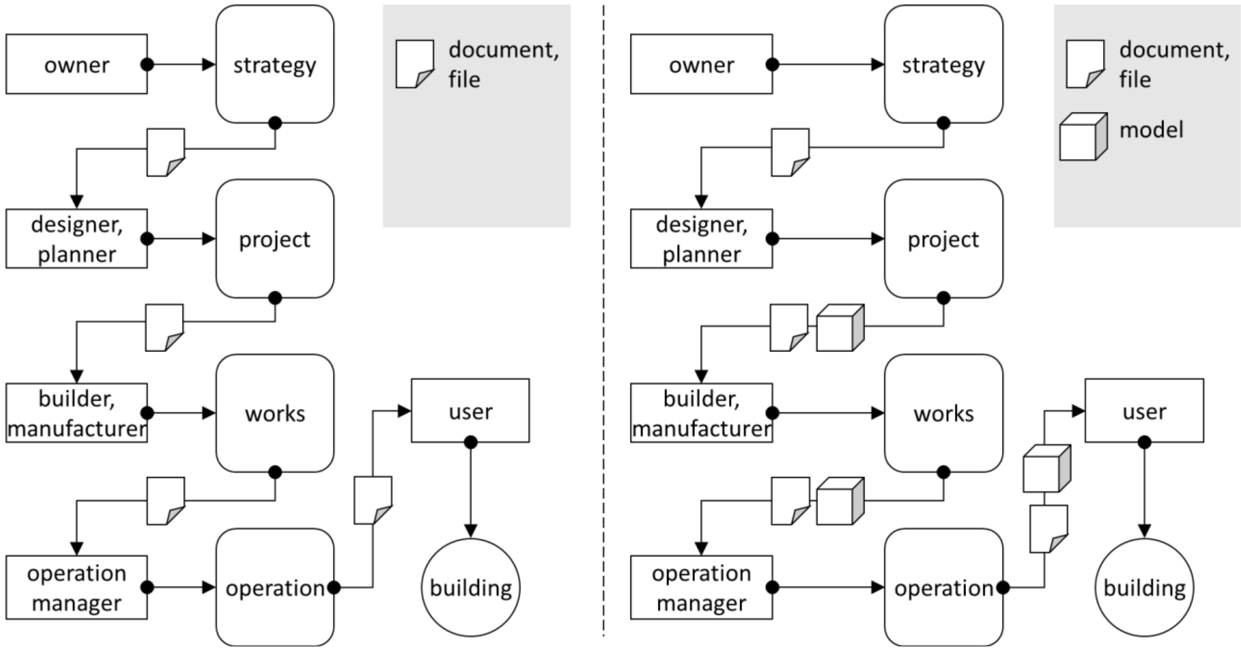


Figure 95. Paper-based or digital CAD process (on the left) and digital BIM process, use of the graphical model alone (on the right)

6.3.4. Common Data Environment

To solve this problem, nowadays it is consolidated the need of creating in addition to the graphical models, specific Common Data Environments (CDE; PAS 1192-2:2013; UNI 11337:2017) to support the entire informative process (collection, reprocessing and management of the data). The construction process is composed by an high number of documents in addition to the graphical models (design) such as reports, calculations, proceedings, etc. These documents support and integrate the design. With this data flow system based on the CDE, it is possible effectively involve every stakeholder (owner, designer and planner, manufacturer and builder, operation manager, users) and every phase of the construction process (strategy, project, work, operation).

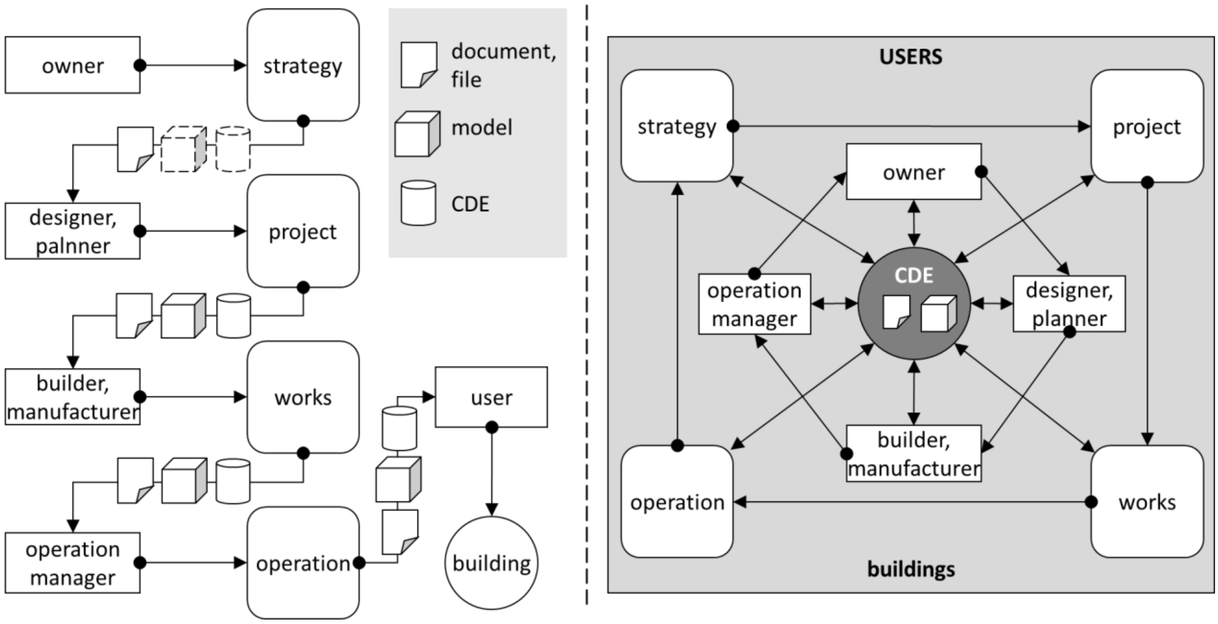


Figure 96. BIM process with CDE on each phase (on the left), BIM process with centralised CDE (on the right)

In these environments of digital sharing (CDE) it is possible to manage in an integrated way, files, models (graphics) and documents, referred to a project, an intervention or an Asset, during the time. The use of these environments, leads to abandon the traditional process flows based on antagonistic relationships (several CDEs defined in each phase of the process – Figure 96 on the left) encouraging collaborative processes supported by a unique centralized CDE or interoperable/unified CDE, where all the actors involved can operate. Furthermore, the centralized/unified CDE can be related to the entire life cycle of the process, from the strategy to the deconstruction and reuse of the land (Figure 96 – right side).

Solution benchmarking shows that CDE have not overcome the challenge of answering to the diversity of construction sector expectations or including highly ergonomic collaboration-based interfaces. In the same time, more and more companies are therefore trying to tackle this challenge by using free or non-professional platforms, which can put data at risk while not answering fully to CDE principles.

6.3.5. Digital platform

In the era of data management (Big Data and Smart cities) an instrument to process sharing that is bounded to the limit of a singular work or that is related to a singular asset (building or infrastructure) like the CDE, can be viewed as a limitation. It needs to be integrated into a coordination system defined at a higher level to unlock the potentiality introduced with the BIM methodology and the related instruments.

For this reason it is now spreading the term data management “platform” (UNI 11337:2017), able to overcome actual limits of the CDE (focused on a specific project and/or work), as places to collect, manage and organise all the information provided from objects, models and from the different Common Data Environments.

Through the concept of platform, big and small properties (public and private), manager, contracting stations, etc. can deal with the integrated management of their real estate portfolio instead of focusing on a single asset and/or on a specific work. In the same way, the credit sector, the insurances, the investors, the production (manufactures), etc. can extract, from the sector platforms and the consequent organisation of the information from the entire chain, specific information and statistical information to develop new strategies, optimise the processes, etc.

According to this trend, starting from 2013 have been developed several BIM-based platforms, defined at national level. Each one with different specifications but with shared objectives, i.e. manage and coordinate data provided by the construction chain overcoming the limits of the single asset and/or of the single work. Moreover, several private and public initiatives have been devoted to the development of specific solutions for the above-mentioned issues. However, the majority of these projects are still under development and there is not an existing specific solution able to address the highlighted issues.

Table 11. National and European digital platforms

Date	Name	Nationality	Description
2013	INNOVance	Italy	BIM Library for manufacturer objects, standard attributes; BIM Server for models and projects, tenant management; BIM-GIS platform for buildings, infrastructures and site.
2015	BIMToolkit NBL Library	United Kindom	BIM Library; Standard definition of geometric an informative attributes (LOD-LOI); valuation Toolkit of objects LOD and LOI level in a model.
2017	Kroqi	France	CDE Server, project oriented; Basic integrated services and open API to allow the development of integrated and structure services/applications from private and public bodies.

The focus of these national platforms is to promote the development and the diffusion of the digitalisation in the construction sector, particularly for SMEs

On the one hand, these platforms aim to centralise the rules and a portion of the costs related to the structure and to the investments in the public sector. On the other hand, the platforms promote the use of open formats, stimulating the competition between software houses, and consequently reducing the creation of leading positions or monopoly position in the software market to avoid what already happened with CAD software.

6.3.6. European digital platform

Nowadays it is clear the need for the market to have platforms defined at a higher level in comparison to the concept of CDE (focused on single goods or works) and to have national platforms that are above both. Hence, it is clear that in a supranational common market – EU – can be identified the same issues related to the collection, sharing, transparency and management of data provided by lower levels. These issues highlight possible benefits in the development of a platform environment above every single national and local market, able to promote the interaction between the operators and that can facilitate the information flow beyond the state borders.

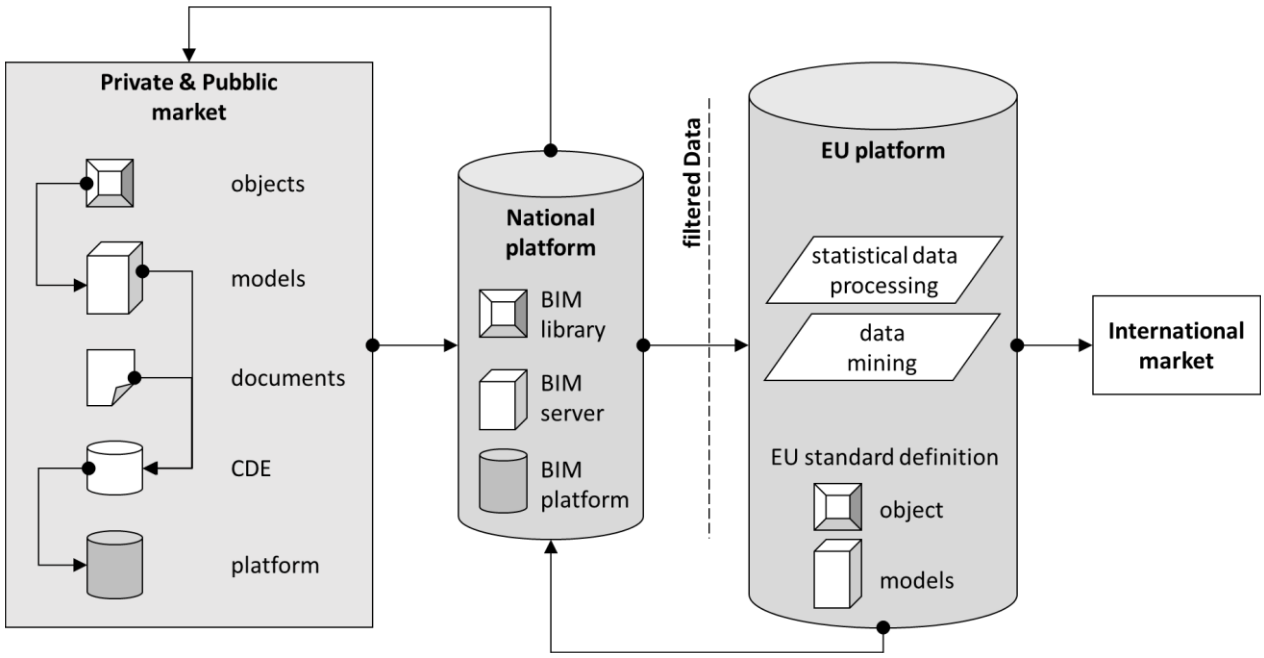


Figure 97. Objects, models, documents, CDE, platforms and Data flow in private and public markets to national and European digital platforms

A European Digital Platform for the construction sector can deal with the development and the competitiveness in internal markets (free circulation of information and knowledge in the European area). Moreover, it can represent a crucial point to improve the competitiveness of the European companies on international markets (systemisation of the information for the European area), with specific reference to the emerging markets like Asia and Africa. Hence, this platform can be used as a means to compete as European system against the international big players like the USA, China, Japan, Canada, Australia and Russia.

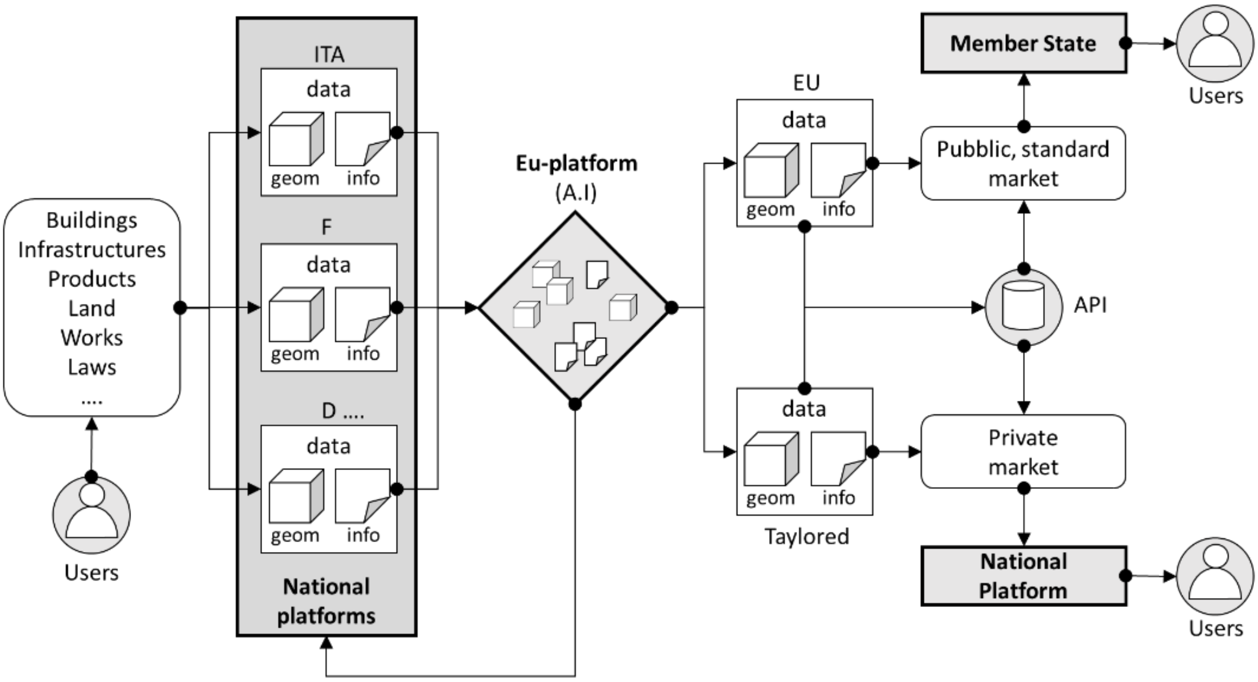


Figure 98. European Digital Construction Platform

A digital platform defined at European level would allow a potential alignment of the different national platforms to access high-level data while connecting to other potential local and/or national market. This would result in a direct improvement in the actual performance and in the possible services based on the platforms.

On the implementation side, the defined environment would encourage the development of collaborative information sharing standards practices for use in domestic or international markets. This ensures open competition in the supply chain and the open sharing of information across software platforms.

This initiative can create a digital ecosystem dedicated to construction connected with national initiatives start-ups and software vendors towards an open and neutral construction marketplace for the whole lifecycle of a building. This would also allow to promote BIM initiatives (public or private) all along the value chain, and therefore give a mass access to information.

On a technical side, it would allow the definition of a data mining function that nowadays represents a critical point for the development of a standardized language for the construction sector in the European area. Moreover, it would allow the definition of unified informative structure for buildings and for construction products (digital CE marking of products; Regulation 350/2011; Afnor PR XP P07-150; UNI 11337-3:2016). The platform would allow the production of new knowledge useful for the development of new voluntary technical standards for CEN and from this last for ISO providing a European contribution instead of a multitude of national contributions. An improvement of the standards can boost the markets development and increase the quality of the products as reported in the Regulation 1025/2012 *"European standardisation also helps to boost the competitiveness of enterprises by facilitating in particular the free movement of goods and services, network interoperability, means of communication, technological development and innovation…"*.

6.3.7. Conclusions

The digitalization of the construction sector allows the construction chain to enter in the industry 4.0 paradigm as all the other industrial and/or services sectors.

The characteristic aspects of this fourth industrial revolution is the management and the valorisation of data. In the construction, the digital management of the information is linked to BIM objects and BIM models, to Common Data Environment based on the process and to the digital platform defined in the sector and at national level.

The construction of a supranational platform framework would allow to the construction to overcome actual barriers of languages, knowledge, regulations (nationally and locally) allowing the definition of common data structures that can be standardised for the construction products (CPR 305/2011).

The project to be developed would not only focus on the possibility to define a vendor neutral data exchange environment and linked technologies and formats but also on their management practices (collaboration, safety, workflows, versions,…) and uses (Construction Permit, Code Checking,…).

As part of the key objectives that the resulting platforms can achieve, we can already identify :

- Give a mass access to technologies and information allowing all construction actors to experiment collaboration before full implementation
- Allow SMEs to develop digital solutions by themselves and for themselves inducing a positive context to stimulate digital solution offer dedicated to construction issues.
- Promote open formats / standards while limiting native systems monopoles
- Create focal pointsfor all EU standardization elements and initiatives (Platforms, Data Dictionary, Process, open DATA…)
- Bring together clients, construction value-chain, public services and software vendors

A market platform at European level would provide to the Member States an open and controlled infrastructure that can represent a fundamental data source also for the public market. This platform would improve the functionalities of national platforms and of the commercial solutions developed by the single software houses providing the access to data collected at European level without the need of paying for their collection, maintenance and updating. Moreover, the European platform would stimulate the Member States that are still lacking in the development of a national platform to develop it, accelerating the digitalization process of the construction sector in the entire European area.

Base on this vision, we believe that the Digital Construction Platform initiative can greatly contribute to develop an innovative, competitive and growing construction sector while improving sector competitiveness and export capability.

7. Conclusion, Discussion and Outlook

Starting from the research questions presented in the introduction it is possible summarising the main results obtained during the PhD path and presented in this thesis. The results here reported must be related to the research boundaries explained in the identification of the research questions and clarified during the development of the thesis.

- How can we collect, share and generate knowledge with value on the construction sector domain?

This question opens the way to the study of social implications in the introduction of processes supported by digital technologies in the construction sector. This dimension was explored including different disciplines proposing a theoretical framework able to represent both the organisational structure in construction processes and the spatio-temporal fragmentation of these last (chapter 2.1). Figure 11 and Figure 12 represent these two perceptions defining the answer to this first question. The fragmented business configuration that characterises the project and the distribution in terms of both space and time, indicate the need to manage knowledge integrating sources from multiple projects. Hence, to collect, share and generate knowledge with value on the construction sector domain there is the need to create an environment (knowledge base) that should be neutral with reference to the project needs providing a way to limit the disruption of the knowledge management process at the end of each project. The answer to this first question highlights the value of the next one that deals with the technical and societal possibilities of creating such environment.

- How can we collect and parse data and information provided by different sources along the construction process, addressing its spatiotemporal fragmentation and its long time distribution?

Starting from the theoretical study related to the area of knowledge that address the first research question, this research identifies in a multi-level approach to collaborative environments and platforms the direction to embrace the complexity of the construction sector without imposing limits to the day-to-day work. The research presents a novel framework for the development of such collaborative environments identifying these structures as platforms in the extents already clarified in chapter 1.4.2. In detail, to collect and parse data and information from different sources (that may be provided by different projects), this research highlights the need to introduce a “translation layers” that embeds AI processes to translate unstructured and/or not standard information in a collection well-defined data structure. Thus, the answer to this question is twofold. On the one hand, there is the need to adopt a multi-level approach “separating” the knowledge based layer to the project layers. On the other hand, there is the need to define automated process able to translate data and information from the project layers to the knowledge based one. Both these aspects have been explored and addressed in the thesis (chapters 4 and 5)

- How can we guide the construction sector in provide such data and information?

Chapter 6.1 proposes an evolutionary vision on collaborative environments identifying a possible future change in the perception on these lasts. In particular, the change in the perception from the collaboration itself to the use of services based on the optimal organisation and use of data and information can change the willingness to share data and information to obtain better services and consequently obtain an increased value. Thus, to guide the construction sector in providing data and information, this research identifies the development of services as a critical passage. The development of services does not represent the focus on this work and has not been addressed during the research. Nevertheless, the framework presented in chapter 4 as well as the applications developed in chapter 5 pave the way to the development of services with value in the sector/domain level.

- How can we maximise the value of data and information collected and how can we use these assets in the generation of knowledge?

The analysis proposed in chapter 3 highlighted the need to change the approach in the use of data and information generated during the construction process according to their poor quality when analysed for analytics purposes. On the other hand, the organisation of the information and of the sources of information (such as text files, images, etc.) is critical to maximise the value of mining and analysis activities. Hence, the experimentation presented in chapter 5 demonstrate a possible path to follow in order to change the perspective on data quality in the construction process with the scope of maximising the value of data and information through the application of analytics processes that can generate novel knowledge re-usable in the sector and company domains. Thus, this research identifies in the a priori optimisation of data quality and data organisation a possible way to maximise the value of data and information collected from the project sources. The development of experimental applications devoted to the use of data and information assets in the generation of knowledge is out of the scope in this work. Nevertheless, the literature review presented in chapter 2.2 highlights the possibilities in the use of data and information to generate knowledge.

7.1. Outcomes and original contributions

The main original contributions (part of which have been published by me in collaboration with my colleagues in international conferences and journals (chapter 7.3), but which never before appeared in mainstream literature as of our knowledge) can be identified as follows:

- Development of an theoretical schema integrating the hypertext organisation schema by Nonaka and Takeuchi (1995), the cultural historical activity theory and the heuristic decision theory. This schema has been integrated with a spatio-temporal perspective on knowledge management allowing a holistic vision comprising business organisation, time distribution and spatial distribution of knowledge creation and consumption activities in the construction sector (chapter 2, Figure 11 and Figure 12).
- Development of a multi-level framework devoted to the definition of a construction digital platform to pave the way for knowledge generation and management activities starting from data and information collection and analysis. The framework comprises not only the platform itself but also the interrelations with collaborative environments according to the multi-level criteria (chapter 4, Figure 55 and Figure 58).
- Development of specific and novel components of the framework namely:
    - A hybrid classification and clustering approach to organise heterogeneous datasets of digital documents paving the way for the development of dedicated applications and services (chapter 5.1).
    - A machine learning image recognition process devoted to semantic enrichment of building information models with the objective of support the translation activities between the different levels identified in the framework (chapter 5.2).
- Definition of an evolutionary perspective on collaborative environments in the construction sector (chapter 6, Figure 87).

To summarise and link the output presented in chapters 2 and 4, Figure 99 reports a compared vision of the integrated hypertext organisation schema (Figure 11), the high level architecture of the multi-level system (Figure 53) and the system framework (Figure 58).

This representation can be linked to the case studies presented in chapter 5 according to the schema reported in Figure 60. The two components of the framework should be integrated in the “translation layers” providing an automated means to reorganise the data from the project level to the knowledge management and analytics one, i.e. to the knowledge based platform.

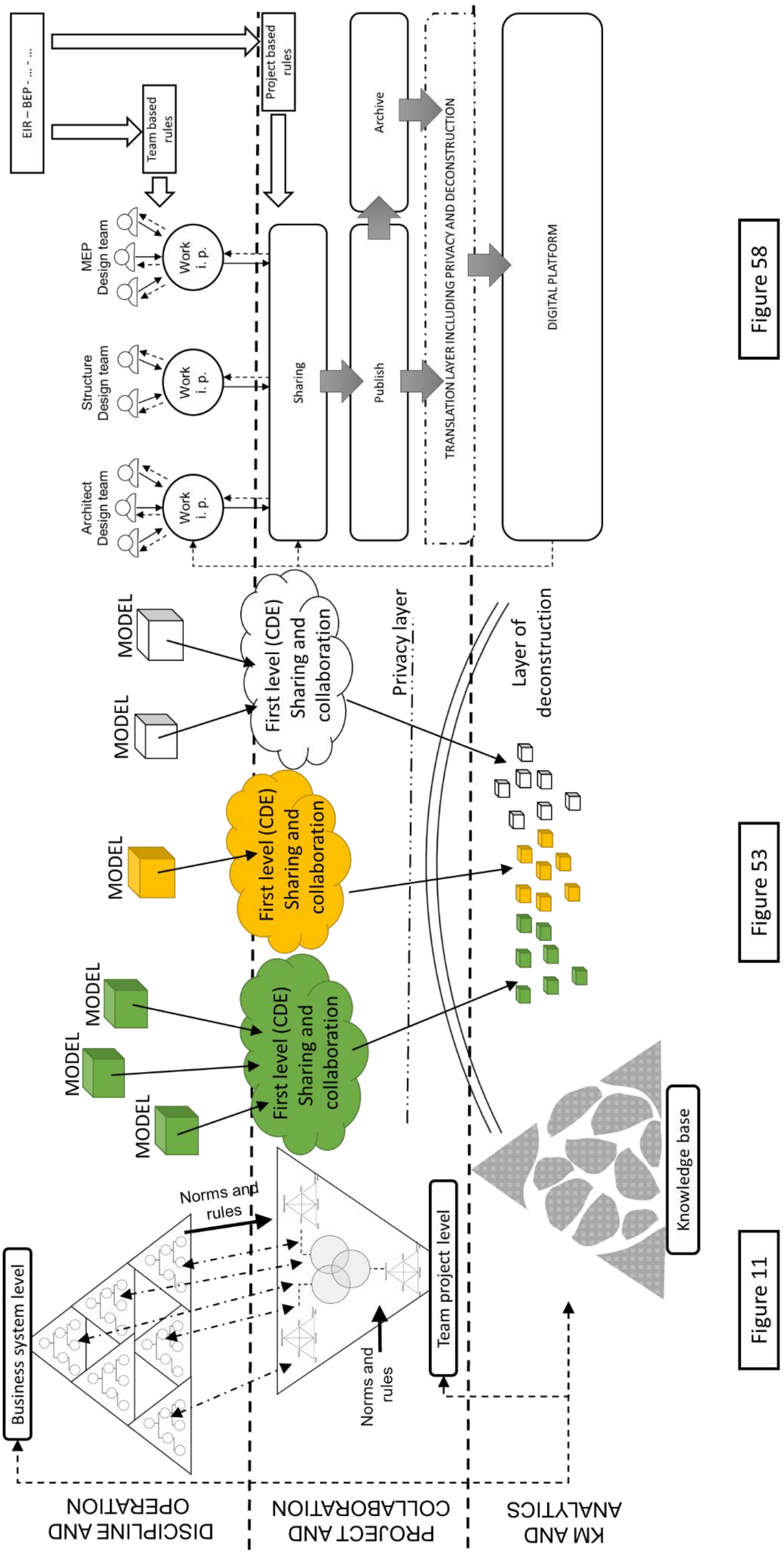


Figure 99. Integration of the theoretical framework presented in Figure 11, the high level architecture of Figure 53 and the system framework of Figure 58

7.2. Drawbacks, open issues and future works

While the results exposed in this thesis can help in defining the path for the construction of digital platforms in the construction sector, they underline new open questions that can be identified as future lines of research.

The experimentations proposed in chapter 5 and the evidences from the literature cited in chapter 4 demonstrated the capability of the techniques in addressing the complexity behind the need of aggregating data and information provided by distributed projects. However, both the developed experimentations and the studies in the literature are still limited to a small area compared to the articulated structure of data and information generated in the construction sector. Hence, future works need to explore deeper the possible technical applications related to the proposed framework. Moreover, there is the need of collecting data from the industry according to the proposed approach to understand the effective application of KDD processes as already started in some studies (e.g. Petrova *et al.*, 2018).

This thesis does not explore the business model related to the proposed framework. Hence, costs of implementation and management must be studied and evaluated according to possible business models associated to the platform. The reason is that the proposed framework does not define a specific technology related to the data structure and/or the connected infrastructure. Hence, according to the different configurations (e.g of use semantic web, use of graph databases, etc.) the costs, the management activities and the related business models can change substantially. Nevertheless, the issues related to the implementation of platforms according to the proposed framework can be related to a political drawback that is the property and management of the data treated by and/or through the platform framework. In fact, the proposed result already includes the need to identify a privacy layer to filter the sensible information and protect the final user. However, as demonstrated by the recent events (e.g. the case of Facebook) and by the increasing requirements from the EU on privacy (EU, 2016e), this theme is all but simple and need a specific and depth study including legal, political, and economic evaluations.

The examples proposed in chapter 5 highlight a critical point that consolidate the need of a multi-level approach in the definition of collaborative and knowledge based platforms. In fact, e.g. the classification of documents is bounded to a specific language and a specific culture where the documents can be classified and clustered. However, even if the approach can be generalised, it is not feasible to apply the same structure to a general dataset. It is required to have several datasets, created in a homogeneous social context (for example regional or national) and starting from this dataset collect the information in a higher level and so on. Obviously, what is required is the coordination of the different levels to promote a precise and easy communication between the different levels of the platform. Thus, e.g. even if the classification is developed according to a specific context and cultural area, it can be defined according to a general structure that is provided by the higher level of the platform hierarchy.

At the level of development reached in this research, the algorithms proposed to introduce new information and to overcome data isolation can only exist in the higher level of the structure where the data are analysed and treated in a statistical way. Nevertheless, the results pave the way for the introduction of similar approaches at the lower collaboration level, with a direct integration with project under developments. In this way, the issues related to a code based interoperability can be eliminated basing not only the analysis but the entire collaboration on the intrinsic semantic content of models and documents stored in a common data environment. However, other issues have to be addressed such as legal aspects related to the intervention of the machine on the project information and the reliability of the system in imputing this information.

7.3. Dissemination, national and international impacts

During the development of the research presented in this thesis, the author worked in the development of several publications that helped in shaping and refining the results proposed. In particular, can be listed the following publications:

- Pavan, C. Mirarchi (2015), Digital information and hedonic models to improve decision making processes (BIM-GIS), in ISTeA 2016, Maggioli, 138-147.
- S. Della Torre, C. Mirarchi, A. Pavan (2017), Il BIM per la conservazione: rappresentare e gestire la conoscenza, Ananke, 82, 108-115.
- Pavan, C. Mirarchi (2017), Overcome data island in the construction sector, ISTeA 2017 Back to 4.0: rethinking – the digital construction industry, Florence, Italy.
- C. Mirarchi, D. Pasini, A. Pavan, B. Daniotti (2017), Automated IFC-based processes in the construction sector: a method for improving the information flow, in Proceedings of the joint conference on computing in construction (JC3), Heraklion, Crete, 491-498.
- C. Mirarchi (2018), A spatio-temporal perspective to knowledge management in the construction sector, New frontiers of construction management workshop, 8-9 November, Ravenna.

Moreover, the results presented in this thesis defined the basis for the development of a proposal (called DigiPLACE project, leaded by Politecnico di Milano and Coordinate by Prof. Alberto Pavan) for the European project DT-ICT-13-2019 about the development of digital platforms in the construction sector. The proposal involved 25 partners including the main research institutes at European level, the main associations related to the construction sector chain and three of the main European countries (Table 12 and Figure 100). The main vision of the project is contained in chapter 6.3.

Table 12. List of DIgiPLACE partners

Acronym	Extended name
POLIMI	Politecnico di Milano
CECE aisbl	Committee for European Construction Equipment
CSTB	Centre Scientifique et Technique du Batiment
FEDERCOSTRUZIONI	Federcostruzioni
MS France	Ministry for the Ecological and Inclusive Transition (MTES)
MS Italy	Ministero delle infrastrutture e dei trasporti (MIT)
MS Germany	German Federal Ministry of Transport and Digital Infrastructure
ECTP	European Construction, built environment and energy efficient buildings Technology Platform
BBRI	Belgian Building Research Institute
EBC	European Builders Confederation
FIEC	Federation de l'Industrie Europeenne de la Construction
ANCE	Associazione Nazionale Costruttori Edili
UL	Univerza v Ljubljani
BuildingSMART Int.	BuildingSmart International Limited
EFCA	European Federation of Engineering Consultancy Associations
ACE	Architects Council of Europe
INDRA	INDRA Sistemas s.a

ENCORD	European Network of Construction Companies for Research and Development
CPE	Construction Products Europe
TECHNALIA	Fundacion Tecnalia Research & Innovation
VTT	VTT Technical Research Centre of Finland
CU/BRE	Cardiff University/ Building Research Establishment
LIST	Luxemburg Institute of Science and Technology
NTNU	Norwegian University of Science and Technology
TNO	Netherlands Organisation for applied scientific research

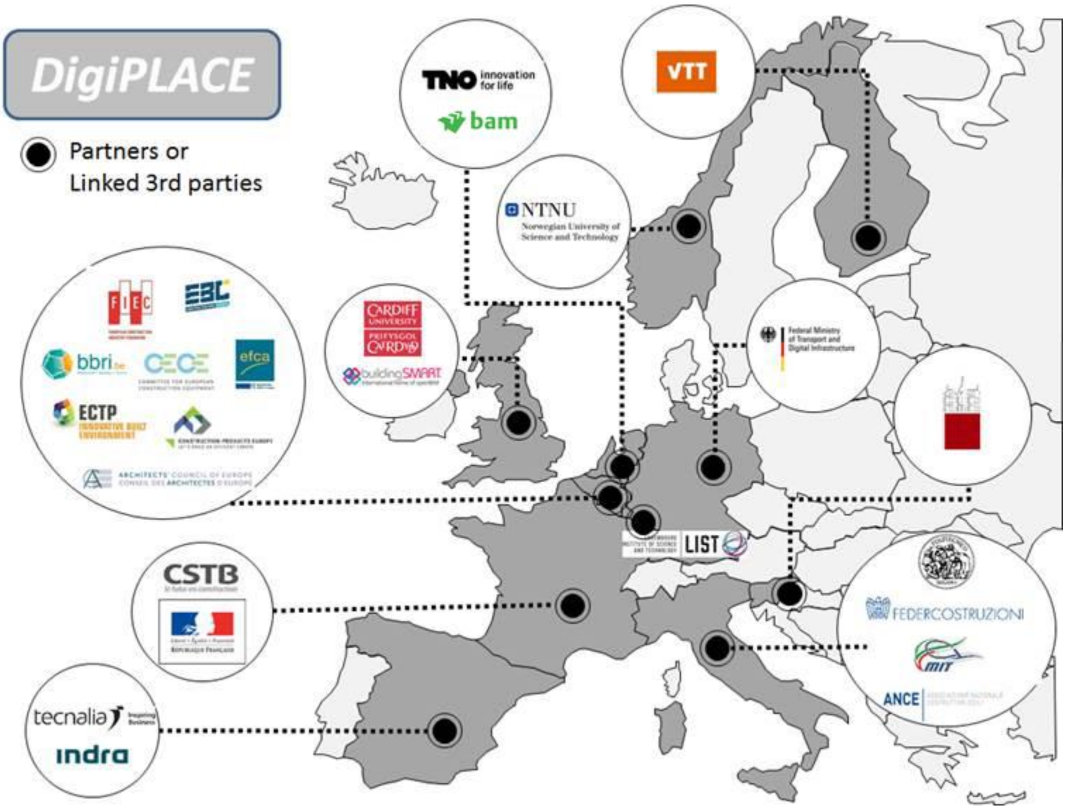


Figure 100. Partners involved in the DigiPlace proposal

8. Bibliography

Aamodt, A. (1994) ‘Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches’, *Artificial Intelligence Communication*, 7(1), pp. 39–59. doi: 10.3390/s120811154.

Acca Software (2018) *usBIM*. Available at: https://www.acca.it/bim-management-software (Accessed: 18 March 2018).

Achten, H. and Beetz, J. (2009) ‘What happened to collaborative design?’, in *Proceedings of eCAADe 2009 - Computation: The new realm of architectural design*. Istanbul, pp. 357–365.

Afsarmanesh, H., Garita, C., Hertzberg, L. O. and Santos-Silva, V. (1997) ‘Management of Distributed Information in Virtual Enterprises - the Prodnet Approach’, in *Proceedings of the 4th International Conference on Construction Engineering*. Nottingham, UK.

Alavi, M. and Leidner, D. E. (2001) ‘Knowledge management and knowledge management systems: Conceptual foundations and research issues’, *MIS Quarterly*, 25(1), pp. 107–136. doi: 10.2307/3250961.

Allplan (2018) *Nemetscheck Group - Allplan BIM Plus*. Available at: https://www.allplan.com/products/allplan-bimplus/ (Accessed: 18 March 2018).

Alreshidi, E., Mourshed, M. and Rezgui, Y. (2016) ‘Cloud-Based BIM Governance Platform Requirements and Specifications : Software Engineering Approach Using BPMN and UML’, *Journal of Computing in Civil Engineering*, 30(4). doi: 10.1061/(ASCE)CP.1943-5487.0000539.

ANCE (2017) *Osservatorio congiunturale sull’industria delle costruzioni*.

Anderberg, M. R. (1973) *Clustering Analysis for Applications*. New York: Academic Press.

Anumba, C. J. (2009) ‘Towards next-generation knowledge management systems for construction sector organisations’, *Construction Innovation*, 9(3), pp. 245–249. doi: 10.1108/MBE-09-2016-0047.

Apache (2016) *Cassandra*. Available at: http://cassandra.apache.org/ (Accessed: 9 March 2018).

Apache (2018a) *CouchDB*. Available at: http://couchdb.apache.org/ (Accessed: 9 March 2018).

Apache (2018b) *Hadoop*. Available at: http://hadoop.apache.org/ (Accessed: 23 March 2018).

Apache (2018c) *HBase*. Available at: http://hbase.apache.org/ (Accessed: 9 March 2018).

Arashpour, M., Wakefield, R., Abbasi, B., Lee, E. W. M. and Minas, J. (2016) ‘Off-site construction optimization: Sequencing multiple job classes with time constraints’, *Automation in Construction*, 71, pp. 262–270. doi: 10.1016/j.autcon.2016.08.001.

Atzeni, P., Ceri, S., Paraboschi, S. and Torlone, R. (1999) *Database Systems: Concepts, Languages and Architectures*. Maidenhead, Berkshire: McGraw-Hill.

Autodesk (2018a) *BIM 360*. Available at: https://bim360.autodesk.com/ (Accessed: 18 March 2018).

Autodesk (2018b) *Collaboration for Revit*. Available at: https://www.autodesk.com/products/collaboration-for-revit/overview (Accessed: 18 March 2018).

Babič, N. C., Podbreznik, P. and Rebolj, D. (2010) ‘Integrating resource production and construction using BIM’, *Automation in Construction*, 19, pp. 539–543. doi: 10.1016/j.autcon.2009.11.005.

Bar-Hillel, Y. (1960) ‘The present status of automatic translation of languages’, *Advances in Computers*, 1, pp. 91–163.

Barazzetti, L., Banfi, F., Brumana, R., Gusmeroli, G., Oreni, D., Previtali, M., Roncoroni, F. and Schiantarelli, G. (2015) ‘BIM from Laser Clouds and Finite Element Analysis: Combining Structural

Analysis and Geomteric Complexity', in *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XL-5/W4, 2015 3D Virtual Reconstruction and Visualization of Complex Architectures, 25-27 February 2015, Avila, Spain.*

Bazjanac, V. and Kiviniemi, A. (2007) 'Reduction, simplification, translation and interpretation in the exchange of model data', *Proceedings of the 24th Conference on Bringing ITC Knowledge to Work*, 78(Gsa 2003), pp. 163–168. doi: 10.1017/CBO9781107415324.004.

Beetz, J., van Berlo, L., de Laat, R. and van den Helm, P. (2010) 'bimserver. org–An Open Source IFC Model Server', *Proceedings of the CIB W78 2010: 27th International Conference–Cairo, Egypt*, (Weise 2006), pp. 16–18. Available at: http://bimserver.org/wp-content/uploads/2010/11/Beetz_Berlo-CIB-W78_Cairo.pdf.

Beetz, J., Van Leeuwen, J. P. and De Vries, B. (2006) 'Towards a topological reasoning service for ifcbased building information models in a semantic web context', in *Joint International Conference on Computing and Decision Making in Civil and Building Engineering*. Montréal.

Beetz, J., Leeuwen, J. Van and De Vries, B. (2009) 'IfcOWL: A case of transforming EXPRESS schemas into ontologies', *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, 23(1), pp. 89–101. doi: https://doi.org/10.1017/S0890060409000122.

Belsky, M., Sacks, R. and Brilakis, I. (2016) 'Semantic Enrichment for Building Information Modeling', *Computer-Aided Civil and Infrastructure Engineering*, 31(4), pp. 261–274. doi: 10.1111/mice.12128.

Bentley (2018) *Project Wise Overview*. Available at: https://www.bentley.com/en/products/brands/projectwise (Accessed: 18 March 2018).

Bernecker, S. and Pritchard, D. (eds) (2011) *The Routledge Companion to Epistemology*. London and New York: Routledge.

Berners-Lee, T. (2009) *Linked Data*. Available at: https://www.w3.org/DesignIssues/LinkedData.

Berry Michael, W. (2004) 'Automatic Discovery of Similar Words', in *Survey of Text Mining: Clustering, Classification and Retrieval*. New York: Springer-Verlag, pp. 24–43.

Bhatla, A., Choe, S. Y., Fierro, O. and Leite, F. (2012) 'Evaluation of accuracy of as-built 3D modeling from photos taken by handheld digital cameras', *Automation in Construction*. Elsevier B.V., 28, pp. 116–127. doi: 10.1016/j.autcon.2012.06.003.

Bicevska, Z. and Oditis, I. (2016) 'Towards NoSQL-based Data Warehouse Solutions', *Procedia Computer Science*. Elsevier Masson SAS, 104, pp. 104–111. doi: 10.1016/j.procs.2017.01.080.

BIMForum (2016) 'Level of development specification 2016'.

Blei, D. M. and Lafferty, J. D. (2009) 'Topic Models', *Text Mining: Classification, Clustering, and Applications*, pp. 71–89. doi: 10.1145/1143844.1143859.

Blignaut, S. (2006) 'Addressing Communication Silo's using Complecity TEchniques and Social Network Analysis', in Kazi, A. S. and Wolf, P. (eds) *Real-Life Knowledge Management: Lessons from the Field*. KnowledgeBoard - VTT Technical Research of Finland, pp. 83–101.

Bolisani, E. and Scarso, E. (1999) 'Information technology management: a knowledge-based perspective', *Technovation*, 19, pp. 209–217.

Bolpagni, M. and Ciribini, A. L. C. (2016) 'The Information Modeling and the Progression of Data-Driven Projects', *Proceedings of the CIB World Building Congress 2016. Volume III. Building up business operations and their logic. Shaping materials and technologies*, pp. 296–307. Available at: http://www.bimthinkspace.com/2016/07/the-many-faces-of-lod.html.

Bonduel, M. (2018) *Towards a PROPS ontology, W3C Linked Building Data Community Group*. Available at: https://github.com/w3c-lbd-cg/lbd/blob/gh-pages/presentations/props/presentation_LBDcall_20180312_final.pdf (Accessed: 22 March 2018).

Bouhali, R. and Laurent, A. (2015) 'Exploiting RDF Open Data Using NoSQL Graph Databases', in Chbeir, R., Manolopoulos, Y., Maglogiannis, I., and Alhajj, R. (eds) *11th IFIP International Conference on Artificial Intelligence Applications and Innovations, AICT-458*. Bayonne, France: Springer, pp. 177–190. doi: 10.1007/978-3-319-23868-5_13.

Boutell, M. R., Luo, J., Shen, X. and Brown, C. M. (2004) 'Learning multi-label scene classification', *Pattern Recognition*, 37(9), pp. 1757–1771. doi: 10.1016/j.patcog.2004.03.009.

Bresnen, M., Edelman, L., Newell, S., Scarborough, H. and Swan, J. (2003) 'Social practices and the management of knowledge in project environments', *International Journal of Project Management*, 21(3), pp. 157–166.

Brilakis, I., Soibelman, L. and Shinagwa, Y. (2005) 'Material-based construction site image retrieval', *Journal of Computing in Civil Engineering*, 19(4), pp. 341–355.

Browne, J., Sackett, P. J. and Wortmann, J. C. (1994) *The system of Manufacturing: a Perspective Study, Report to the DG XII of the CEC, European Commission*.

Brüggemann, B. M., Holz, K. and Molenthin, F. (2000) 'Semantic documentation in engineering', in *Proceeding of the ICCCBE-VIII*. Palo Alto, CA, pp. 828–835.

BSI (2007) *BS 1192:2007, Collaborative production of architectural , engineering and construction information – Code of practice*. UK: The British Standards Institution.

BSI (2013) *PAS 1192-2:2013, Specification for information management for the capital/delivery phase of construction projects using building information modelling*. UK: The British Standards Institution.

BSI (2014) *PAS 1192-3:2014, Specification for information management for the operational phase of assets using building information modelling*. UK: The British Standards Institution.

Building smart (2010) *Information Delivery Manual Guide to Components and Development Methods, buildingSMART*.

Building Smart (2014) *Technical Roadmaps*. Available at: https://www.buildingsmart.org/standards/technical-vision/technical-roadmaps/ (Accessed: 21 March 2018).

Building Smart (2018a) *IFC Introduction*. Available at: https://www.buildingsmart.org/about/what-is-openbim/ifc-introduction/ (Accessed: 18 March 2018).

Building Smart (2018b) *Linked Data Working Group*. Available at: http://www.buildingsmart-tech.org/future/linked-data (Accessed: 21 January 2019).

Building Smart (2018c) *MVS overview summary*. Available at: http://www.buildingsmart-tech.org/specifications/mvd-overview (Accessed: 31 August 2018).

Cabena, P., Hadjnian, P., Stadler, R., Verhees, J. and Zanasi, A. (1998) *Discovering data mining from concept to implementation*. Prentice Hall.

Cabinet Office (2011) *Government Construction Strategy*. UK. doi: Vol 19.

Caldas, C. H. and Soibelman, L. (2003) 'Automating hierarchical document classification for construction management information systems', *Automation in Construction*, 12(4), pp. 395–406. doi: 10.1016/S0926-5805(03)00004-9.

Caldas, C. H., Soibelman, L. and Gasser, L. (2005) 'Methodology for the Integration of Project Documents in Model-Based Information Systems', *Journal of Computing in Civil Engineering*, 19(1), pp. 25–33. doi: 10.1061/(ASCE)0887-3801(2005)19:1(25).

Caldas, C. H., Soibelman, L. and Han, J. (2002) 'Automated Classification of Construction Project Documents', *Journal of Computing in Civil Engineering*, 16(4), pp. 234–243.

Carrara, G., Fioravanti, A. and Nanni, U. (2009) 'Knowledge-based collaborative architectural design', *International journal of design science & technology*, 1(16), pp. 1–16.

Carrillo, P. M., Anumba, C. J. and Kamara, J. M. (2000) 'Knowledge Management Strategy for Construction : Key I.T. and Contextual Issues', *Constructions Informatics Digital Library*. Available at: https://www.researchgate.net/profile/Chimay_Anumba/publication/239603583_KNOWLEDGE_MANAGEMENT_STRATEGY_FOR_CONSTRUCTION_KEY_I._T._AND_CONTEXTUAL_ISSUES/links/00b7d52e2980c983e1000000.pdf.

Carrillo, P. M., Robinson, H. S., Anumba, C. J. and Al-Ghassani, A. M. (2003) 'A Framework for Linking Knowledge Management to Business Performance', *Electronic Journal of Knowledge Management*, 1(1), pp. 1–12.

Cerovsek, T. (2011) 'A review and outlook for a "Building Information Model" (BIM): A multi-standpoint framework for technological development', *Advanced Engineering Informatics*. Elsevier Ltd, 25(2), pp. 224–244. doi: 10.1016/j.aei.2010.06.003.

Chang, F., Dean, J., Ghemawat, S., Hsieh, W. C., Wallach, D. A., Burrows, M., Chandra, T., Fikes, A. and Gruber, R. E. (2006) 'Bigtable: A distributed storage system for structured data', in *7th Symposium on Operating Systems Design and Implementation (OSDI '06), November 6-8*. Seattle, WA, USA, pp. 205–218. doi: 10.1145/1365815.1365816.

Charbuck, D. and Young, J. S. (1992) 'The Virtual Workplace', *Forbes*, 150(12), pp. 184–190.

Cheng, J. C. P. and Das, M. (2014) 'A bim-based web service framework for green building energy simulation and code checking', *Journal of Information Technology in Construction*, 19(July), pp. 150–168.

Cheng, T. and Teizer, J. (2012) 'Modeling tower crane operator visibility to minimize the risk of limited situational awareness', *Journal of Computing in Civil Engineering*, 14.

Cherns, A. B. and Bryant, D. T. (1984) 'Studying the client ' s role in construction management', *Construction Management and Economics*, 2, pp. 177–184. doi: 10.1080/01446198400000016.

Chowdhury, G. G. (2003) 'Natural language processing', *Review of Information Science and Technology*, 37(1), pp. 51–89.

Chua, W. F. (1986) 'Radical developments in accounting thought', *The Accounting Review*, 61, pp. 601–632.

Chuang, T., Lee, B. and Wu, I. (2011) 'Applying Cloud Computing Technology To Bim Visualization and Manipulation', in *Proceedings of 28th International Symposium on Automation and Robotics in Construction*, pp. 144–149. Available at: http://www.iaarc.org/publications/2011_proceedings_of_the_28th_isarc_seoul_korea.

Cooper, R., Aouad, G., Lee, A., Wu, S., Fleming, A. and Kagioglou, M. (2005) *Process Management in Design and Construction*. Blackwell Publishing.

Cross, R. (1998) 'Managing for knowledge: managing for growth', *Knowledge Management*, 1(3), pp. 9–13.

CSI and CSC (2010) *UniFormat - A Uniform Classification of Construction Systems and Assembles*.

CSI and CSC (2016) *MasterFormat - Master List of Numbers and Titles for the Construction Industry*.

Curry, E., O'Donnell, J., Corry, E., Hasan, S., Keane, M. and O'Riain, S. (2013) 'Linking building data in the cloud: Integrating cross-domain building data using linked data', *Advanced Engineering Informatics*. Elsevier Ltd, 27, pp. 206–219. doi: 10.1016/j.aei.2012.10.003.

Datacubist (2018) *SimpleBIM - Datacubist*. Available at: http://www.datacubist.com/.

Dave, B. and Koskela, L. (2009) 'Collaborative knowledge management-A construction case study', *Automation in Construction*. Elsevier B.V., 18(7), pp. 894–902. doi: 10.1016/j.autcon.2009.03.015.

Davenport, T. H. and Prusak, L. (1998) *Working Knowledge: How Organizations Manage*

What They Know. Boston, MA: Harvard Business School Press.

Day, M. (2017) *Autodesk Project Quantum: the future of BIM?*, *AEC Magazine*. Available at: https://www.aecmag.com/59-features/1241-autodesk-project-quantum-revit-bim (Accessed: 21 March 2018).

Dayan, R., Pasher, E. and Dvir, R. (2006) 'The Knowledge Management Journey of Israel Aircraft Industry', in Kazi, A. S. and Wolf, P. (eds) *Real-Life Knowledge Management: Lessons from the Field*. KnowledgeBoard - VTT Technical Research of Finland, pp. 25–45.

Dean, J. and Ghemawat, S. (2004) 'MapReduce: Simplified Data Processing on Large Clusters', in *Proceeding of 6th Symposium on Operating Systems Design & Implementation*. San Francisco, CA. Available at: https://www.usenix.org/legacy/publications/library/proceedings/osdi04/tech/full_papers/dean/dean_html/.

DeCandia, G., Hastorun, D., Jampani, M., Kakulapati, G., Lakshman, A., Pilchin, A., Sivasubramanian, S., Vosshall, P. and Vogels, W. (2007) 'Dynamo: Amazon's Highly Available Key-value Store', in *Proceedings of the Symposium on Operating Systems Principles*, pp. 205–220. doi: 10.1145/1323293.1294281.

Delany, S. (2018) *Classification - Uniclass 2015*. Available at: https://toolkit.thenbs.com/articles/classification#classificationtables (Accessed: 14 March 2018).

DellaValle, E., Celino, I. and Cerizza, D. (2009) *Semantic Web: Dai fondamenti alla realizzazione di un'applicazione*. Edited by Pearson.

Delone, W. H. and McLean, E. R. (2003) 'The DeLone and McLean Model of Information Systems Success : A Ten-Year Update', *Journal of Management Information Systems*, 19(4), pp. 9–30.

Deshpande, A., Azhar, S. and Amireddy, S. (2014) 'A framework for a BIM-based knowledge management system', *Procedia Engineering*, 85, pp. 113–122. doi: 10.1016/j.proeng.2014.10.535.

Designing Buildings Ltd. (2018) *BIM toolkit*. Available at: https://www.designingbuildings.co.uk/wiki/BIM_toolkit (Accessed: 28 September 2018).

Devlin, B. (1997) *Data Warehouse: From Architecture to Implementation*. Menlo Park, CA: Addison Wesley Longman.

Diamonds, J. (2011) *Collapse - How societies choose to fail or survive*. London: Penguin Books.

Dickinson, J., Pardasani, A., Ahamed, S. and Krujthof, S. (2009) 'A survey of automation technology for realising as-built models of services', in *1st International Conference on Improving Construction and Use through Integrated Design Solutions*. Espoo, Finland.

Dimyadi, J., Pauwels, P. and Amor, R. (2016) 'Modelling and accessing regulatory knowledge for computer-assisted compliance audit', *Journal of Information Technology in Construction*, 21(July), pp. 317–336.

Dourish, P. and Bellotti, V. (1992) 'Awareness and Coordination in Shared Workshops', in *Proc. of CSCW'92*. Toronto, Canada: ACM Press.

DTI UK (2006) *Construction Statistic Annual*. London.

Eastman, C., Teicholz, P., Sacks, R. and Liston, K. (2011) *BIM Handbook: A guide to Building Information Modeling for owners, managers, architects, engineers, contractors, and fabricators*. Second edi. Edited by John Wiley & Sons. Hoboken, NJ, USA.

ECMA (2017) *The JSON Data Interchange Syntax - Standard ECMA - 404*, *ECMA International*. doi: 10.17487/rfc7158.

Egbu, C. O., Kurul, E., Quintas, P., Hutchinson, V., Anumba, C. and Ruikar, K. D. (2003) 'Knowledge production, resources and capabilities in the construction industry', *Work package 1 - final report, knowledge management for sustainable construction competitiveness project*.

Egbu, C. O. and Sturges, J. (2001) 'Knowledge management in small and medium enterprises in the construction industry: challenges and opportunities', in *Managing Knowledge Conference, University of Leicester*.

Eidal, L. C., Hjorth, F. and Kadal, J. (2006) 'Applying Business Modelling to Knowledge Management', in Kazi, A. S. and Wolf, P. (eds) *Real-Life Knowledge Management: Lessons from the Field*. KnowledgeBoard - VTT Technical Research of Finland, pp. 131–144.

El-Diraby, T. E. (2013) 'Domain ontology for construction knowledge', *Journal of Construction Engineering and Management*, 139(7), pp. 768–784.

Elghamrawy, T. and Boukamp, F. (2008) 'A vision for a framework to support management of and learning from construction problems', in *Proceedings of the 25th CIB W78 International Conference on Information Technology in Construction*. Santiago, Chile.

Engeström, Y. (1987) *Learning by Expanding: an activity-theoretical approach to developmental research*. Helsinki: Orienta-Konsultit.

Engeström, Y. (2001) 'Expansive Learning at Work: toward an activity theoretical reconceptualization', *Journal of Education and Work*, 14(1), pp. 133–156. doi: 10.1080/13639080123238.

Engeström, Y. and Escalante, V. (1996) 'Mundane tool or object of affection? The rise and fall of the Postal Buddy', in Nardi, B. A. (ed.) *Context and Consciousness: Activity Theory and Human-Computer Interaction*. Cambridge: The MIT press.

Engeström, Y., Miettinen, R. and Punamäki, R.-L. (1999) *Perspectives on Activity Theory*. Cambridge: Cambridge University Press.

Enkovaara, E., Heikkonen, A. and Tapionen, T. (1998) *Measurement of maturity and benfits of construction information technology*. Finland: Mimeo.

EU (2016a) 'COM(2016) 176 final - ICT Standardisation Priorities for the Digital Single Market'. European Commission.

EU (2016b) 'COM(2016) 178 final - European Cloud Initiative- Building a competitive data and knowledge economy in Europe'. European Commission.

EU (2016c) 'COM(2016) 179 final -Accelerating the digital transformation of government'. European Commission.

EU (2016d) 'COM(2016) 180 final - Digitising European Industry - Reaping the full benefits of a Digital Single Market'. European Commission.

EU (2016e) *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Da*.

Fantino, E. and Stolarz-Fantino, S. (2005) 'Decision-making: Context matters', *Behavioural Processes*, 69(2), pp. 165–171. doi: 10.1016/j.beproc.2005.02.002.

Farias, T. M. de, Roxin, A. and Nicolle, C. (2018) 'A rule-based methodology to extract building model views', *Automation in Construction*, 92(May), pp. 214–229. doi: 10.1016/j.autcon.2018.03.035.

Fassi, F., Achille, C., Mandelli, A., Rechichi, F. and Parri, S. (2015) 'A new idea of BIM system for visualization, web sharing and using huge complex 3d models for facility management', *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XL-5/W4, pp. 359–366. doi: 10.5194/isprsarchives-XL-5-W4-359-2015.

Fassi, F., Achille, C., Monti, C., Fregonese, L., Parri, S., Rechichi, F., Mandelli, A., Gaudio, F. and Galbusera, L. (2013) 'A yard of the future: the main spire of Milan Cathedral', in *BUILT HERITAGE 2013 monitoring Conservation Management*.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996a) 'From data mining to knowledge discovery: an overview', in Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R. (eds) *Advanced in knowledge discovery and data mining*. Menlo Park: AAAI Press, pp. 1–30.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996b) 'From Data Mining to Knowledge Discovery in Databases', *AI Magazine*, 17(3), pp. 37–54. doi: 10.1609/aimag.v17i3.1230.

Fernie, S., Weller, S., Green, S. D., Newcombe, R. and Williams, M. (2001) 'Learning across business sectors: context, embeddedness and conceptual chasms', in *In 17th Annual ARCOM Conference, University of Salford, 5-7 September 2001*, pp. 557–565. Available at: http://www.reading.edu.my/web/FILES/innovativeconstructionresearchcentre/icrc-18-i-ARCOM.pdf.

Fioravanti, A., Novembri, G. and Rossini, F. L. (2017) 'Improving Proactive Collaborative Design Through the Integration of BIM and Agent-Based Simulations', in *Proceedings of the 35th International Conference on Education and Research in Computer Aided Architectural Design in Europe*. Rome, pp. 103–108.

Fortune Knowledge Group (2018) 'Artificial Intelligence : The Next Frontier in Drug Discovery'. Available at: https://www-01.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=17016517USEN&.

Frakes, W. B. and Baeza-Yates, R. (1992) *Information Retrieval: Data Structure and Algorithms*. Prentice Hall.

Franco, L. A., Cushman, M. and Rosenhead, J. (2004) 'Project review and learning in the construction industry: embedding a problem structuring method within a partnership context', *European Journal of Operational Research*, 152, pp. 586–601.

Fregonese, L., Achille, C., Adami, A., Fassi, F., Spezzoni, A. and Taffurelli, L. (2015) 'Bim: an integrated model for planned and preventive maintenance of architectural heritage', in *2015 Digital Heritage, Volume 2, 28 September - 2 October*. Granada, Spain. doi: 10.1109/DigitalHeritage.2015.7413832.

Fuchs, S., Katranuschkov, P. and Scherer, R. J. (2010) 'A Framework for Multi-Model Collaboration and Visualisation', in *Proc. ECPPM, 14-16 September*. Cork, Ireland: Taylor & Francis, pp. 115–120. doi: 10.1201/b10527-20.

Furukawa, Y., Curless, B., Seitz, S. M. and Szeliski, R. (2009) 'Reconstructing building interiors from images', in *IEEE 12th International Conference on Computer Vision*, pp. 80–87. doi: 10.1109/ICCV.2009.5459145.

Gann, D. M. and Salter, A. J. (2000) 'Innovation in project-based, service-enhanced firms: the construction of complex products and systems', *Research Policy*, 29, pp. 955–972.

Garrett, J. H., Fenves, S. J. and Stasiak, D. M. (1996) 'A WWW-based regulation broker', in *CIB proceedings publication 198: Construction on the information Highway*. Rotterdam: CIB, pp. 219–230.

Gessert, F., Wingerath, W., Friedrich, S. and Ritter, N. (2017) 'NoSQL database systems: a survey and decision guidance', *Computer Science - Research and Development*. Springer Berlin Heidelberg, 32(3–4), pp. 353–365. doi: 10.1007/s00450-016-0334-3.

Goh, Y. M. (2009) 'Case-based reasoning approach to construction safety hazard identification: adaptation and utilization', *Journal of Construction Engineering and Management*, 136(2), pp. 170–178.

Goh, Y. M. and Chua, D. (2009) 'Case-based reasoning for construction hazard identification: case representation and retrieval', *Journal of Construction Engineering and Management*, 135(11), pp. 1181–1189.

Google (2018) *The Open Images Dataset*. Available at: https://github.com/openimages/dataset (Accessed: 20 March 2018).

Graphisoft (2018) *Nemetscheck Group - Graphisoft BIM Cloud*. Available at: http://www.graphisoft.com/bimcloud/overview/ (Accessed: 18 March 2018).

Grillitsch, W. and Muller-Stingl, A. (2006) 'Enabling Cross-Project Knowledge Creation through "Knowledge Oriented Project Supervision"', in Kazi, A. S. and Wolf, P. (eds) *Real-Life Knowledge Management: Lessons from the Field*. KnowledgeBoard - VTT Technical Research of

Finland, pp. 165–179.

Guo, W., Soibelman, L. and Garrett, J. H. (2009) 'Visual pattern recognition supporting defect reporting and condition assessment of wastewater collection systems', *Journal of Computing in Civil Engineering*, 23(3), pp. 160–169.

Gupta, A. and McDaniel, J. (2002) 'Creating competitive advantage by effectively managing knowledge: A framework for knowledge management', *Journal of Knowledge Management Practice*, 3.

Gupta, V. and Lehal, G. S. (2009) 'A Survey of Text Mining Techniques and Applications', *Journal of Emerging Technologies in web Intelligence*, 1(1), pp. 60–76.

Hadsell, R., Chopra, S. and LeCun, Y. (2006) 'Dimensionality reduction by learning an invariant mapping', *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1735–1742. doi: 10.1109/CVPR.2006.100.

Han, J., Kamber, M. and Pei, J. (2012) *Data Mining: Concepts and Techniques*. Edited by Morgan Kaufmann. San Francisco, CA.

Harris, S. and Seaborne, A. (2013) *SPARQL 1.1 Query Language*. Available at: https://www.w3.org/TR/sparql11-query/ (Accessed: 21 March 2018).

Hartmann, A. and Bresnen, M. (2011) 'The emergence of partnering in construction practice: an activity theory perspective', *Engineering Project Organization Journal*, 1, pp. 41–52. doi: 10.1080/21573727.2010.549609.

Hecht, R. and Jablonski, S. (2011) 'NoSQL evaluation: A use case oriented survey', in *International Conference on Cloud and Service Computing, CSC*, pp. 336–341. doi: 10.1109/CSC.2011.6138544.

Hernández, M. A. and Stolfo, S. J. (1998) 'Real-world Data is Dirty: Data Cleansing and The Merge/Purge Problem', *Data Mining and Knowledge Discovery*, 2(1), pp. 9–37. doi: 10.1023/A:1009761603038.

Hevner, A. R., March, S. T., Park, J. and Ram, S. (2004) 'Design Science in Information System Research', *MIS Quarterly*, 28(1), pp. 75–105.

Hobbs, R. W. (1996) 'Leadership through collaboration', *AIArchitect*, 3, p. 11.

Horenburg, T. and Günthner, W. A. (2013) 'Construction Scheduling and Resource Allocation based on Actual State Data', *Computing in Civil Engineering*.

Hu, Z. Z., Zhang, X. Y., Wang, H. W. and Kassem, M. (2016) 'Improving interoperability between architectural and structural design models: An industry foundation classes-based approach with web-based tools', *Automation in Construction*. Elsevier B.V., 66, pp. 29–42. doi: 10.1016/j.autcon.2016.02.001.

Huang, A. (2008) 'Similarity measures for text document clustering', in *Proceedings of the New Zealand Computer Science Research Student Conference 2008*, pp. 49–56.

Huber, D., Akinci, B., Adan, A., Anil, A., Okorn, B. and Xiong, X. (2011) 'Methods for automatically modeling and representing as-built building information models', in *Proceedings of 2011 NSF Engineering Research and Innovation Conference*. Atlanta, Georgia. Available at: https://www.ri.cmu.edu/pub_files/2011/1/2011-huber-cmmi-nsf-v4.pdf.

Hypertable (2014) *Hypertable*. Available at: http://www.hypertable.org/ (Accessed: 9 March 2018).

IEEE (1990) 'Standard Computer Dictionary. A Compilation of IEEE Standard Computer Glossaries'. New York: Institute of Electrical and Electronics Engineers.

Ifcopenshell (no date) *ifcopenshell*. Available at: http://academy.ifcopenshell.org/ (Accessed: 9 April 2018).

Iiritano, S. and Ruffolo, M. (2001) 'Managing the knowledge contained in electronic documents:

a clustering method for text mining', in *12th International Workshop on Database and Expert Systems Applications*. Munich, Germany: IEEE Comput. Soc, pp. 454–458. doi: 10.1109/DEXA.2001.953103.

Ilich, M., Becerik, B. and Aultman, B. (2006) 'Online Collaboration: Why Aren't We Using Out Tools?', *Means, Methods and Trends, Publication of Architectural Engineering Institute and the Construction Institute of ASCE*. Available at: http://i-lab.usc.edu/documents/Online Collaboration-Why Aren't We Using Our Tools.pdf.

Ilin, I. V, Kalinina, O. V, Levina, A. I. and Iliashenko, O. Y. (2016) 'Approach to Organizational Structure Modelling in Construction Companies', *MATEC Web of Conferences*. Edited by V. Andreev, 86.

Inman, W. H. (1996) *Building the Data Warehouse*. New York: Wiley.

Iordanov, B. (2010) 'HyperGraphDB: A generalized graph database', in *International Conference on Web-Age Information Management*. doi: 10.1007/978-3-642-16720-1_3.

Ismail, A. and Scherer, R. J. (2018) 'A graph-based approach for management and linking of BIM models with further AEC domain models', in Karlshøj, J. and Scherer, R. J. (eds) *eWork and eBusiness in Architecture, Engineering and Construction, ECCPM*. Copenhagen, Denmark: CRC Press, pp. 377–383.

ISO (2007) 'ISO 12006-3: Building construction - organization of information about construction works - Part 3: Framework for object-oriented information'.

ISO (2013) 'ISO 16739: Industry Foundation Classes (IFC) for data sharing in the construction and facility management industries'.

ISO (2015) *ISO 12006-2: Building construction - Organization of information about construction works - Part 2: Framework for classification*.

ISO (2016) 'ISO 29481 - Building Information Models -- Information Delivery Manual'.

ISO (2018) *ISO 19650-1 - Organization of information about construction works — Information management using building information modelling — Part 1: Concepts and Principles*.

Jain, A. K. and Dubes, R. C. (1988) *Algorithms for Clustering Data*. Englewood Cliffs, NJ: Prentice Hall.

Jain, A. K. and Flynn, P. J. (1996) 'Image Segmentation Using Clustering', in Ahuja, N. and Bowyer, K. (eds) *Advances in Image Understanding: A Festschrift for Azriel Rosenfeld*. IEEE Computer Society Press, pp. 65–83.

Jain, A. K., Murty, M. N. and Flynn, P. J. (1999) 'Data Clustering: A review', *ACM Computing Surveys (CSUR)*, 31(3), pp. 264–323.

Jernigan, F. (2007) *Big BIM, little bim: the practical approach to building information modeling; integrated practice done the right way*. 1st edn. 4Site Press.

Jiao, Y., Wang, Y., Zhang, S., Li, Y., Yang, B. and Yuan, L. (2013) 'A cloud approach to unified lifecycle data management in architecture, engineering, construction and facilities management: Integrating BIMs and SNS', *Advanced Engineering Informatics*. Elsevier Ltd, 27(2), pp. 173–188. doi: 10.1016/j.aei.2012.11.006.

Jimes, C. and Lucardie, L. (2003) 'Reconsidering the tacit-explicit distinction - A move toward functional (tacit) knowledge management', *Electronic Journal of Knowledge Management*, 1(1), pp. 23–32.

Joachims, T. (1998) 'Text categorization with support vector machines: Learning with many relevant features', in Nédellec, C. and Rouveirol, C. (eds) *Machine Learning: ECML-98. ECML 1998. Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence)*. Berlin, Heidelberg: Springer. doi: https://doi.org/10.1007/BFb0026683.

Jonassen, D. H. and Hernandez-Serrano, J. (2002) 'Case-based reasoning and instructional design: Using stories to support problem solving', *Educational Technology Research &*

Development, 50(2), pp. 65–77. doi: 10.1007/BF02504994.

Kamara, J. M., Anumba, C. J. and Carrillo, P. M. (2002) 'A CLEVER approach to selecting a knowledge management strategy', *International Journal of Project Management*, 20(3), pp. 205–211. doi: 10.1016/S0263-7863(01)00070-9.

Kamaruzzaman, S. N., Zawawi, A. M. A., Shafie, M. O. and Noor, A. M. (2016) 'Assessing the readiness of facilities management organizations in implementing knowledge mangament systems', *Journal of Facilities Management*, 14(1), pp. 69–83. doi: 10.1108/MBE-09-2016-0047.

Kargul, A., Günthner, W., Bügler, M. and Borrmann, A. (2015) 'Web based field data analysis and data-driven simulation application for construction performance prediction', *Journal of Information Technology in Construction*, ITcon Vol.(August), pp. 479–494. Available at: http://www.itcon.org/2015/28.

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R. and Fei-Fei, L. (2014) 'Large-scale video classification with convolutional neural networks', in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1725–1732. doi: 10.1109/CVPR.2014.223.

Katranuschkov, P., Weise, M., Windisch, R., Sebastian, F. and Scherer, R. J. (2010) 'BIM-based generation of multi-model views.', in *Proceedings of the CIB W78 Conference 2010: 27th International Conference 16-18 November*. Cairo, Egypt.

Kazi, A. S. (2005) *Knowledge Management in the Construction Industry: A Socio-Technical Perspective*. London: Idea Group.

Kazi, A. S. and Koivuniemi, A. (2006) 'Sharing through Social Interaction: The Case of YIT Construction Ltd', in Kazi, A. S. and Wolf, P. (eds) *Real-Life Knowledge Management: Lessons from the Field*. KnowledgeBoard - VTT Technical Research of Finland, pp. 65–79.

Kelly, J. (2014) *Big Data: Hadoop, Business Analytics and Beyond*. Available at: http://wikibon.org/wiki/v/Big_Data:_Hadoop,_Business_Analytics_and_Beyond (Accessed: 23 March 2018).

Keras (2018) *Keras Packages*. Available at: https://keras.io/ (Accessed: 20 March 2018).

Kim, W., Choi, B. J., Hong, E. K., Kim, S. K. and Lee, D. (2003) 'A Taxonomy of Dirty Data', *Data Mining and Knowledge Discovery*, 7(1), pp. 81–99. doi: 10.1023/A:1021564703268.

Klein, L., Li, N. and Becerik-Gerber, B. (2012) 'Imaged-based verification of as-built documentation of operational buildings', *Automation in Construction*, 21(1), pp. 161–171. doi: 10.1016/j.autcon.2011.05.023.

Klyne, G., Carroll, J. J. and McBride, B. (2014) *RDF 1.1 Concepts and Abstract Syntax*. Available at: https://www.w3.org/TR/rdf11-concepts/ (Accessed: 21 March 2018).

Kogut, B. and Zander, U. (1992) 'Knowledge of the firm, combinative capabilities and the replication of technology', *Organization Science*, 3(3), pp. 383–397.

Kolodner, J. (1993) *Case-Based Reasoning*. San Mateo, CA: Morgan Kaufmann.

Konstantas, D., Bourrieres, D. J. P., Léonard, M. and Boudilida, N. (2005) 'Interoperability of Enterprise Systems and Applications', in *Proceedings of the International Conference on Interoperability for Enterprise Systems and Applications (I-ESA)*. Geneva: Springer-Verlag.

Koskela, L. and Vrijhoef, R. (2001) 'Is the current theory of construction a hindrance to innovation?', *Building Research & Information*, 29(3), pp. 197–207. Available at: http://dx.doi.org/10.1080/09613210110039266%5Cnhttp://www.tandfonline.com/doi/abs/10.1080/09613210110039266?src=recsys%5Cnhttp://www.tandfonline.com/doi/pdf/10.1080/09613210110039266.

Krizhevsky, A., Sutskever, I. and Geoffrey E., H. (2012) 'ImageNet Classification with Deep Convolutional Neural Networks', in *Advances in Neural Information Processing Systems 25 (NIPS2012)*, pp. 1090–1098. doi: 10.1109/5.726791.

Kululanga, G. K. and McCaffer, R. (2001) 'Measuring knowledge management for construction organizations', *Engineering, construction and architectural management*, 8(5/6), pp. 346–354.

Kumar, V. and Viswanadham, N. (2007) 'A SBR-based decision support system framework for construction supply chain risk management', in *Proceeding of 2007 IEEE International Conference on Automation Science and Engineering*. Scottsdale, AZ.

Kvan, T. (2000) 'Collaborative design, what is it?', *Automation in Construction*, (9), pp. 409–415.

Laakso, M. and Kiviniemi, A. (2012) 'The IFC standard - a review of history, development, and standardization', *Journal of Information Technology in Construction*, 17, pp. 134–161.

Laney, D. (2001) *3d data management: Controlling data volume, velocity, and variety, Tech. rep, META Group*. doi: 10.1016/j.infsof.2008.09.005.

Lawrence, S., Giles, C. L., Tsoi, A. C. and Back, A. D. (1997) 'Face recognition: A convolutional neural-network approach', *IEEE Transactions on Neural Networks*, 8(1), pp. 98–113. doi: 10.1109/72.554195.

Lecun, Y., Bengio, Y. and Hinton, G. (2015) 'Deep learning', *Nature*, 521(7553), pp. 436–444.

Lee, S. and Chang, L. (2005) 'Digital image processing methods for assessing bridge painting rust defects and their limitations', in *Proc. of the International Conference on Computing in Civil Engineering*. Cancun, Mexico: American Society of Civil Engineers.

Lee, Y.-C., Eastman, C. M. and Lee, J.-K. (2015) 'Validations for ensuring the interoperability of data exchange of a building information model', *Automation in Construction*. Elsevier B.V., 58, pp. 176–195. doi: 10.1016/j.autcon.2015.07.010.

Leont'ev, A. N. (1978) *Activity, Consciousness, and Personality*. Englewood Cliffs, NJ: Prentice-Hall.

Leung, S. wing, Mak, S. and Lee, B. L. P. (2008) 'Using a real-time integrated communication system to monitor the progress and quality of construction works', *Automation in Construction*, 17(6), pp. 749–757. doi: 10.1016/j.autcon.2008.02.003.

Lewis, A. and Diamond, B. (2006) 'Knowledge Management for Organisations Dealing with Critical Incidents', in Kazi, A. S. and Wolf, P. (eds) *Real-Life Knowledge Management: Lessons from the Field*. KnowledgeBoard - VTT Technical Research of Finland, pp. 237–251.

Liang, M. and Hu, X. (2015) 'Recurrent convolutional neural network for object recognition', *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3367–3375. doi: 10.1109/CVPR.2015.7298958.

Licklider, J. C. R. (1960) 'Man-Computer Symbiosis', *IRE Transactions on Human Factors in Electronics*, (March), pp. 4–11. doi: 10.1109/THFE2.1960.4503259.

Liebich, T. and Wix, J. (1999) 'Highlights of the development process of Industry Foundation Classes', *Durability of building materials and components*.

Lin, K.-Y. and Soibelman, L. (2007) 'Knowledge-Assisted Retrieval of Online Product Information in Architectural/Engineering/Construction', *Journal of Construction Engineering and Management*, 133(11), pp. 871–879. doi: 10.1061/(ASCE)0733-9364(2007)133:11(871).

Lin, Y.-C., Wang, L.-C., Tserng, H.-P. and Jan, S.-H. (2005) 'Enanching knowledge and experience exchange through construction map-based knowledge management system', in *Construction research congress 2005: Broadening perspectives, American society of civil engineers*. San Diego, California: American Society of Civil Engineers.

Lin, Y. C., Wang, L. C. and Tserng, H. P. (2006) 'Enhancing knowledge exchange through web map-based knowledge management system in construction: Lessons learned in Taiwan', *Automation in Construction*, 15(6), pp. 693–705. doi: 10.1016/j.autcon.2005.09.006.

Lin, Y. and Tserng, H. (2003) 'Knowledge Management and its application to Lean Construction', in *... International Group for Lean Construction (...*, pp. 1–12. Available at: http://www.leanconstruction.dk/media/17764/Knowledge Management and its Appication to Lean

Construction.pdf.

Liu, B. (2006) *Web data mining: Exploring hyperlinks, contents, and usage data (data-centric systems and applications)*. New York, NY, USA: Springer Verlag.

Llach, D. C. (2015) *Builder of the Vision: Software and the Imagination of Design*. Routledge.

Logothetis, S., Delinasiou, A. and Stylianidis, E. (2015) 'Building Information Modelling for Cultural Heritage: a Review', in *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Science*. Taipei, Taiwan.

LOV (2018) *Linked Open Vocabularies (LOV)*. Available at: http://lov.okfn.org/dataset/lov/ (Accessed: 21 March 2018).

Lu, Y., Li, Q. and Xiao, W. (2013) 'Case-based reasoning for automated safety risk analysis on subway operation: case representation and retrieval', *Safety Science*, 57, pp. 75–81.

Lundvall, B. Å. and Johnson, B. (1994) 'The learning economy', *Journal of Industry Studies*, 1, pp. 23–42.

Madjarov, G., Kocev, D., Gjorgjevikj, D. and Džeroski, S. (2012) 'An extensive experimental comparison of methods for multi-label learning', *Pattern Recognition*, 45(9), pp. 3084–3104. doi: 10.1016/j.patcog.2012.03.004.

Mahdjoubi, L., Moobela, C. and Laing, R. (2013) 'Providing real-estate services through the integration of 3D laser scanning and building information modelling', *Computers in Industry*. Elsevier B.V., 64(9), pp. 1272–1281. doi: 10.1016/j.compind.2013.09.003.

Makris, A., Tserpes, K., Andronikou, V. and Anagnostopoulos, D. (2016) 'A Classification of NoSQL Data Stores Based on Key Design Characteristics', *Procedia Computer Science*, 97, pp. 94–103. doi: 10.1016/j.procs.2016.08.284.

Manning, C. D., Raghavan, P. and Schütze, H. (2008) *Introduction to information retrieval*. New York, NY, USA: Cambridge University Press.

De Mantaras, R. L., Mcsherry, D., Bridge, D., Leake, D., Smyth, B., Craw, S., Faltings, B., Maher, M. Lou, Cox, M. T., Forbus, K., Keane, M., Aamodt, A. and Watsoni, I. (2005) 'Retrieval, reuse, revision and retention in case-based reasoning', *Knowledge Engineering Review*, 20(3), pp. 215–240. doi: 10.1017/S0269888906000646.

Maree, G. A., Roux, D. J. and Marais, M. A. (2006) 'Beneath the Surface of Conscious Patterns: Using Narrative to Characterise the Culture of Innovation at a Leading R&D Organization', in Kazi, A. S. and Wolf, P. (eds) *Real-Life Knowledge Management: Lessons from the Field*. KnowledgeBoard - VTT Technical Research of Finland, pp. 105–125.

Markey, S. C. (2012) *Deploy an OpenStack private cloud to a Hadoop MapReduce environment*. Available at: https://www.ibm.com/developerworks/cloud/library/cl-openstack-deployhadoop/ (Accessed: 17 March 2018).

Matsokis, A. and Kiritsis, D. (2010) 'An ontology-based approach for product lifecycle management', *Computers in Industry*, 61(8), pp. 787–797.

McCarthy, T. J., Kahn, H. J., Elhag, T. M. S., Williams, A. R., Milburn, R. and Patel, M. B. (2000) 'Knowledge management in the designer/constructor interface', in Fruchter, R., Peña-Mora, F., and Rodis, R. M. K. (eds) *Proceedings of the 8th International Conference on Computing in Civil and Building Engineering*. Reston, VA: American Society of Civil Engineers, pp. 836–843.

Meadows, D. H., Meadows, D. L., Randers, J. and Behrens, W. W. (1972) *The Limits to Growth: A report for the Club of Rome's Project on the Predicament of Manking*. Rome: Potomac Book.

Megahed, N. A. (2015) 'Towards a Theoretical Framework for HBIM Approach in Historic Preservation and Management', *International Journal of Architectural Research*, 9(3), pp. 130–147.

Mell, P. and Grance, T. (2011) 'The NIST Definition of Cloud Computing: Recommendations of the National Institute of Standards and Technology', *NIST Special Publication 800-145*. Gaithersburg: National Institute of Standards and Technology.

Mertone, R. K. (1965) *On the shoulder of giants*. University of Chicago Press.

Michael Lewis (2017) *The undoing project*. Penguin Books.

Miettinen, R., Kerosuo, H., Korpela, J., Mäki, T. and Paavola, S. (2012) 'An activity theoretical approach to BIM-research', in Gudnason, G. and Scherer, R. (eds) *eWork and eBusiness in Architecture, Engineering and Construction, ECCPM*. London, UK: Taylor & Francis Group, pp. 777–781. doi: 10.1201/b12516-124.

Miller, J. J. (2013) 'Graph Database Applications and Concepts with Neo4', in *Proceedings of the Southern Association for Information systems Conference*. Atlanta, GA, USA, pp. 141–147. Available at: http://sais.aisnet.org/2013/MillerJ.pdf.

Mirarchi, C., Pasini, D., Pavan, A. and Daniotti, B. (2017) 'Automated Ifc-Based Processes in the Construction Sector: a Method for Improving the Information Flow', in *Proceedings of the Joint Conference on Computing in Construction (JC3)*. Heraklion, Crete, Greece, pp. 491–498.

Mirarchi, C., Pavan, A. and Cianciulli, A. (2018) 'Semi-automated object identification between building information models and price lists: Social media technologies (hashtag) applied to Building Information Modeling (BIM)', in *BIM World*. Paris, France.

Mirończuk, M. M. and Protasiewicz, J. (2018) 'A recent overview of the state-of-the-art elements of text classification', *Expert Systems with Applications*, 106, pp. 36–54. doi: 10.1016/j.eswa.2018.03.058.

MongoDB (2018) *MongoDB*. Available at: https://www.mongodb.com/ (Accessed: 9 March 2018).

Moody, D. and Walsh, P. (1999) 'Measuring The Value Of Information: An Asset Valuation Approach', in *Seventh European Conference on Information Systems (ECIS'99)*. Frederiksberg, Denmark, pp. 1–17.

Motik, B., Grau, B. C., Horrocks, I., Wu, Z., Fokoue, A. and Lutz, C. (2012) *OWL 2 Web Ontology Language Profiles (Second Edition)*. Available at: https://www.w3.org/TR/owl2-profiles/ (Accessed: 21 March 2018).

Musso, S. F. and Franco, G. (2014) 'The "Albergo dei Poveri" in Genova: conserving and using in the Incertainty and in the Provisional', in Della Torre, S. (ed.) *ICT per il miglioramento del processo conservativo*. Firenze: Nardini, pp. 41–50.

Naumann, F. (2013) 'Duplicate Detection'. Information System Group (HPI), Potsdam University.

Naumann, F. and Rolker, C. (2000) 'Assessment Methods for Information Quality Criteria', in *Proceedings of the 2000 Conference on Information Quality*, pp. 1–18.

Naumann, F. and Sattler, U. K. (2006) 'Information Quality Fundamentals, Techniques, and Use'. Munich, Germany: EDBT Tutorial.

NBS (2015) *BIM toolkit*. Available at: https://toolkit.thenbs.com/ (Accessed: 1 April 2017).

NBS (2018) *BIM toolkit*. Available at: https://toolkit.thenbs.com/ (Accessed: 25 September 2018).

Nemati, H. R., Steiger, D. M., Iyer, L. S. and Herschel, R. T. (2002) 'Knowledge warehouse: An architectural integration of knowledge management, decision support, artificial intelligence and data warehousing', *Decision Support Systems*, 33(2), pp. 143–161. doi: 10.1016/S0167-9236(01)00141-5.

Neo4j (2018) *Neo4j*. Available at: https://neo4j.com/ (Accessed: 9 March 2018).

Ng, H. S., Toukourou, A. and Soibelman, L. (2006) 'Knowledge discovery in a facility condition assessment database using text clustering', *Journal of Infrastructure Systems*, 12(1), pp. 50–59.

Nicolle, C. and Cruz, C. (2011) 'Semantic Building Information Model and multimedia for facility management', *Information Systems and Technologies, Lecture Notes in Business Information Processing*, pp. 14–29. doi: 10.1007/978-3-642-22810-0_2.

Niharika, S., Latha, V. S. and Lavanya, D. R. (2012) 'A survey on text categorization', *International Journal of Computer Trends and Technology*, 3(1), pp. 39–45. Available at: http://www.ijcttjournal.org/Volume3/issue-1/IJCTT-V3I1P108.pdf.

Nitithamyong, P. and Skibniewski, M. J. (2004) 'Web-based construction project management systems: How to make them successful?', *Automation in Construction*, 13(4), pp. 491–506. doi: 10.1016/j.autcon.2004.02.003.

Nonaka, I. and Takeuchi, H. (1995) *The Knowledge-creating Company: How Japanese Companies Create the Dynamics of Innovation*. New York and Oxford: Oxford University Press.

Nonaka, I., Toyama, R. and Byosière, P. (2001) 'Theory of Organizational Knowledge Creation: Understanding the Dynamic Process of Creating Knowledge', *In: DIERKES, M.; et al. Handbook Organizational Learning and Knowledge*, pp. 491–517.

OECD (2000) *Small and Medium-sized Enterprises: Local Strength, Global Reach, Policy Brief*. doi: 10.1177/0022146511418950.

Oh, M., Lee, J., Hong, S. W. and Jeong, Y. (2015) 'Integrated system for BIM-based collaborative design', *Automation in Construction*. Elsevier B.V., 58, pp. 196–206. doi: 10.1016/j.autcon.2015.07.015.

Oliver, A. A. and Huber, D. (2010) 'Reconstruction of Wall Surfaces Under Occlusion and Clutter in 3D Indoor Environments', in *Proceedings of 3D Imaging,Modeling, Processing, Visual- ization and Transmission (3DIMPVT)*. Hangzhou, China, pp. 1–33. doi: 10.1109/3DIMPVT.2011.42.

OmniClass (2013) *OmniClass - A Strategy for Classifying the Built Environment*. Available at: http://www.omniclass.org/ (Accessed: 14 March 2018).

opensourcebim (2016) *BIMserver*. Available at: www.bimserver.org.

Oreni, D., Brumana, R., Cuca, B. and Gergopoulos, A. (2013) 'HBIM for conservation and management of built heritage: Towards a library of vaults and wooden beam floors', in *CIPA 2013 XXV International Symposium, ISPRS Annals, Volume 164*, pp. 1–6.

Orlikowski, W. J. and Baroudi, J. J. (1991) 'Studying information technology in organizations: Research approaches and assumptions', *Information Systems Research*, 2, pp. 1–28.

Otto, M. (2018) *Treetagger Python module*. Available at: https://github.com/miotto/treetagger-python (Accessed: 22 November 2018).

Pan, J., Anumba, C. J. and Ren, Z. (2004) 'Potential application of the semantic web in construction', in *Proceedings of the 20th Annual Conference of the Association of Researchers in Construction Management (ARCOM)*. Heriot-Watt University, Edinburgh, UK, pp. 923–929.

Park, J. and Cai, H. (2017) 'WBS-based dynamic multi-dimensional BIM database for total construction as-built documentation', *Automation in Construction*. Elsevier B.V., 77, pp. 15–23. doi: 10.1016/j.autcon.2017.01.021.

Pathirage, C. (2007) *A Structured Approach To Manage The Tacit Knowlwdge of Construction Employees*. University of Salford.

Pauwels, P., Krijnen, T., Terkaj, W. and Beetz, J. (2017) 'Enhancing the ifcOWL ontology with an alternative representation for geometric data', *Automation in Construction*. Elsevier B.V., 80, pp. 77–94. doi: 10.1016/j.autcon.2017.03.001.

Pauwels, P. and Oraskari, J. (2015) *IFCtoRDF converter*. Available at: https://github.com/IDLabResearch/IFC-to-RDF-converter.

Pauwels, P. and Roxin, A. (2016) 'SimpleBIM : From full ifcOWL graphs to simplified building graphs', in *11th European Conference on Product and Process Modelling*, pp. 11–18.

Pauwels, P. and Terkaj, W. (2016) 'EXPRESS to OWL for construction industry: Towards a recommendable and usable ifcOWL ontology', *Automation in Construction*. Elsevier B.V., 63, pp. 100–133. doi: 10.1016/j.autcon.2015.12.003.

Pauwels, P., Zhang, S. and Lee, Y. C. (2017) 'Semantic web technologies in AEC industry: A literature overview', *Automation in Construction*. Elsevier B.V., 73, pp. 145–165. doi: 10.1016/j.autcon.2016.10.003.

Pavan, A. (2017) 'Common Data Environment (CDE). An efficient and effective use in the Italian standard UNI 11337:2017', in Ciribini, A. L. C., Alaimo, G., Capone, P., Daniotti, B., Dell'Osso, G., and Nicolella, M. (eds) *Re-shaping the construction industry*. Santarcangelo di Romagna: Maggioli, pp. 60–69.

Pavan, A., Antisari, C. V., Mirarchi, C. and Rizzarda, C. (2017) 'Design optioneering and BIM : a digital revolution', *Area*, 150, pp. 78–83.

Pavan, A. and Mirarchi, C. (2017) 'Overcome Data Island in the Construction Sector', in Ciribini, A., Alaimo, G., Capone, P., Daniotti, B., Dell'Osso, G., and Nicolella, M. (eds) *Re-shaping the construction industry*. Maggioli, pp. 70–82.

Pavan, A., Odorizzi, P., Bolognesi, C., Napoleone, V., Trebbi, C. and Mirarchi, C. (2017) 'A BIM-based per-checking model for digital building paperwork: consistency administrative, graphics and documents', in *Building Smart International Summit*. Barcelona, Spain. doi: 10.13140/RG.2.2.12715.44327.

Peffers, K., Tuunanen, T., Gengler, C. E., Rossi, M., Hui, W., Virtanen, V. and Bragge, J. (2006) 'The Design Science Research Process: A Model for Producing and Presenting Information Systems Research', *The Proceedings of Design Research in Information Systems and Technology DESRIST 2006*, pp. 83–106. doi: 10.2753/MIS0742-1222240302.

Pellicer, E., Pellicer, T. M. and Catalá, J. (2009) 'An integrated control system for SMEs in the construction industry', *Revista de la Construcción*, 8(2), pp. 4–17.

Petrova, E., Svidt, K., Jensen, R. L. and Pauwels, P. (2018) 'From patterns to evidence: Enhancing sustainable building design with pattern recognition and information retrieval approaches', in Karlshøj, J. and Scherer, R. (eds) *Proceedings of the 12th european conference on product and process modelling (ECPPM 2018), 12-14 september*. Copenhagen, Denmark: CRC Press.

Al Qady, M. and Kandil, A. (2014) 'Automatic clustering of construction project documents based on textual similarity', *Automation in Construction*. Elsevier B.V., 42, pp. 36–49. doi: 10.1016/j.autcon.2014.02.006.

Radulovic, F., Poveda-Villalón, M., Vila-Suero, D., Rodríguez-Doncel, V., García-Castro, R. and Gómez-Pérez, A. (2015) 'Guidelines for Linked Data generation and publication: An example in building energy consumption', *Automation in Construction*. Elsevier B.V., 57, pp. 178–187. doi: 10.1016/j.autcon.2015.04.002.

Rasmussen, E. (1992) 'Clustering Algorithms', in Frakes, W. B. and Baeza-Yates, R. (eds) *Information retrieval: Data Structures and Algorithms*. Englewood Cliffs, NJ: Prentice Hall, pp. 419–442.

Rasmussen, M. H., Hviid, C. A. and Karlshøj, J. (2017) 'Web-based topology queries on a BIM model', in *5th LDAC workshop, 13-15 November*. Dijon, France.

Rasmussen, M. H., Pauwels, P., Hviid, C. A. and Karlshøj, J. (2017) 'Proposing a Central AEC Ontology That Allows for Domain Specific Extensions', in *Lean and Computing in Construction Congress (LC3) - Joint Conference on Computing in Construction (JC3)*. Heraklion, Crete, Greece, pp. 237–244.

Raz, O., Buchheit, R., Shaw, M., Koopman, P. and Faloutsos, C. (2004) 'Detecting semantic anomalies in truck weigh-in-motion traffic data using data mining', *Journal of Computing in Civil Engineering*, 18(4), pp. 291–300.

Redis (2018) *Homepage Redis database*. Available at: https://redis.io/ (Accessed: 7 March 2018).

Redmond, A., Hore, A., Alshawi, M. and West, R. (2012) 'Exploring how information exchanges can be enhanced through Cloud BIM', *Automation in Construction*. Elsevier B.V., 24, pp. 175–183. doi: 10.1016/j.autcon.2012.02.003.

République Francaise (2018) *Kroqui platform*. Available at: https://www.kroqi.fr/ (Accessed: 26 September 2018).

Rezgui, Y. (2006) 'Ontology driven knowledge management using information retrieval techniques', *Computing in Civil Engineering*, 20, pp. 261–270.

Rezgui, Y., Boddy, S., Wetherill, M. and Cooper, G. (2011) 'Past, present and future of information and knowledge sharing in the construction industry: Towards semantic service-based e-construction?', *CAD Computer Aided Design*. Elsevier Ltd, 43(5), pp. 502–515. doi: 10.1016/j.cad.2009.06.005.

Rezgui, Y., Hopfe, C. J. and Vorakulpipat, C. (2010) 'Generations of knowledge management in the architecture, engineering and construction industry: An evolutionary perspective', *Advanced Engineering Informatics*. Elsevier Ltd, 24(2), pp. 219–228. doi: 10.1016/j.aei.2009.12.001.

Rezgui, Y. and Miles, J. (2011) *Harvesting and managing knowledge in construction: from theoretical foundations to business applications*. Spon Press.

RIBA (2013) *RIBA plan of work 2013*, *Royal Institute of British Architects*. UK.

Roos, J., Roos, G., Dragonetti, N. C. and Edvinsson, L. (1998) *Intellectual capital: Navigating in the new business landscape*. New York: New York University Press.

Ruffino, M. (2006) 'Sharing Knowledge between Generations: a Case Study about Managing Apprenticeship in Italian Districts', in Kazi, A. S. and Wolf, P. (eds) *Real-Life Knowledge Management: Lessons from the Field*. KnowledgeBoard - VTT Technical Research of Finland, pp. 183–193.

Sacks, R., Kaner, I., Eastman, C. M. and Jeong, Y.-S. (2010) 'The Rosewood experiment — Building information modeling and interoperability for architectural precast facades.', *Automation in Construction*, 19(4), pp. 419–432.

Sacks, R., Ma, L., Yosef, R., Borrmann, A., Daum, S. and Kattel, U. (2017) 'Semantic Enrichment for Building Information Modeling: Procedure for Compiling Inference Rules and Operators for Complex Geometry', *Journal of Computing in Civil Engineering*, 31(6), p. 04017062. doi: 10.1061/(ASCE)CP.1943-5487.0000705.

Saitta, S., Kripakaran, P., Raphael, B. and Smith, I. F. (2008) 'Improving system identification using clustering', *Journal of Computing in Civil Engineering*, 22(5), pp. 292–302.

Salton, G. (1991) 'Development in Automatic Text Retrieval', *Science*, 253, pp. 974–980.

Sawhney, A., Mukherjee, K. K., Rahimian, F. P. and Goulding, J. S. (2014) 'Scenario thinking approach for leveraging ICT to support SMEs in the Indian construction industry', *Procedia Engineering*, 85, pp. 446–453. doi: 10.1016/j.proeng.2014.10.571.

Scarborough, H., Swan, J., Laurent, S., Bresnen, M., Edelman, L. and Newell, S. (2004) 'Project-based learning and the role of learning boundaries', *Organization studies*, 25, pp. 1579–1600.

Scarborough, H., Swan, J. and Preston, J. (1998) *Knowledge management: a literature review*, *Issues in People Management*. London.

Schank, R. C. (1983) *Dynamic Memory: A Theory of Reminding and Learning in Computers and People*. New York: University Press.

Schank, R. C. (2014) *Inside Case-based Explanation*. New York: Psychology Press.

Scherer, R. J. and Reul, S. (2000) 'Retrieval of project knowledge from heterogeneous AEC documents', in *Proceedings of the ICCCBE-VIII*. Palo Alto, CA, pp. 812–819.

Schneider, G. F. (2017) 'Towards Aligning Domain Ontologies with the Building Topology Ontology', in *5th LDAC workshop, 13-15 November*. Dijon, France.

Scikit-learn (2018) *Scikit-learn packages*. Available at: http://scikit-learn.org/stable/

(Accessed: 27 September 2018).

Sebastiani, F. (2002) 'Machine learning in automated text categorization', *ACM Computing Surveys*, 34(1), pp. 1–47.

Sergey, B. and Lawrence, P. (1998) 'The Anatomy of a Large-Scale Hypertextual Web Search Engine', in. doi: 10.1038/ncomms3193.

Shafiq, M. T., Matthews, J. and Lockley, S. R. (2012) 'Requirements for Model Server Enabled Collaborating on Building Information Models', in *First UK Academic Conference on BIM*. Newcastle upon Tyne.

Shafiq, M. T., Matthews, J. and Lockley, S. R. (2013) 'A study of BIM collaboration requirements and available features in existing model collaboration systems', *Journal of Information Technology in Construction*, 18, pp. 148–161.

Shahandashti, S. M., Razavi, S. N. S. N., Soibelman, L., Berges, M., Caldas, C. H., Brilakis, I., Teizer, J., Vela, P. A., Haas, C., Garrett, J., Akinci, B. and Zhu, Z. (2010) 'Data-Fusion Approaches and Applications for Construction Engineering', *Journal of Construction Engineering and Management*, 137(10), pp. 863–869. doi: 10.1061/(ASCE)CO.1943-7862.0000287.

Shirazi, B., Langford, D. A., & Rowlinson, S. M. (1996) 'Organizational structures in the construction industry', *Construction Management & Economics*, 14(3), pp. 199–212.

Sibenik, G. (2016) 'Building information modelling based interdisciplinary data exchange : a case study', in *BIM Academic Forum, Glasgow*.

Singh, V., Gu, N. and Wang, X. (2011) 'A theoretical framework of a BIM-based multi-disciplinary collaboration platform', *Automation in Construction*. Elsevier B.V., 20(2), pp. 134–144. doi: 10.1016/j.autcon.2010.09.011.

Skyrme, D. J. and Amidon, D. M. (1997) *Creating the knowledge-based business*. London: Business Intelligence.

Soibelman, L. and Caldas, C. H. (2000) 'Project extranets for construction management: the American experience', in *Proceedings of En- tac-2000*. Salvador, Brazil.

Soibelman, L. and Kim, H. (2002) 'Data Preparation Process for Construction Knowledge Generation through Knowledge Discovery in Databases', *Journal of Computing in Civil Engineering*, 16(1), pp. 39–48. doi: 10.1061/(ASCE)0887-3801(2002)16:1(39).

Soibelman, L., Wu, J., Caldas, C., Brilakis, I. and Lin, K. Y. (2008) 'Management and analysis of unstructured construction data types', *Advanced Engineering Informatics*, 22(1), pp. 15–27. doi: 10.1016/j.aei.2007.08.011.

Stato Italiano (2010) *Decreto del presidente della repubblicato del 5 ottobre 2010 n. 207: Regolamento di esecuzione e attuazione del decreto legislativo 12 aprile 2006, n. 163, recante "Codice dei contratti pubblici relativi a lavori, servizi e foniture in attuazione delle Di*.

Stato Italiano (2016) *Decreto Legislativo 28 Aprile 2016, n. 50 - Attuazione delle direttive 2014/23/UE, 2014/24/UE e 2014/25/UE sull'aggiudicazione dei contratti di concessione, sugli appalti pubblici e sulle procedure d'appalto degli enti erogatori nei settori dell'acqua, de*. Italy.

Steinbach, M., Karypis, G. and Kumar, V. (2000) 'A Comparison of Document Clustering Techniques', *KDD workshop on text mining*, 400, pp. 1–2. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4721382.

Stephan, G., Andreaas, A., Völker, J. and Studer, R. (2011) 'Ontologies and the Semantic Web', in Domingue, J., Fensel, D., and Hendler, J. A. (eds) *Handbook of Semantic Web Technologies*. Berlin, Heidelberg: Springer, pp. 507–579.

Sun, Y., Chen, Y., Wang, X. and Tang, X. (2014) 'Deep Learning Face Representation by Joint Identification-Verification', *Advances in neural information processing systems*, pp. 1988–1996. doi: 10.1109/CVPR.2014.244.

Tan, H. C., Carrillo, P. M., Anumba, C. J., Bouchlaghem, N., Kamara, J. M. and Udeaja, C. E.

(2007) 'Development of a Methodology for Live Capture and Reuse of Project Knowledge in Construction', *Journal of Management in Engineering*, 23(1), pp. 18–26.

Tang, P., Huber, D., Akinci, B., Lipman, R. and Lytle, A. (2010) 'Automatic reconstruction of as-built building information models from laser-scanned point clouds: A review of related techniques', *Automation in Construction*, 19, pp. 829–843.

Tao, F., Cheng, J., Qi, Q., Zhang, M., Zhang, H. and Sui, F. (2018) 'Digital twin-driven product design, manufacturing and service with big data', *The International Journal of Advanced Manufacturing Technology*, 94(9–12), pp. 3563–3576. doi: 10.1007/s00170-017-0233-1.

Teicholz, P. (1999) 'Vision of future practice', in *Berkeley-Stanford work- shop on defining a research agenda for AEC process/product devel- opment in 2000 and beyond*. Stanford University, Stanford, California.

Tixier, A. J. P., Hallowell, M. R., Rajagopalan, B. and Bowman, D. (2016) 'Automated content analysis for construction safety: A natural language processing system to extract precursors and outcomes from unstructured injury reports', *Automation in Construction*. Elsevier B.V., 62, pp. 45–56. doi: 10.1016/j.autcon.2015.11.001.

TNO (2018) *BIMbots*. Available at: http://bimbots.org/ (Accessed: 25 September 2018).

Tollmar, K., Sandor, O. and Schömer, A. (1996) 'Supporting Social Awareness @ Work Design and Experience', *Computer Supported Cooperative Work (CSCW)*, pp. 298–307.

Törmä, S. (2013) 'Semantic linking of building information models', in *Proceedings - 2013 IEEE 7th International Conference on Semantic Computing, ICSC 2013*, pp. 412–419. doi: 10.1109/ICSC.2013.80.

Törmä, S. (2017) *Linked Building Data: An overview*. Available at: https://www.w3.org/community/lbd/.

Della Torre, S., Mirarchi, C. and Pavan, A. (2017) 'Il BIM per la conservazione: Rappresentare e gestire la conoscenza', *Ananke*, 82, pp. 108–115.

Trimble (2018) *Trimble Connect*. Available at: https://connect.trimble.com/ (Accessed: 18 March 2018).

Troxler, P. and Lauche, K. (2006) 'Learn, Exchange, Develop', in Kazi, A. S. and Wolf, P. (eds) *Real-Life Knowledge Management: Lessons from the Field*. KnowledgeBoard - VTT Technical Research of Finland, pp. 7–22.

Tversky, A. and Kahneman, D. (1971) 'Belief in the law of small numbers', *Psychological Bulletin*, 76(2), pp. 105–110. doi: 10.1037/h0031322.

Tversky, A. and Kahneman, D. (1973) 'Availability: A heuristic for judging frequency and probability', *Cognitive Psychology*, 5(2), pp. 207–232. doi: 10.1016/0010-0285(73)90033-9.

Tversky, A. and Kahneman, D. (1974) 'Judgment under Uncertainty: Heuristics and Biases', *Science*, 185(4157), pp. 1124–1131. doi: 10.1126/science.185.4157.1124.

UNI (2017a) *UNI 11337- 4 - Building and civil engineering works - Digital management of the informative processes - Part 4: Evolution and development of information within models, documents and objects*. Italy.

UNI (2017b) 'UNI 11337-1:2017 - Edilizia e opere di ingegneria civile - Gestione digitale dei processi informativi delle costruzioni - Parte 1:Modelli, elaborati e oggetti informativi per prodotti e processi'. Ente Italiano di Normazione.

UNI (2017c) 'UNI 11337-4:2017 - Edilizia e opere di ingegneria civile - Gestione digitale dei processi informativi delle costruzioni - Parte 4: Evoluzione e sviluppo informativo di modelli, elaborati e oggetti'. Ente Italiano di Normazione.

UNI (2017d) 'UNI 11337-5:2017 - Edilizia e opere di ingegneria civile - Gestione digitale dei processi informativi delle costruzioni - Parte 5:Flussi informativi nei processi digitalizzati'. Ente Italiano di Normazione.

Vaishnavi, V., Kuechler, B. and Petter, S. (2004) *Design Science Research in Information Systems*. Available at: http://www.desrist.org/design-research-in-information-systems/. (Accessed: 10 February 2019).

Venugopal, M., Eastman, C. M., Sacks, R. and Teizer, J. (2012) 'Semantics of model views for information exchanges using the industry foundation class schema', *Advanced Engineering Informatics*, 26(2), pp. 411–428.

Vicknair, C., Macias, M., Zhao, Z., Nan, X., Chen, Y. and Wilkins, D. (2010) 'A comparison of a graph database and a relational database', in *Proceedings of the 48th Annual Southeast Regional Conference on - ACMSE*. Oxford, MS, USA, p. 1. doi: 10.1145/1900008.1900067.

Volk, R., Stengel, J. and Schultmann, F. (2014) 'Building Information Modeling (BIM) for existing buildings — Literature review and future needs', *Automation in Construction*. Elsevier B.V., 38, pp. 109–127. doi: 10.1016/j.autcon.2013.10.023.

Vygotsky, L. (1978) *Mind in Society: the development of higher psychological processes*. Cambridge: Cambridge University Press.

W3C (2018) *Linked Building Data Community Group*. Available at: https://www.w3.org/community/lbd/ (Accessed: 22 March 2018).

W3C OWLWG (2009) *OWL 2 Web Ontology Language Document Overview*. Available at: https://www.w3.org/TR/2009/WD-owl2-overview-20090327/ (Accessed: 21 March 2018).

Walton, J. and Whicker, L. (1996) 'Virtual Enterprise: Myth and Reality', *Journal of Control*, 22(8), pp. 22–25.

Wang, K. C., Wang, W. C., Wang, H. H., Hsu, P. Y., Wu, W. H. and Kung, C. J. (2016) 'Applying building information modeling to integrate schedule and cost for establishing construction progress curves', *Automation in Construction*. Elsevier B.V., 72, pp. 397–410. doi: 10.1016/j.autcon.2016.10.005.

Wang, R. Y. and Strong, D. M. (1996) 'Beyond Accuracy: What Data Quality Means to Data Consumers', *Journal of Management Information Systems*, 12(4), pp. 5–33. doi: 10.1080/07421222.1996.11518099.

Watson, A. (2011) 'Digital buildings - Challenges and opportunities', *Advanced Engineering Informatics*, 25, pp. 573–581.

WebFinance Inc (2018a) *The Business Dictionary*. Available at: http://www.businessdictionary.com (Accessed: 17 March 2018).

WebFinance Inc (2018b) *The Business Dictionary - law of small numbers definition*. Available at: http://www.businessdictionary.com/definition/law-of-small-numbers.html (Accessed: 17 August 2018).

Wetherill, M., Rezgui, Y., Boddy, S. and Cooper, G. (2007) 'Intra and inter-organisational knowledge services to promote informed sustainability practices', *Computing in Civil Engineering*, 21, pp. 78–79.

Whitley, E. (2000) 'Tacit and Explicit Knowledge: Conceptual Confusion around the Commodification of Knowledge', in *Knowledge Management: Concepts and Controversies conference, University of Warwick*.

Wikidata (2018) *Wikidata*. Available at: https://www.wikidata.org/wiki/Wikidata:Main_Page.

Wilkinson, P. (2005) *Construction Collaboration Technologies, The Extranet Evolution*. Taylor & Francis.

Willett, P. (1988) 'Recent trends in hierarchical document clustering: A critical review', *Information processing & management*, 24(5), pp. 577–597.

Winter, R. (2008) 'Design science research in Europe', *European Journal of Information Systems*, 17, pp. 470–475. doi: 10.1057/ejis.2008.44.

Wolf, P. and Kazi, A. S. (2006) 'Communities of Practice - A Case Study from the Automotive

Industry', in Kazi, A. S. and Wolf, P. (eds) *Real-Life Knowledge Management: Lessons from the Field*. KnowledgeBoard - VTT Technical Research of Finland, pp. 49–61.

Wong, J., Wang, X., Li, H., Chan, G. and Li, H. (2014) 'A review of cloud-based BIM technology in the construction sector', *Journal of Information Technology in Construction*, 19(September), pp. 281–291.

Wong, J. and Yang, J. (2010) 'Research and Application of Building Information Modelling (BIM) in the Architecture, Engineering and Construction (AEC) industry: A review and direction for future research.', in Loughborough University, U. K. and Pennsylvania State University (eds) *Proceedings of the 6th International Conference on Innovation in Architecture, Engineering & Construction (AEC)*, pp. 356–365.

Wood, W. H. (2000) 'The development of modes in textual design data', in *Proceedings of the ICCCBE-VIII*. Palo Alto, CA, pp. 882–889.

Working Group 2 (2017) *Digitising European Industry Working Group 2 Digital Industrial Platforms*.

World Wide Web Consortium (W3C) (2015) *Semantic Web: inference*. Available at: http://www.w3.org/standards/semanticweb/inference (Accessed: 31 August 2018).

X Commissione, P. (2016) 'Indagine conoscitiva su Industria 4.0: quale modello applicare al tessuto industriale italiano. Strumenti per favorire la digitalizzazione delle filiera industriali nazionali - X Commissione Permanente'.

Xiong, X. and Huber, D. (2010) 'Using Context to Create Semantic 3D Models of Indoor Environments', in *Procedings of the British Machine Vision Conference (BMVC)*, p. 45.1-45.11. doi: 10.5244/C.24.45.

Yalcinkaya, M. and Singh, V. (2015) 'Patterns and trends in Building Information Modeling (BIM) research: A Latent Semantic Analysis', *Automation in Construction*. Elsevier B.V., 59, pp. 68–80. doi: 10.1016/j.autcon.2015.07.012.

Yu, W. Der and Hsu, J. Y. (2013) 'Content-based text mining technique for retrieval of CAD documents', *Automation in Construction*. Elsevier B.V., 31, pp. 65–74. doi: 10.1016/j.autcon.2012.11.037.

Yurchyshyna, A. and Zarli, A. (2009) 'An ontology-based approach for formalisation and semantic organisation of conformance requirements in construction', *Automation in Construction*. Elsevier B.V., 18(8), pp. 1084–1098. doi: 10.1016/j.autcon.2009.07.008.

Zaniolo, C. (1984) 'Database relations with null values', *Journal of Computer and System Sciences*, 28(1), pp. 142–166. doi: 10.1016/0022-0000(84)90080-1.

Zhong, B. T., Ding, L. Y., Luo, H. B., Zhou, Y., Hu, Y. Z. and Hu, H. M. (2012) 'Ontology-based semantic modeling of regulation constraint for automated construction quality compliance checking', *Automation in Construction*, 28, pp. 58–70. doi: 10.1016/j.autcon.2012.06.006.

Zou, Y., Kiviniemi, A. and Jones, S. W. (2017) 'Retrieving similar cases for construction project risk management using Natural Language Processing techniques', *Automation in Construction*. Elsevier, 80, pp. 66–76. doi: 10.1016/j.autcon.2017.04.003.