CODIFICATION CHALLENGES FOR DATA SCIENCE IN CONSTRUCTION

Ranjith K. Soman
PhD Student, Centre for Systems Engineering and Innovation, Department of Civil and Environmental Engineering, Imperial College London, London, SW7 2AZ, UK
(corresponding author). Email: ranjithks17@imperial.ac.uk
https://orcid.org/0000-0003-3967-9121

Jennifer K. Whyte, PhD, FICE
Laing O'Rourke/Royal Academy of Engineering Professor in Systems Integration and Director, Centre for Systems Engineering and Innovation, Department of Civil and Environmental Engineering, Imperial College London, London, SW7 2AZ, UK. Email: j.k.whyte@imperial.ac.uk
http://orcid.org/0000-0003-4640-2913

Abstract: New forms of data science, including machine learning and data analytics, are enabled by machine-readable information but are not widely deployed in construction. A qualitative study of information flow in three projects using Building Information Modelling (BIM) in the late design and construction phase is used to identify the challenges of codification which limit the application of data science. Despite substantial efforts to codify information with 'Common Data Environment (CDE)' platforms to structure and transfer digital information within and between teams, participants work across multiple media in both structured and unstructured ways. Challenges of codification identified in this paper relate to *software usage* (interoperability, translation, modelling, and file-based sharing), *information sharing* (unstructured information, document control, workarounds, process change, and multiple CDEs), and *construction process information* (loss of constraints and low level of detail). This paper contributes to the current understanding of data science in construction by articulating the codification challenges and their implications for data quality dimensions, such as accuracy, completeness, accessibility, consistency, timeliness, and provenance. It concludes with practical implications for developing and using machine-readable information and directions for research to extract insight from data and support future automation.

Keywords: Building information modelling (BIM), Codification; Artificial Intelligence (AI); Automation; Data science; Machine readability; Construction.

1. Introduction

Machine-readable information is enabling new forms of data science, including machine learning and data analytics. These methods offer value to the construction sector through resource and waste optimization, data-driven design, prescriptive analytics for rule checking, visual analytics, performance predictions, operational analytics, and more (Bilal et al., 2016). Yet the construction sector is not taking advantage of these developments as data science is not widely deployed. Bilal et al. (2016) have identified poor data management as one of the main factors which limit the application of data science in construction. The sector is actively trying to overcome this limitation through Building Information Modelling (BIM), an approach to incrementally building structured (and, hence, machine-readable) information throughout the project cycle (Eastman et al., 2008; Jordani, 2010), and where it is machine readable, such structured information could support the use of data science. Recent ideas such as the 'digital twin' are also predicated on the availability of such structured information (both geometries and behaviours; see Bolton et al., 2018). Research has begun to develop approaches to, and document practices of, codifying information (to convert it into a structured and machine-readable format) to improve delivery practices (e.g. in relation to the construction phase (Goedert and Meadati, 2008), workspace planning (Choi et al., 2014), construction defects (Kwon et al., 2014), and lessons learned in a project (Oti et al., 2018)). These studies have advanced knowledge regarding frameworks and methods to codify construction information. In addition to these studies, there is a need for work to extend understanding of the issues which prevent codification of information to improve uptake of data science in construction.

By examining information sharing across projects which use BIM in the late design and construction phase, this paper aims to identify the codification challenges which arise in practice by examining information sharing across projects that use BIM. In this paper, codification is defined as the process of conversion of information into a structured and machine-readable format to support the application of data science. Information refers to the collection of data contextualized with relevant schema and semantics so that insights can be made from the data. The codification challenges are the issues which reduce the machine readability and quality of construction information and, in turn, limit the uptake of data science in the sector. Inspired by work which frames BIM use as a complex social activity (Cao et al., 2014; Dossick and Neff, 2010), this paper builds on and contributes to strands in the literature focused on data quality, machine readability, and BIM adoption and implementation. Data analyses suggest codification challenges which are organizational as well as technical in nature, relating to software use, information sharing, and construction process information.

To develop the contribution, the rest of this paper is divided into four sections. Section 2 provides a brief background on data quality and machine readability in projects and organizational issues associated with BIM adoption and data quality. Section 3 describes the cases and method used in the study. Section 4 presents the findings. These findings are discussed further in relation to the literature in section 5, and the conclusions are presented in section 6.

2. Background

As the construction sector becomes increasingly digital, most information is stored digitally and is accessible through servers or a common data environment (CDE) held in firms or projects (British Standards Institution, 2018; Preidel et al., 2016). However, being digitally accessible does not mean that the information is machine readable as the semantics may not be embedded in the data (Hendler and Pardo, 2012). Semantics could be derived from the data using advanced machine learning techniques such as natural language processing and deep learning (Carrillo et al., 2011; Wang, 2017). Nonetheless, this is a resource-intensive process which requires training models for achieving satisfactory accuracy and has costs associated with it. The result of this process may also not be of high-quality (Wang, 2017). Despite the existence of such data-cleaning algorithms across sectors, poor data quality is costing $3.1 trillion in the United States (Quintero et al., 2015). In addition, the poor quality of data is increasing operational costs, decreasing revenue, and resulting in missed commercial opportunities (Loshin, 2010). Within construction, Sacks et al. (2017) have described how the quality of input information influences semantic enrichment of BIM when using machine learning. Moreover, Farias et al. (2018) have shown the effects of poor quality data resulting in wrong inferences when they tried to extract building views using a rule-based method. Whyte et al. (2016) articulated how managing change in large datasets becomes a focus in an era of 'big data' in which project information is increasingly characterized by volume, velocity, and variety. Recent work has further characterizes such data as also including characteristics of veracity and value (e.g. Younas, 2019). These issues of data quality occur because construction data is heterogeneous, and its veracity is not always known. Data cleaning related to the variety (heterogeneity) and veracity characteristics of big data is especially difficult when compared to data cleaning related to other characteristics such as volume and velocity (Fan, 2015; Janssen et al., 2017). Therefore, there is a need to keep the data of the highest quality and in a machine-readable format (maintaining the data relationships) as far as possible to have the best inferencing.

2.1 Quality and machine readability

What constitutes a good quality dataset? According to Wang and Strong (1996), a good quality dataset is the one that has enough information embedded in it for a particular use by the user. Researchers have set out multiple dimensions to assess the data quality concerning big data analytics (Batini and Scannapieco, 2016; Cai and Zhu, 2015; Delone and McLean, 2014; Naumann and Rolker, 2000; Wang and Strong, 1996). For this paper, the focus is on the following data quality dimensions based on Batini and Scannapieco (2016) as they best reflect the implications of codification challenges. *Accuracy* is the closeness of the measured/represented data and reality. There are two kinds of accuracies: semantic and syntactic. Semantic accuracy relates to the closeness of the data value to reality, whereas syntactic accuracy refers to the closeness of the data representation to the expected data type/model. *Completeness* is the measure of information content present in the data compared to the extent of information content required to be present in the data to perform a particular task. *Temporal* dimensions refer to currency, volatility, and timeliness. *Currency* relates to the promptness of data updates. *Volatility* refers to the frequency at which the data variance occurs. *Timeliness* refers to the suitability of the current data to perform a task. *Consistency* refers to uniformity and constancy of data with respect to the semantic rules defined over multiple data items. *Accessibility* refers to the ability of data to be accessed by a user (human user or computer program) and generate information from it. *Data provenance* is the description of the origins of data and the process by which it is manipulated. Jayawardene et al. (2015) have conducted a systematic literature review on the extensive data quality dimensions and consolidated overlapping dimensions of quality. This means that some of the data quality dimensions are interdependent. For example, the data quality dimension related to semantic accuracy might depend on the timeliness dimension as the data may be accurate with respect to time. However, the same value may be inaccurate at a different time. At the outset, a dataset is considered to be of good quality when the measure of these dimensions is high, leading to better inferences.

What constitutes machine-readable data? Data in a structured format that could easily be processed by computers are considered as 'machine-readable'. Berners-Lee (2006) has stated a set of' 'rules' for creating structured data so that it can be connected and interpreted easily by machines. The first one is to index the data to make it digitally accessible by storing it on online servers so that it can be easily accessed by computers as well as people. This relates to the accessibility dimension of data quality. Indexing the data and storing it online increases the accessibility as it is easy to find. The second rule is to structure the data with relevant schemas for easy interpretation by machines. This step makes the data structured such that semantic relations are embedded within it, resulting in better inferencing and, thus, improving the syntactic accuracy of the data. The third rule is to make the schemas public and machine readable by using open-source schemas to describe the data model so that interpretations can be made by computers without any proprietary data interfaces. Proprietary data formats limit data inferencing as the schema by which data is modelled is only accessible to few applications. Therefore, using open-source schema would increase the measure of accessibility dimension as more applications can use the open schema to derive the context of data for inferencing. The last rule is to link the data with other datasets so that better inferences can be made by deriving the context information. This improves the consistency dimensions of data quality as the same data is linked to multiple datasets. Linking would ensure that there are no conflicts in the data about a concept stored in multiple databases. Based on these rules, structured information can be classified into five types in the increasing order of machine readability, as shown in Table 1. Increasing the machine readability of the data, in turn, increases the data quality as the dimensions relating to accessibility, accuracy, and consistency are improved in the process.

<<Insert Table 1>>

2.2 Limits of existing research

What is limiting the generation of good quality machine-readable information within the sector? It does not appear to be technical development as novel technical solutions are being developed by construction informatics researchers with a focus on the integration of data in the sector, for example, through the use of data standards (Krijnen and Beetz, 2017; Pazlar and Turk, 2008), cloud-based BIM (Beetz et al., 2010; Singh et al., 2011; Zhang et al., 2017), and linked-data technologies (Kim et al., 2018; Pauwels et al., 2015; Pedro et al., 2017, Zhang and Beetz, 2015). It does not appear to be policy interventions either. Standards and public mandates have placed BIM at the heart of the information management required to coordinate processes in project delivery and operation of infrastructure, making BIM central to digital tools and workflows in projects (British Standards Institution, 2018; Sacks et al., 2018). Instead, the literature suggests that the issues may be both organizational and technical in nature.

Prior research has assessed BIM adoption across different markets, using a model of diffusion area, macro-maturity components, macro-diffusion dynamics, and so forth, and validated this model by applying it to assess BIM adoption amongst 21 countries (Kassem and Succar, 2017). This work has determined the BIM project objectives, critical success factors, and operative critical success factors for effective implementation of BIM (Chegu, Badrinath, and Hsieh, 2019). It has identified the success factors for adoption of BIM in a company, selection of projects within the company to implement BIM, and selection of BIM services and software (Won et al., 2013). It has surveyed the degree of implementation of BIM statistically by evaluating the level of BIM implementation and quality of collaboration and communication in BIM-enabled projects, and linking discussing its influence on uptake of integrated delivery systems (Chang et al., 2017), and developing strategies for using BIM to reduce rework in construction (Hwang, et al., 2019) and improve collaboration through the development of BIM-based platforms by analyzing requirements and details of elements needed for a collaborative work model (Zhang et al., 2017). In addition, Gu and London (2010) have created a collaborative BIM decision framework to facilitate BIM adoption through a four-part method. The framework first defines the scope, purpose, roles, relationships, and project phases, followed by developing a work process roadmap, identifying the technical capabilities and the limitations of tools, and finally customizing these to suit the capabilities and skillsets of the project team. Building on this study, Singh et al. (2011) have determined technical requirements for a BIM server to serve as a collaboration platform. These studies have extended the existing knowledge base, and they give a deeper understanding of the problems associated with BIM implementation, and suggested steps for the effective implementation of BIM. However, data quality issues emerging from the problems associated with BIM implementation is relatively less studied.

Previous research on data quality within the construction sector has studied semantic and syntactic accuracy of BIM, BIM quality assurance processes for the design stage and data quality issues in the design model, and completeness of information in BIM for facility management. Solihin et al. (2015) have identified requirements for good quality for BIM in an Industry Foundation Class (IFC) format. Building on this study, Lee et al. (2018) have presented a semantic rule-checking process to ensure data quality pertaining to semantic and syntactic accuracy is maintained whilst BIM in an IFC format is exchanged. In another study, the quality of the information in the design phase was assessed using a structured and quantifiable process based on a BIM quality assurance by Donato et al. (2017). Mirarchi and Pavan (2019) have analyzed the data quality issues concerning accuracy, consistency, and completeness dimensions of the BIM models created during the design. For facility management, Zadeh et al. (2017) have proposed a framework to assess the quality of BIM, focusing on the completeness dimension. These studies have advanced the knowledge on data quality issues associated with BIM models. Aa limitation of these studies is that they do not address the data quality issues emerging due to the wider practice of using document-based and model-based information sharing in the sector.

To address the data quality issues emerging due to such document-based and model-based information-sharing practices, it is necessary to study information flow across project teams in detail. By better characterizing the practice, such empirical work can then inform future technical developments (e.g. Hartmann, 2008) and address challenges raised in prior work in areas such as automated scheduling (e.g. Han and Golparvar-Fard, 2017). Previous research on the use of BIM in organizations articulates antecedents to BIM uptake (Taylor, 2007) and identifies organizational issues which affect BIM implementation, such as organizational divisions (Dossick and Neff, 2010). It describes how practices are always ‘hybrid’, overlaying a range of old and new media and processes (Harty and Whyte, 2010; Whyte, 2011), with the roles of construction professionals also evolving (Akintola et al., 2017; Jaradat et al., 2013; Sebastian, 2011). Such work draws attention to the organizational factors associated with information use, whereby technological integration cannot be assumed to foster closer collaboration across companies (Dossick and Neff, 2010). This literature, which understands BIM use as a complex socialized activity (Cao et al., 2014), provides an approach that can be used to study the codification challenges and design quality issues which emerge in leading practice.

3. Research method

To understand the challenges of codification in construction, three construction projects are studied qualitatively to investigate the digital tools and workflows used in the projects, structured and unstructured information flows in these projects, and the problems associated with the information flow. These three leading projects are a multi-storey residential student apartment block in the United Kingdom (Case 1), a metro rail infrastructure project in India (Case 2), and a major water infrastructure megaproject in the United Kingdom (Case 3). The multi-storey residential apartment project (Case 1) is an exemplary project exhibiting the use of BIM in the United Kingdom, constructed by a leading contractor and using state-of-the-art offsite manufacturing approaches in construction. The metro rail project in India (Case2) is pioneering BIM implementation amongst the metro projects in India, incorporating learning on digital implementation from global megaprojects. The water infrastructure project (Case 3) is one of the biggest construction projects in the United Kingdom, using innovative technological solutions to futureproof construction and deliver a physical asset as well as a digital asset for operation. Early in the study, the first author, who collected the data, also visited another infrastructure project and a commercial retrofit project, and these three projects were then chosen due to their significance and because they use a level of digital collaboration categorized as BIM level 2. These projects follow the BIM level 2 recommendations set out by the mandate (it is mandated in Case 3 and seen as best practice in the other cases). Although the metro rail project (Case 2 is in a country which does not have a regulatory framework for BIM level 2 recommendation, the owner required the adoption of BIM level 2 as international best practice, which justifies our choice for selecting the case. Information sharing across project teams in the late design and construction phases of the three projects is studied qualitatively by visiting the projects, analyzing internal and publicly available documents, observing meetings, and conducting informal and formal interviews on the use of product and process information during the construction stage (refer Appendix 1).

Within each project, there was an initial setup meeting to present the study and identify interviewees. The interview protocol covered questions of communication, software tools used, BIM, collaboration, and information flow. The data analysis phase overlapped with the data collection phase. The taped interviews were transcribed, and field notes were typed up. These were read and reread between the project meetings. Summaries of interviews were sent back to the interviewees for member checks. All the data was organized into cases and stored into the qualitative analysis software. These methods draw on a qualitative case study approach (Eisenhardt and Graebner, 2007), building insights across the three cases from multiple sources (site visits, documents, field notes, and interview transcripts).

- Multi-storey residential student apartment block in the United Kingdom (Case 1): To study this case, the first author visited the construction site and offices of the projects, had informal conversations with the digital and planning engineers, examined construction documents and models, and studied the software tools used to understand the information embedded in the BIMs, construction schedules, and other reports such as design calculation, method statements, and so on. These documents were centrally hosted on a CDE, and the first author had access to it whilst being there at the office.
- Metro rail infrastructure project in India (Case 2): To examine this case, documents such as the BIM execution plan, presentation documents for training, and press releases were studied to understand digital information management practices. The project manager, chief site engineer, casting yard engineer, and BIM consultants, who form a cohort of the key decision makers during the construction stage, were interviewed informally to get an insight into the extent of codification in the information flow during their daily work practices. Field notes were taken during the interview. In addition to these interviews, the casting yard, viaduct construction site, and a station site were visited to understand the on-ground practice of various activities. Further insight into this case was obtained through a workshop, co-organized by the authors, with 40 participants, including client representatives of six major Indian metro-rail projects along with technology providers and delivery teams. The workshop provided a perspective on the digitization of this project in the broader landscape of Indian metro rail construction.
- Water infrastructure project in the United Kingdom (Case 3): To understand the codification challenges in this case, eight semi-structured interviews were conducted. All eight interviewees had more than ten years of experience in the construction sector and had worked with different major projects in the United Kingdom and abroad. The interviewees' areas of expertise covered design, planning, project engineering, digital engineering, prefabricated construction, and information management. All interviewees had teams working

with them on their areas of specialization and also interacted with the other stakeholders in the project. These characteristics make them ideal for case-based research. Following the semi-structured approach ensured participants would talk broadly on their experiences with information flow using digital collaboration tools. Seven out of the eight interviews were recorded, and transcripts were made from the recordings. In addition to the interviews, the first author conducted multiple visits over two weeks to the project office, observing meetings and the work practice. The first author also had access to CDE and documents such as a construction programme, look-ahead schedule, and method statements.

Data analysis took place in three steps. First, each of the cases was separately analyzed. Second, the cases were compared and contrasted. The initial analyses were conducted during data collection, so early analyses focused and directed later data collection. The 'within case' analyses and the 'cross-case comparisons' also led to an iteration between these steps of data analysis as the comparison across cases was instructive in directing analytic attention within cases. Finally, the third step was a more in-depth analysis of the third case study (for which there was more detailed information).

A starting assumption of this current study is that there are data quality issues caused by the way digital tools and workflows are used in late design and construction stages. This research thus addresses the questions: How do codification challenges arise because of the different digital workflows and working practices across projects? How do these lead to data quality issues? 'Within case' analysis of the multi-storey student apartment raised issues of data interoperability, information loss, use of 2D CAD, and lack of detail in the schedule. In the metro rail project, the issues of data interoperability, use of 2D CAD, and lack of detail in the schedule were also present, but there were also issues of unstructured communication channels, document control, and lack of skills to adopt digital technologies. In the water infrastructure project, additional issues were identified concerning problems with CDE, lack of process codification, and long processing times. In the water infrastructure project, the design workflow, work package plan, construction programme, BIM models, and drawings in the CDE were studied to understand the level of detail and machine readability of the documents. Coding was done on the field notes and interview transcripts to identify different issues related to codification and information sharing. The software was used to track the patterns emerging from these data. These codes were organized to find themes. The identified themes were then analyzed based on the data quality dimensions to understand their implications about data quality.

4. Codification challenges in construction

Table 2 summarizes the codification challenges observed from studying the projects. Low machine readability of data is a significant challenge for codification, which was observed across the projects. Product information is well codified through BIM, CAD drawings, analysis models, and so on in all the projects. However, the codified information is distributed amongst different formats and databases, limiting the application of analytics. In addition, multiple modes of communication, multiple CDEs, and lack of process change also limit the codification of information in the projects studied. Different codification challenges observed in the cases have been mapped in Table 2. These topics are discussed in detail in this section.

<<Insert Table 2 here>>

4.1 Software usage

This section presents the codification challenges related to software usages such as interoperability, information loss during conversions, and multiple modelling techniques during the late design and construction phase. The implication of these challenges on the data quality with respect to dimension accuracy, completeness, accessibility, and data provenance is explained in this section.

4.1.1 Interoperability

Interoperability was raised as a central problem by the interviewees, especially when working with multiple CAD tools in projects, resulting in data loss during format conversions. Even within the same software environment, there are problems related to data compatibility whilst working between different software versions.

> *"Sometimes the drawings where I am using this MicroStation, but sometimes they were drafted from the client, let's say, in Autodesk. And transferring things from Autodesk to MicroStation, you lose data […] the 2019 will open the 2018, 2017, and 2016. But when the 2020 comes into play, then you cannot open it anymore with the 2019 files that are generated with the 2020" (Technical manager, C3I5)*

Here, the drawings are made using different CAD tools such as AutoCAD and MicroStation. However, the data may not be opened (or edited) using the same tools with which they were created. This creates problems of interoperability and loss of information when data in one format is converted to another. Similar problems also occur when using multiple versions of the same tool.

This issue was observed in all three cases but at different scales. Uneven distribution of tools would result in issues of data interoperability. The student apartment (Case 1) and the metro project (Case 2) had predominantly used tools from a single software vendor (Autodesk for the student apartment and Bentley for the metro project) for executing most of their tasks, resulting in better interoperability when compared to the water project (Case 3), which uses tools from different software vendors. The scale of the project has an influence on this diverse distribution of software use. The scale of the student apartment was smaller than that of the metro and water projects, with a leading firm involved in both design and construction, resulting in evenness of software usage (contractor office visit, C1S1). Even though the metro project had different firms engaged in design and construction, the information management was handled by an owner support organization, resulting in evenness in the data (BIM consultant 1, C2I4). The water project, on the other hand, had multiple firms working on different phases of the projects with their own sets of tools, resulting in issues of data interoperability (digital engineer, C3I3).

Data interoperability is a significant problem when it comes to data quality and machine readability. If the data is locked to a proprietary format, it limits the application of data science. Software vendors provide a proprietary Application Program Interface (API) to access the data. However, access to the data through APIs is limited, and information access is limited to proprietary domain-specific programs. This limits data science as different systems cannot talk to each other and derive the context from the information. Although there are open-data formats available for the exchange of information, these are not often used in the construction phase and are only submitted at specific data drops. Evidence also shows that there is a loss of information when converting between formats because of inefficient exporters and importers. This problem reflects data quality related to the accessibility dimension. As long as this data remains inaccessible, algorithms make inferences with limited data, resulting in incomplete inferences.

4.1.2 Information loss during the conversion

For structured information flow through a CDE, the files are converted to a PDF format. The original file may be uploaded as a supporting document, but this is not a necessary requirement.

> *"If we're conveying CAD information, it's being uploaded as a supporting file. There's a facility in [CDE1] that when you upload a PDF, you can also upload a secondary file." (Information manager, C3I4)*

The student apartment and the water project followed a workflow which required documents to be uploaded as a PDF to the CDE, resulting in the loss of information during conversions (access to CDE C1D4, access to CDE1, C3D2). However, the metro project used a workflow without this requirement, resulting in retaining the information (digital project management, C2D2).

The conversion from native formats to a PDF format results in the loss of semantic relationships embedded in the file. The loss of the semantic relationship between datasets results in data silos and limits machine readability. This aspect of the loss of information results in incomplete information and lowers the data quality related to the completeness dimension. Furthermore, during the conversions, the metadata related to the original file is lost. This lowers the data quality dimension associated with data provenance.

4.1.3 Multiple modelling techniques

Software tools allow different methods for creation of information. However, not all methods lead to the information being reusable. The modelled information would have multiple uses, which may not be known to the creator of information.

> *"So, if they use the wrong tool to model something, you don't have the appropriate dataset to it [...] when you go into your authoring tool, do I use a slab tool or do I go and use a generic solid modelling tools then try and attach a dataset to it[…]if they're not, therefore we have to go in there and say, well I can't just say there's a slab now, that's just a piece of geometry" (Digital engineer, C3I3)*

For example, a slab could be modelled as a generic solid model with a dataset attached to it or as a slab component. From the human point of view, the information contained in both models is the same. However, during automated quantity take-off, the slab modelled as a generic solid would not be considered as the computer cannot classify it as a slab.

> *"For tunneling purposes, when you try to extract 2D drawings from 3D BIM models, those drawings are not as correct and as detailed as they used to be. They have glitches, they have errors" (Technical manager, C3I5)*

This issue was observed predominantly in the water project. Limited observation of this issue in the student apartment and the metro project can be attributed to the absence of multiple firms in modelling the data.

Organizational divisions in large projects lead to lower machine readability and data quality because of differences in modelling approaches and software tools used. Software tools offer different approaches to model the same information at the same level of detail. Moreover, modelling approaches used by the firms are guided by the norms and practices followed in the firm. These norms may be different for the firms who use the data. Whilst examining the cases of the student apartment and metro project, where the information modelling is performed by a single firm, the issues such as interoperability and improper modelling of information were limited. However, in the water project, where the information modelling spanned over different firms, there was evidence of issues of interoperability of data and improper information modelling. Therefore, the information created could be used as a digital submission but limits further use. Even when the information is present in the model in the correct format, the level of detail of the modelled information is less than desired, making the information not fit for further use. This problem reduces the quality of the data concerning accuracy and completeness. The fact that the data exists but not in the way it was supposed to be a case of syntactic inaccuracy.

4.2 Information sharing

This section presents the codification challenges in construction, such as unstructured information sharing, drawing and file-based sharing, document control issues, and lack of process change. The implications of these challenges on the data quality dimensions are presented in this section.

4.2.1 Unstructured information sharing

Information shared over modes such as meetings, reports, e-mails, etc., contains relevant data for decision making. This information is embedded in documents, is shared in a human-readable format (documents, PowerPoint presentations, drawings, etc.), and is in formats which are both human and machine readable (e.g. spreadsheet, BIM, etc.). The main limitation of the information shared through these unstructured channels is its accessibility, which is limited to people involved in the meeting or e-mail conversation.

> *"A guy made a design on a spreadsheet for quantities. Some people knew about it; he logged it as well. And I didn't know that at all. So, at the very end of the day, on the eleventh hour when I have finished everything, by the way, we have this spread sheet, and it's exactly what I wanted to do" (Technical manager, C3I5)*

In this case, the information required for a task was already available. However, it was not accessible for the person who needed it, who was not part of the group within which the information was shared. This led to the recreation of the information and loss of productive time. The unstructured information-sharing issue was observed in all cases. The office visits in the student apartment (C1S1), meetings with project participants in the metro project (C2I1, C2I2, C2I3), and interviews with participants in the water project (Table 3) revealed the problem of unstructured information in the projects.

<<Insert Table 3>>

Even when there are structured workflows for information sharing, project participants find it easier to use the unstructured channels of communication. They often find structured information flow through the CDE slow and complicated. Despite being more traceable and accountable compared to unstructured information-exchange practices, the complexity of the new structured methods for information sharing and the poor understanding of workflows across the teams lead to the use of a combination of structured and unstructured channels for information sharing. This issue, however, has an implication on data quality, lowering it with respect to the accessibility dimension since data is not available in a common repository. Instead, it is distributed in different silos and e-mail databases, and the access is limited. In addition, it introduces inconsistency as the same information is distributed amongst different databases which are not connected or synchronized. Tracing the source of the data and its history is also difficult when using unstructured channels for information sharing, thereby reducing the quality of data associated with the provenance dimension. These issues limit the data science as the datasets for drawing inferences are siloed and disconnected.

4.2.2 Drawings and file-based sharing

Even though the projects follow BIM level 2, the engineers interviewed are more comfortable performing submissions and approvals using drawings rather than model-based information sharing. This is mainly because they find it intuitive to use drawings.

> *"I use, I'm not very good at, but I use all the navigator tools that we've got here. But I prefer to use AutoCAD because I find it a lot easier" (Project engineer, C3I1)*

Despite having a BIM coordination tool, Bentley Navigator, the project engineers resort to using the CAD tool because they find it easier. In the contractor's offices of the projects studied, the engineers had drawings on their desks and the CAD software opened on the monitors. If they find errors in the drawings, they modify and edit them first on paper and then on the computer.

> *"Because I have to open up every drawing individually and print them all, or even if you do it the other way around, it's very slow anyway. And then, once I've printed them, reviewed them all, you could do it on there [computer] but it's not the best way because we haven't got the technology. I haven't got a big screen" (Project engineer, C3I1)*

The engineers use laptops with small screens. Some of the hot desks have an additional screen; however, these were also small (less than 23 inches). This is highly inconvenient when reviewing large drawings as they pan and zoom to detect mistakes. This issue was also observed in the student apartment and metro projects.

During the visit to the contractor's office (C1S1) in the student apartment, the first author observed multiple discussions between engineers using drawings as a common representation medium. On the site, a tablet-based application was used to open the drawing. Similarly, in the metro project, the workflow for design, review, and approval (C2D2) presents how drawings are reviewed and processed using the CDE. Such evidence points towards the drawing and file-based information sharing in the construction phase.

The file-based sharing impacts the data quality dimensions associated with temporality and consistency. Most of the file-based manipulations happen within the computer and are uploaded only when complete. Hence, there is a mismatch between the rate of the volatility of data and currency of data. The volatility is high as the data is manipulated on the users' desks (for example, they are manipulating the information in a printed drawing). However, the currency of the data is low as it is uploaded as a batch. That means the data is updated at a lesser speed than it is varied. This has an implication on the data quality associated with the timeliness dimension as the data in the CDE is not the latest version. File-based sharing impacts consistency too. The files act as individual entities and have information from related files in them. Unlike a model, this information is not connected. Therefore, when the source is updated, the information in the file may not be updated, which introduces inconsistency problems.

4.2.3 Document control bottlenecks

Document control plays a vital role in the flow of structured information in the projects, and document control professionals are tasked with managing the access, version control, and availability of documents. Before being uploaded to the CDE, such documents must be approved by the relevant authority (depending on the document). The quantity of documents uploaded to CDEs in the projects studied is enormous, and in each project, processing information becomes a significant task, with bottlenecks in the process leading to workarounds and data quality issues. As the authorization of documents is limited to specific individuals, they tend to get asked for a huge volume of authorizations, and this slows down the information flow.

> *"So, there's certain people who are responsible for issuing information or authorize certain communications, and if they're not available then things can stop. Or they may have a high volume of these authorizations to do that it takes them a lot of time to get through" (Project planner, C3I2)*

This process of authorization means that multiple versions of designs can be circulating in different parts of the project. For example, whilst one design is in use by the construction team, in the meantime, the designers could have progressed the design, and the latest design is not uploaded as it is in the queue to get approved.

> *"Design teams and checkers and approvers could be progressing designs but then it'd be held up when someone say high up needed to actually approve the whole design" (Digital engineer, C3I3)*

> *"It's just making sure that I'm getting the latest information and no-one's updating it in the background and then the right versions are going onto [CDE1] […] it's no longer the most current version anymore by the time I'm reviewing it" (Project engineer, C3I1)*

> *"[…]when this revision has been updated from the designer to revision 10 and I sit here on my desk checking the revision 1 and the designer has the revision ten, then that revision 10 is internal […] because he keeps on updating but he hasn't he hasn't put it on the [CDE1]." (Technical manager, C3I5)*

This shows that the information available in the CDE is not the latest version, thereby reducing data quality associated with the timeliness dimension. The CDE has provisions for labelling status of a document as work-in-progress. However, even with that, it is hard to ascertain whether the information at hand is the latest as the work-in-progress documents can only be accessed by internal teams. This accessibility becomes a dimension of data quality which such approval processes make challenging. Additionally, even when the information is internally approved, it does not get stored on the CDE. Another approval is necessary to share the information with other stakeholders, thus limiting other stakeholders and algorithms from accessing this information for further decision making. In the metro project, an innovative approach to address this issue for drawings was implemented, placing a quick response (QR) code (a matrix barcode) in the document and a mobile app to scan the QR code and inform the user whether the drawing is the latest version or not. However, the document status must be continuously updated to make this useful and is limited to the issued drawings and not the work-in-progress drawings.

Although it seems to be straightforward from the outset, many users find document control frustrating because submission ends up being a long process even when all the attributes are correct. For example, engineers submit the packages to the document controller in their firm, who sends it to the document controller in the other firm (receiving end), which is then sent to the team lead and, finally, to the user who would get the useful information out of that package. This is a long process with checks and iterative cycles involved in each stage. The document controller makes sure that the files in the CDE have the relevant attributes before they are published in the CDE. If there are missing attributes, the submission is rejected. In the case of a Request For Information (RFI), if the document controller does not understand the request, it gets rejected. At times, it takes more than two to three weeks for the document to reach the recipient whilst following the document control workflow. This is essentially slowing down the whole information flow by implementing a system which was supposed to speed up the process.

> *"It took us three weeks to actually get the package with all the right documents and revisions in there, and that's a long time; again […] I gave it to my document control and my document control sent it to the other company's document control, the document control there sends it to whoever the lead is, and the lead then sends it on to whoever's doing the work – and that may take a week or*

two. And that's completely wasted time, and no one in the middle of that process has done any work[...] And actually, by the time it gets to the people who are reviewing the actual technical data, it may be three or four weeks later [...] In fact, I did it yesterday, I sent a load of RFIs through to [designer] informally, five minutes after I'd sent it through my document control." (Principal engineer, C3I8)

To bypass this obstacle, workers send the information through an unstructured channel in addition to the structured workflows. This is because of poor understanding of document control workflows amongst the project participants regarding the requirement of these structured workflows and the CDE. The slow processes and the need for completing the task before deadlines force employees not to follow document control workflows.

"There is generally a poor understanding of document control requirements, certification requirements [...] we're finding that general good practice that people should have brought with them from other projects is being conveniently put to one side for the purposes of expediting the work that people are being asked to do" (Information manager, C3I4)

In addition, the workflows in the CDE are complex and not intuitive, making it difficult for users to follow the protocols for document control.

"Just because of the way they need to store it in certain places and stuff like that, and it can't be done… The way that [CDE1] is set up here, I believe it is not easy to use" (Project engineer, C3I1)

"It seems very complex, you open them up, there's lots of things going on […] I just want to know where I can get my latest drawing" (Digital engineer, C3I3)

This has resulted in employees bypassing the workflow, which leads to system conflicts and further delays in the processes and, at times, in the information being stored on the CDE.

"I think someone within the doc management system had obviously circumnavigated it somehow, to get the drawings out. And then when we were trying to get the said revisions for our set out, the system wouldn't allow it because directory hadn't been properly created." (Technical manager, C3I5)

Not following the document control workflows leads to information loss in the CDE. This is a major setback to codification as the data is stored in an unstructured way which is difficult to access. This problem was found in the metro project as well. Conversations with the project manager (C2I1), chief site engineer (C2I2), and BIM consultants (C2I4, C254, C2I6) revealed that the project team weren't exposed to structured information flow used with digital technologies in the past. This made the implementation of CDE-based structured workflows difficult despite the training given to the participants, resulting in a combination of structured and unstructured workflows in the project.

The complexity of workflows and document control measures has implications for data quality. There are shared norms, values, and expectations for the users regarding the tools, such as speed, easy communication, transparency, and so on, which were developed based on their previous experiences of collaboration. When the new tool does not meet the expected qualities, it reduces their productivity, and users move back to the older ways of information sharing to expedite the task. When users find it difficult to utilize the CDE for structured information flow, they bypass the workflows to get the work expedited. This leads to the loss of metadata, document trails, and information dependencies as these unstructured communication channels offer limited or no codification. In addition, the document control workflow itself makes the process slow. Document control bottlenecks have multiple implications for data quality. Firstly, the value for the timeliness dimension is lowered as the data which is published might not be the latest. Hence, the inferences are based on old data, which leads to false interpretations. Secondly, this lowers the semantic accuracy of the data as its attributes might no longer be true. This also introduces the problem of consistency. Depending on which database employees look at, they see different values. For instance, one CDE to which the information was packaged would have the latest value, whilst the one which must go through another document controller would have a different value. This tampers with the idea of a 'single source of truth'. In addition, when users circumnavigate the workflow, there are more data quality issues related to unstructured information sharing such as accessibility and provenance.

4.2.4 Lack of process change

Even though structured information flow is digitized through the introduction of a CDE, the process enabling information flow remains unchanged. For example, for a piece of information to be approved, it must be printed, associated with a cover sheet, and signed.

> *"We're actually going to export that out of the CDE, we're going to print it out, we're going to staple it together, we're going to put our own cover sheet on the front of it, with the exact same details on the back and we're going to go off and go and get three signatures, scan it back in, put it back into [CDE1] and submit it."(Project engineer, C3I1)*

Printing and scanning the document results in loss of metadata. A scanned document in a PDF format has little machine-readable information in it. Inferring the contents from a scanned document is also resource intensive when compared to its original form. In the process of printing and scanning, the content becomes digitally accessible but not machine readable, thereby limiting the application of data science.

"What I'm finding now is it's not actually speeding everything up, it's sort of making everything a lot slower; which I find very frustrating" (Principal engineer, C3I8)

The lack of change in the processes reduces the value in the adoption of a CDE as it increases the time to do these tasks rather than a total reduction in time.

In the water project, three CDEs were used for the project, which created issues such as double handling, data inconsistencies, and so on. The presence of multiple CDEs in a project is another example of the lack of process change. Multiple CDEs resembles the paper-based workflows such as document flows between the designer and contractor, another set of document flows within the contractor's office, and another set of document flows to the clients for approvals.

'Our client prefers [CDE1], and we have the designer who stores things in [CDE2], so we have both of those tools, and we have to balance between those two. That can be very confusing when we have two platforms' (technical manager, C3I5).

For the information submissions to the clients, CDE1 was used; for the information from the design consultants to the contractor, CDE2 was used; and for internal file handling and sharing with the contractor, CDE3 was used. CDE1 and CDE3 came from the same vendor. However, CDE2 was from a different vendor.

"Everything had to be taken out of one data environment and pushed into another. One of the issues with that is the consistency or the compliance or knowing the latest versions of information" (Digital engineer, C3I6)

The existence of multiple CDEs within a project introduces the problem of data inconsistencies. Documents must be taken out of one CDE and placed in another. When the volume of information is huge, with each file having multiple versions, it is difficult to maintain consistency of documents across multiple CDEs. This means that information in a CDE might not be accurate and up to date, leading to incorrect interpretations when data analytics is performed on it.

"As the contractor, then we have to deliver it to a completely separate, disconnected CDE […] we're double-handling" (Digital engineer, C3I3)

When the CDEs are disconnected, the document trail is lost when a document is moved from one CDE to another, leading to the loss of traceability.

Lack of process change has multiple implications on data quality as well. Printing and scanning remove metadata and data relationships from the files. A scanned version of the file would also have very little machine-readable information embedded in it and would require resource-intensive methods to extract insights from it.

This reduces the data quality dimensions such as accessibility (as the metadata and data relationships are removed), completeness (information is not complete), and provenance (document trail is lost in the process). In addition, having multiple CDEs introduces the data quality issue associated with provenance as files residing in multiple CDEs are disconnected, and the relationships of that particular file with another file are lost in the transition process. There are further issues with synchronization of the information when it is distributed in multiple CDEs. This introduces the data quality issues associated with consistency, which limits the quality of findings made by inference algorithms.

4.3 Construction process information

This section presents the codification challenges related to construction process information. The data analysis shows that product information is relatively well structured as BIM models, analysis models, and CAD drawings. However, construction process information is relatively less structured and detailed when compared to product information. Process information is codified as Gantt chart models in scheduling software and then linked to the BIM model. The detailed process information is not structured into a model. Instead, it is shared as method statements in less structured PowerPoint presentations and PDF documents. Sharing information in these unstructured formats has implications for data quality, which are presented in this section.

4.3.1 Loss of constraint information

In construction, the constraints for any activity execution are discussed during team meetings as the constraints span between different teams. For example, logistics constraints span amongst prefabrication, logistics, and site teams. These discussions lead to the removal of constraints by rearranging the start and end times for the activities. These are then translated to Gantt charts as an output.

> *"So, from an engineering perspective, we have to interpret engineering information, whether it be drawing or written constraints, written narratives and interpret those into a Gantt chart. So, we physically need that information to know what we're building and what the constraints in building it are." (Project planner, C3I2).*

The above statement from the project planner provides evidence on the processes used to convert the constraints into a Gantt chart for communication. However, during this process, many of the constraints themselves, and thus context information for rearranging the activities, are lost. This is because Gantt charts can hold only precedence constraints. Other constraints, such as disjunctive (where activities cannot overlap) and logical constraints, are not embedded into the Gantt model. Instead, they are retained only as tacit information by project participants involved in team meetings. This is a case of incomplete information within the dataset as this information is only accessible to the meeting participants. For example, one of the meetings in the water project had an issue with piling, where the pile-driving equipment did not have access to the site for a specific date as there was another activity going on which limited the width of the site access road. At the meeting, this was raised:

> *"Access chamber works will conflict with access road for pile work, piling work package has to be moved back 2 weeks." (Progress review meeting, C3M3)*

Here, there is a dependency between access chamber works, and the piling work package as the access chamber work would reduce the road width. Therefore, the piling activity was delayed to a later date. The constraint was removed. However, the knowledge that there was a constraint is not recorded, and thus the presence of that constraint is not codified. This means such constraints are not machine readable as the access to this information was limited to the participants of a particular meeting. If an automatic scheduler is used to reschedule these activities, this rescheduling activity would not have access to this information, resulting in an unrealistic schedule.

Similar issues were observed in all cases. This issue introduces data quality problems associated with accessibility (information is limited to people who attended the meeting) and data completeness (the model does not include any constraint information; hence, it is incomplete).

4.3.2 Low level of detail

The precedence information codified into Gantt charts is linked with BIM to create 4D BIM simulations. However, the lack of detail in work packaging and associated information (such as constraints and resources) results in misinterpretation from the 4D models. The metro project follows a 5D BIM workflow (digital project management PPT—C2D2), where the schedule is linked to a BIM model, Bill of Quantities (BOQ), and an Enterprise Resource Planning (ERP) system to compare the cost based on the quantities versus the cost stated in the work orders from the subcontractors. The progress information is also linked to this model to ensure that the work is done before sanctioning the bills for the work orders.

> *"Work package for three spans were linked to a work order. Model showed the deck for a span was completed before the pier supporting it was completed because the work package for the first span was reported as completed." (Field notes- BIM Consultant 2, C2I5)*

<<Insert Figure 2>>

The deck of the metro can be completed only when the pier supporting it is completed, as shown in Figure 2. However, the system recorded the deck assembly to be completed when the pier was not completed. This was because work packages for the deck and pier were different as they were done by different subcontractors, and the level of detail of the work package is low. A subcontractor who dealt with deck assembly had a part of the work package completed, but the lack of detail in work packaging triggered the computer to record the whole work package as completed, resulting in the error. This is a clear case of lack of detail in the model leading to wrong inferencing.

Similar issues were observed in the student apartment (Case 1) by examining the schedule data in the construction programme (C1D1) and the water project (Case 3) by examining the construction programme update (C3D4) and observing the review meeting (C3M3). These issues are caused by low data quality due to incomplete information related to the completeness dimension.

5. Discussion

This section discusses the software usage, information sharing, and construction process information codification challenges which limit the uptake of data science in construction, drawing on the evidence from the empirical study. The discussion relates the findings to the literature on BIM use in practice (e.g. Dossick and Neff, 2010; Harty and Whyte, 2010) and other strands of research on data quality, machine readability, and BIM adoption and implementation to articulate how these new analyses contribute by extending understanding of codification challenges. Furthermore, building on and extending Batini and Scannapieco (2016), it shows how these codification challenges are then mapped to their data quality dimensions, such as accuracy, completeness, timeliness, consistency, accessibility, and data provenance.

5.1 Software usage

The findings on software usage show that, despite significant digitization of work processes, data remains fragmented into different domains and formats because of the multiple software tools in use across the organizations involved in construction. Work in the construction information technology community is pioneering new data management solutions to improve interoperability (Hu et al., 2016; Pauwels et al., 2010; Pazlar and Turk, 2008; Redmond et al., 2012), and it is disappointing to find that construction projects still suffer from poor quality data as a result of problems of interoperability caused by the existence of multiple domain-specific tools and modelling practices. In their work, Dossick and Neff (2010) have previously shown how the organizational and cultural divisions between the designers and builders, contractors, and subcontractors stifle collaborative work. This paper shows these issues are not resolved. In the projects studied, organizational and cultural divisions between the firms involved in the late design and construction stages of projects cause software usage problems (interoperability, information loss during format conversion, multiple modelling techniques). Whilst there may be shared norms and tools within a firm for modelling information, these norms differ across the firms which modelled project information. Multiple modelling techniques (as described in 4.1.3) and data created using different software (as described in 4.1.1) result in datasets which are not interoperable and require format conversions, resulting in loss of information and low machine readability. The water project (Case 3) had multiple firms working on the data over different phases of the project, with interoperability problems more prevalent in this case in comparison with the student apartment (Case 1), in which a single leading firm was involved with the creation and use of the model. Although the metro project (Case 2) had different firms over the different phases of the project, digital data creation was handled through a single owner support organization, limiting the impact of this problem. Software usage problems are found to lead to challenges of codification for data science; hence, this work extends prior insights by Dossick and Neff (2010) to show how organizational and cultural divisions between designers and builders not only stifle collaborative work and joint problem-solving but also result in fragmented datasets in construction, leading to data silos and data loss and, thus, resulting in poor data quality which is more difficult to use in data science.

As it is relatively unusual and potentially undesirable to have one firm or owner with overall control of the model, to enable more distributed working, developers of new tools or digitally enabled processes should consider the implication of organizational separation in the sector in addition to the technical requirements. In their work, Dossick and Neff (2010) have described the influence of strong leadership to hold people together on a project to improve collaboration despite professional segregation. Similarly, a set of common practices and a larger vision for the data creation and management should be laid out in the project to ensure the data meets the necessary quality to enable its use without loss of information in between. To achieve better-quality data in projects, practitioners must focus beyond the individual scope of their multiple firms towards the common goals of the project.

5.2 Information sharing

The analyses suggest that the construction sector has not yet made the transition from document-based to model-based ways of organizing digital data. The use of drawings and file-based sharing, unstructured information sharing, printing and scanning of documents, multiple CDEs, and so forth in information sharing has a significant impact on the machine readability of data. Paper-based practices are institutionalized in the sector, and while they are being replaced by digital ways of working, this change is slow, with users of construction information still conditioned to work with drawings and PDFs and unstructured information sharing. Even in projects which are championing newer BIM-based workflows using CDEs, this work finds it is difficult to replace these practices, as evidenced by the problems associated with information sharing (section 4.2). The complexity and long processing times involved in these workflows force users to shift back to existing practices and workarounds to expedite their work. The findings from this paper also support the previous characterization of users in construction combining new structured methods of information sharing along with the prior practice of unstructured information sharing when they were hindered by bottlenecks in processes, such as document control. Thus, we can characterize the project participants use a range of new and existing practices together as 'hybrid practices' (Harty and Whyte, 2010), and their circumnavigation of workflows results in unstructured information sharing (as shown in a previous discussion in Whyte et al. (2016)). However, this paper goes beyond such studies to characterize the implications for data quality and to highlight, building on Hartmann (2008), the potential to develop newer workflows and digitally enabled processes which address the challenges faced by practitioners.

5.3 Construction process information

Regarding construction process information, this paper shows that the process codification is limited to the master planning or phase planning level and lacks the level of detail and linkage required for the application of data science tools. Whilst product information is relatively well codified in BIM, process information is less well detailed, and there is a lack of constraint information. These challenges are identified by researchers implementing 4D BIM. For example, Han and Golparvar-Fard (2017) have stated that the process modelling methods fail to document field issues to be made available for further analysis; for instance, the 4D BIM's "Model Breakdown Structure typically does not match operational details or require creating complicated namespaces which, without visual representations, are difficult to communicate" (p. 1733). Giretti et al. (2012) have further reported the lack of correlation between the resources employed hourly and work progress. This led to the decomposition of tasks into sub-tasks to determine causal relationships between the involved variables so the whole progress could be determined. This study suggests that to overcome such reported issues, the methods for codifying construction process information must be more detailed. The institutionalized practice of planning being limited to master planning and phase planning, without the focus on granular planning such as look-ahead planning and weekly planning, is causing this codification challenge, with the lack of semantic relationships embedded in the model limiting the application of automatic schedulers. These issues suggest both a change in the modelling of process information in construction, with the need to develop tools which support modelling of complex constraint information, and also a change in the practice to codify the process information in greater detail so that data science could be employed to augment decision making in construction.

5.4 Machine readability of construction datasets

This section discusses the machine readability of the construction datasets. Common construction datasets are classified based on the set of 'rules' for creating structured data described by Berners-Lee (2006) in Table 4.

<<Insert Table 4>>

Most of the construction information observed from the cases satisfies the requirement for a one-star category. The observed projects use a CDE for storing and managing project data, resulting in indexing the data and storing it on online servers, resulting in one-star data. CDE makes the data easier to retrieve for the computers to make inferences on them. However, the complexity of the new structured methods for information sharing using CDE, and the poor understanding of workflows across the teams, leads to the use of a combination of structured and unstructured channels for information sharing, as discussed in section 4.2. This aspect reduces the machine readability of the information distributed over unstructured channels as the information is not indexed nor available on a common server. The same issue occurs when the users circumnavigate the workflows to get the work expedited. Similarly, codification challenges associated with construction process information also lower the machine readability as the information is not recorded (lack of detail and loss of construction information) and, hence, not indexed or stored in online servers. These issues make the information inaccessible for inferencing.

With regard to the structure of the construction information, construction datasets in the form of BIM, project management information in project management software (Primavera P6, Asta powerproject, etc.), outputs from Microsoft tools such as Excel, and so forth are structured, satisfying the requirements for two-star data. However, construction data are also unstructured in the forms of PDF documents, drawings, and other file-based formats, as described in sections 4.2.2 and 4.2.4. The lack of structure in the datasets makes inferencing from them difficult, leading to the need for complex algorithms. Where the construction data is structured, the data structure is in proprietary formats which require the APIs to access the semantic relationships within the data. The observed projects do not use open formats or open standards for publishing the data. Proprietary tools for the authorship of construction data are far more advanced and easier to use than the open-source tools. Hence, construction projects resort to using the tested and robust proprietary tools, resulting in issues associated with interoperability and loss of information presented in section 4.1. Thus, the construction information rarely achieves three-star classification, as mentioned by Berners-Lee (2006). In conclusion, the maximum level of machine readability of the construction datasets in the observed projects is two-star, with most of the information with a one-star rating.

Low machine readability of the data has implications on data quality. When the construction information is not stored on servers due to information sharing issues or lack of detail in the models, the accessibility of that data is affected, thus reducing the accessibility dimension of the data quality. When information is stored in CDE (satisfying conditions for one-star data) as PDF documents, the structure of the data is not maintained, resulting in data quality issues associated with syntactic accuracy and consistency. Lack of a data structure removes the context from the information, thus resulting in the need for complex algorithms to infer the contexts and infer from data. This issue also introduces problems associated with consistency as the data value for a field might be different in different files, and the lack of context limits the computer programs to detect it. This problem is further worsened as the datasets are not linked since the links are lost when a file is moved from one CDE to another. Storing the information in proprietary formats also reduces the accessibility dimension for data quality.

5.5 Implications for data quality

The codification challenges discussed earlier have many significant implications for data quality. To unpack these in this section, they are mapped onto the different quality dimensions.

<<Insert Table 5>>

Accuracy: The organizational and cultural divisions between different teams results in problems associated with multiple modelling techniques, leading to data quality issues concerning the syntactic accuracy of the data. This was evident from the dataset when different people had different perceptions of the model, as in the case of the slab example in 4.1.3. Similarly, the hybrid practices associated with information sharing lead to lowering the semantic accuracy of the data as the data with which inferences are made are not accurate due to inefficiencies in information sharing. When there are syntactic errors in the data, this leads to incorrect insights (for example, if a slab is modelled as a geometric object with attributes attached to it and when a software tool is used to compute quantity take-off from the model for all the slabs). The output would be zero as quantity required as the program fails to identify the geometric object as a slab. If this software is integrated with a costing tool used for cash flow analytics, this error gets propagated into that tool. These errors can be removed to an extent using semantic enrichment programs. However, even the accuracy of inferences of semantic enrichment programs is dependent on the quality of input datasets (Sacks et al., 2017).

Completeness: Concerning the completeness of the data, organizational and cultural divisions between the teams resulting in problems such as interoperability, format conversions, multiple modelling techniques and the implication of hybrid practices such as printing and scanning the documents play a role in reducing the data quality. For example, the software usage issues caused by the organizational divisions leads to format conversion resulting data loss leading incomplete dataset. Lack of process change also leads to similar problems such as loss of metadata when documents are printed and scanned. The institutionalized practices of process modelling with low levels of detail and the practice of not codifying constraints in the model result in incomplete data. Inferring insights from incomplete datasets reduces the quality of the output. For example, if the constraints are not codified in a schedule, an automatic scheduler would create an unrealistic schedule. This leads to further problems down the line. In the case of a product model, an incomplete dataset used for a structural capacity prediction would give incorrect results.

Timeliness: This dimension of data quality is mostly affected by the information-sharing practices in construction using hybrid practices by using old and new practices simultaneously. File-based sharing (resulting in slowing down the data updates compared to model-based sharing), document control bottlenecks delaying the submission of data into the system, and the use of multiple CDEs requiring data transfer from one to another (resulting in a delay in fetching the data) all constitute outdated data-skewing analytics. When outdated data is used for analytics, the resulting inference is not time appropriate, and decisions taken with those inferences lead to problems. For example, if the construction plan is made using resource availability data, during the actual construction date, the assigned resource might not be available. When this is managed manually, the planner makes sure this does not happen. However, when performed automatically, it is necessary that the datasets are updated close to real time.

Consistency: Hybrid practices in information sharing, resulting in document control bottlenecks, multiple CDEs, and circumnavigating workflows, induce inconsistency in the data. The data in one CDE might be different from that in another CDE. Similarly, the document control bottlenecks in publishing the data result in different teams and their databases having different versions of the data. This creates a problem during the data analysis phase. For example, for the same content, different values might exist, of which only one is true. This induces problems of semantic accuracy if the computer takes the incorrect value for analysis.

Accessibility: This dimension of data quality is affected by organizational and cultural divisions between teams, hybrid practices, and institutionalized practices in process modelling. Organizational and cultural divisions between teams, resulting in the use of multiple software, multiple modelling techniques, and so forth, affect the accessibility of data as employees must convert the data to access it, and there are instances which show conversions result in loss of data. Hence, the data is only accessible in full content to the team, which created it. Similarly, the hybrid practices, particularly printing and subsequent scanning of the documents, remove the metadata as well as semantic information within those data, resulting in a less-efficient analysis. In addition, the lack of constraint codification restricts the information regarding constraints, making it accessible to the few people who attended the meeting. It is not embedded in the model and, hence, not accessible to any analytic algorithms. This reduces the capability of such algorithms to infer accurate information and determine causalities.

Provenance: This dimension of the data quality is affected by both organizational and cultural divisions in the industry as well as hybrid practices. Fragmentation in the industry leads to many conversions and much manipulation of data to suit purposes, in the process removing the traceability and origins of the data. Similarly, multiple CDEs, printing and scanning of documents and unstructured information sharing also lead to loss of metadata and ends up removing information regarding origins of the data. This limits algorithms from making inferences based on data origins and their incremental manipulation.

This section described the challenges of codification as causes of data quality issues, which have implications for different data quality dimensions and the output of data analytics. While big data techniques suggest future opportunities to use data science with an increasing variety of data of different levels of veracity, data cleaning is a resource-intensive process, and significant training models are required. To improve the uptake of data science in construction, high-quality data is necessary, and achieving this requires overcoming the codification challenges identified here.

6. Conclusions and future work

Codification challenges for data science in construction are found to be related to: 1) software usage—interoperability, information loss during conversion, and multiple modelling techniques, 2) information sharing—unstructured information sharing, drawing and file-based sharing document control, and lack of process change, and 3) construction process information—loss of constraints and low levels of detail. The implication of these challenges was discussed by mapping them to data quality dimensions such as accuracy, completeness, timeliness, consistency, and provenance. Through the identification of the codification challenges in the late design and construction phase of the projects and their mapping to the data quality dimensions, this paper extends the knowledge on data quality issues in construction. It shows how data quality arises from organizational as well as technical practices. The persistence of organizational and cultural divisions, paper-and document-based (as well as digital- and model-based) ways of working, and institutionalized practices of construction process modelling are challenges for the uptake of data science as they lead to the partial codification of information in machine-readable formats. Whilst the fragmented nature of the sector is well understood, this work shows how codification challenges arise because of the different digital workflows and working practices across projects, and how these lead to fragmented data as a result of the use of multiple software packages, poor information sharing and only partially captured construction process information.

While all of the projects studied were BIM enabled, within these projects, information sharing issues and software usage issues emerge as a result of their document-based, and sometimes paper-based, practices. Project-wide standards and policies are created to streamline information sharing through structured workflows, and these workflows are aimed at improving collaboration, but the process of document control raises issues. In addition, this study identifies inefficiencies (such as long processing times, complexity, etc.) in these workflows, which pushes the users to bypass the structured methods in the current workflows. This forces the users to revert to older methods or use a combination of old and new methods, resulting in the generation of unstructured information. These cause data quality issues relating to inaccessibility of data, timeliness of data, and data consistency.

Limited codification in construction process information is a result of the institutionalized practice of scheduling focusing on codifying just precedence relationships into a Gantt chart. The advancements in 4D and 5D BIM have attempted to address this problem to an extent by integrating the resource and cost information to a single platform. However, the data herein suggest that the level of detail of the scheduling is still at a macro-planning level. In addition, the constraint relationships between resources and processes, resources and the site conditions, and site conditions and the processes discussed in weekly meetings are not codified into a model. This limits the data science due to data quality issues associated with completeness and accessibility.

How might codification challenges be overcome to enable greater uptake of data science in construction? The evidence from this paper points towards a need for change in the policy and practice to ensure machine readability and better-quality construction information. The current policies on information modelling are aimed at improving collaboration amongst the project participants. Building on the existing policies, newer policies should ensure that the machine readability and quality of the data are maintained during information sharing. Care should be taken to ensure that newer policies suit the existing work practices in construction so that the workflows recommended by such policies are not bypassed by the users. Newer policies also ensure the transition from file-based information sharing to model-based information sharing, where the model data is shared based on the principles stated by Berners-Lee (2006) to ensure machine readability. Standards under development, such as ISO/DIS 21597—information container for data drop, exchange specification parts 1 and 2—are moving in this direction, where linked data technologies are used to define the relationships between documents and datasets, thereby automating the document control processes (International Organization for Standardization, 2019). These standards would have a positive impact on data quality pertaining to dimensions, timeliness, accessibility, and consistency. Even though policy interventions would have a positive effect on data quality, the construction sector is known for institutionalized practices and resistance to change. Thus, policies must influence practices to enable greater uptake of data science.

Thus, in addition to the policies, it is, therefore, necessary to adopt newer methods in practice to make construction information machine readable. This study examined a student apartment building project constructed by a leading contractor in the United Kingdom using state-of-the-art offsite manufacturing techniques (Case 1), a metro project pioneering digital transformation in India (Case 2), and an innovative water megaproject in the United Kingdom (Case 3). Despite these projects having advanced workflows and innovative approaches to ensure structured information, lack of process change after digitization of workflows was evident in the datasets, which included multiple CDEs, printing and scanning of documents, need for wet signatures, and such. These practices continue to occur due to the lack of trust in digital workflows; thus, there must be a change in this mindset, and trust in digital data must be developed in the construction practice. There is also a need for a shift to model-based information sharing from the file-based information sharing between teams. To address this issue, related to a lack of detail in process information, there is a need for change in the look-ahead scheduling practice to incorporate the codification of different constraint relationships into information. This might involve a significant change in the planning of a progress review process by integrating BIM-based scheduling tools within the planning and progress meetings and modelling different constraints within the software environment to capture it from the discussions.

The current study has identified the codification challenges by examining information sharing in the late design and construction phase and mapped it to the data quality dimensions. The findings from this study inform the researchers, who are developing frameworks and methods to codify construction information, of the organizational issues to be considered in their work. This paper also provides the data quality implications of issues associated with BIM implementation, motivating the researchers focusing on implementation studies to widen their scope from collaboration to include data quality and machine readability. Future research can build on this study to develop recommendations for ensuring the machine readability of construction information generated throughout different stages of the lifecycle of the project. To achieve this, future research should focus on different aspects. Firstly, researchers should elaborate on the organizational issues identified in the current study across different phases of the project and amongst different stakeholders involved in the project. Secondly, accounting for the complexity of information-use trends in construction, there is a need for research in fundamental data science to pave the way for the integration of the disconnected information in the construction sector. This includes developing newer information modelling approaches which adapt to the current work processes as well as support codification of information, such as algorithms for crawling through the disconnected information to draw insights and learning algorithms to detect discrepancies in data (such as with accuracy, completeness, timeliness, consistency, and provenance) and predict their consequences. Finally, researchers and managers of construction projects should work together towards developing workflows and information sharing practices which ensure the machine readability of construction information whilst considering the issues of fragmentation, hybrid practices and institutionalized practices. This paper has provided a foundation for such future research by extending the knowledge on data quality issues in construction through the identification of codification challenges, considering the wider practice of model and document-based information sharing.

7. Data availability statement

To comply with the research ethics process set out by Imperial College London (https://www.imperial.ac.uk/research-ethics-committee/application-process/), an agreement was made with the participants of this study. Sharing of the data, that support the findings of this study, is restricted by the agreement made with the research participants. As a result, the data cannot be shared without the permission of the participants.

8. Acknowledgements

The authors are grateful to the research participants in the three case studies. The PhD research of the first author is co-funded by Bentley Systems UK and through a Skempton Scholarship from the Department of Civil and Environmental Engineering, Imperial College, London. During the development of this paper, this author was supported by the PhD enrichment scholarship from the Alan Turing Institute, the United Kingdom's National Institute for Data Science and AI. The second author acknowledges the support of Laing O'Rourke and the Royal Academy of Engineering for cosponsoring her Professorship.

References

Akintola, A., Venkatachalam, S., and Root, D. (2017). "New BIM Roles' Legitimacy and Changing Power Dynamics on BIM-Enabled Projects." *Journal of Construction Engineering and Management*, 143(9), 04017066.

Batini, C., and Scannapieco, M. 2016. *Data and information quality*. Cham, Switzerland: Springer International Publishing.

Beetz, J., Berlo, v. L. L., Laat, d. R., and Helm, v. d. P. (2010). "BIMSERVER.ORG – An open source IFC Model Server." In *Proc.,CIB W78 2010: 27th International Conference*, Cairo, Egypt.

Berners-Lee, T. 2006. "Linked Data - Design Issues."Accessed:19/03/2019. https://www.w3.org/DesignIssues/LinkedData.html.

Bilal, M., Oyedele, L. O., Qadir, J., Munir, K., Ajayi, S. O., Akinade, O. O., Owolabi, H. A., Alaka, H. A., and Pasha, M. (2016). "Big Data in the construction industry: A review of present status, opportunities, and future trends." *Advanced Engineering Informatics*, 30(3), 500-521.

Bolton, A., Butler, L., Dabson, I., Enzer, M., Evans, M., Fenemore, T., Harradence, F., Keaney, E., Kemp, A., Luck, A., Pawsey, N., Saville, S., Schooling, J., Sharp, M., Smith, T., Tennison, J., Whyte, J., Wilson, A., and Makri, C. (2018). "The Gemini Principles: Guiding values for the national digital twin and information management framework." Centre for Digital Built Britain and Digital Framework Task Group, Cambridge, UK.

British Standards Institution (2018). "ISO19650 - 1 & 2 : Organization and digitization of information about buildings and civil engineering works, including building information modelling (BIM). Information management using building information modelling." Retrieved from: https://bsol.bsigroup.com/Bibliographic/BibliographicInfoData/000000000030333754.

Cai, L., and Zhu, Y. (2015). "The Challenges of Data Quality and Data Quality Assessment in the Big Data Era." *Data Science Journal*, 14(2), 1-10.

Cao, D., Li, H., and Wang, G. (2014). "Impacts of Isomorphic Pressures on BIM Adoption in Construction Projects." *Journal of Construction Engineering and Management*, 140(12), 04014056

Carrillo, P., Harding, J., and Choudhary, A. (2011). "Knowledge discovery from post-project reviews." *Construction Management and Economics*, 29(7), 713-723.

Chang, C.-Y., Pan, W., and Howard, R. (2017). "Impact of Building Information Modeling Implementation on the Acceptance of Integrated Delivery Systems: Structural Equation Modeling Analysis." *Journal of Construction Engineering and Management*, 143(8), 04017044.

Chegu Badrinath, A., and Hsieh, S.-h. (2019). "Empirical Approach to Identify Operational Critical Success Factors for BIM Projects." *Journal of Construction Engineering and Management*, 145(3), 04018140.

Choi, B., Lee, H.-S., Park, M., Cho, Y. K., and Kim, H. (2014). "Framework for Work-Space Planning Using Four-Dimensional BIM in Construction Projects." *Journal of Construction Engineering and Management*, 140(9), 04014041.

Delone, W., and McLean, E. (2014). "The DeLone and McLean Model of Information Systems Success: A Ten-Year Update." *Journal of Management Information Systems*, 19(4), 9-30.

Donato, V., Lo Turco, M., and Bocconcino, M. M. (2017). "BIM-QA/QC in the architectural design process." *Architectural Engineering and Design Management*, 14(3), 239-254.

Dossick, C. S., and Neff, G. (2010). "Organizational Divisions in BIM-Enabled Commercial Construction." *Journal of Construction Engineering and Management*, 136(4), 459-467.

Eastman, C., Teicholz, P., Sacks, R., and Liston, K. 2008. *BIM Handbook: A Guide to Building Information Modeling for Owners, Managers, Designers, Engineers, and Contractors*. Hoboken: John Wiley & Sons, Inc.

Fan, W. (2015). "Data Quality: From theory to practice." *ACM SIGMOD Record*, 44(3), 7-18.

Farias, T. M. d., Roxin, A., and Nicolle, C. (2018). "A rule-based methodology to extract building model views." *Automation in Construction*, 92, 214-229.

Giretti, A., Carbonari, A., Novembri, G., and Robuffo, F. (2012). "Estimation of job-site work progress through on-site monitoring." In *Proc.,9th International Symposium of Automation and Robotics in Construction, ISARC* 2012, Eindhoven, The Netherlands.

Goedert, J. D., and Meadati, P. (2008). "Integrating Construction Process Documentation into Building Information Modeling." *Journal of Construction Engineering and Management*, 134(7), 509-516.

Gu, N., and London, K. (2010). "Understanding and facilitating BIM adoption in the AEC industry." *Automation in Construction*, 19(8), 988-999.

Han, K. K., and Golparvar-Fard, M. (2017). "Potential of big visual data and building information modeling for construction performance analytics: An exploratory study." *Automation in Construction*, 73, 184-198.

Hartmann, T. (2008). "A grassroots model of decision support system implications by construction project teams." Ph.D., Stanford University, California, USA.

Harty, C., and Whyte, J. (2010). "Emerging Hybrid Practices in Construction Design Work: Role of Mixed Media." *Journal of Construction Engineering and Management*, 136(4), 468-476.

Hendler, J., and Pardo, T. A. 2012. "A Primer on Machine Readability for Online Documents and Data."Accessed:10th October 2019. https://www.data.gov/developers/blog/primer-machine-readability-online-documents-and-data.

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*, 66, 29-42.

Hwang, B.-g., Zhao, X., and Yang, K. W. (2019). "Effect of BIM on Rework in Construction Projects in Singapore: Status Quo, Magnitude, Impact, and Strategies." *Journal of Construction Engineering and Management*, 145(2), 04018125.

International Organization for Standardization (2019). "ISO/DIS 21597 Information container for data drop -- Exchange specificaton ", Retrieved from https://www.iso.org/standard/74389.html

Janssen, M., van der Voort, H., and Wahyudi, A. (2017). "Factors influencing big data decision-making quality." *Journal of Business Research*, 70(1), 338-345.

Jaradat, S., Whyte, J., and Luck, R. (2013). "Professionalism in digitally mediated project work." *Building Research & Information*, 41(1), 51-59.

Jayawardene, V., Sadiq, S., and Indulska, M. (2015). "An Analysis of Data Quality Dimensions " *ITEE Technical Report*, The University of Queensland.

Jordani, D. A. (2010). "BIM and FM: The Portal to Lifecycle Facility Management." *Journal of Building Information Modeling*, 13-16.

Kassem, M., and Succar, B. (2017). "Macro BIM adoption: Comparative market analysis." *Automation in Construction*, 81, 286-299.

Kim, K., Kim, H., Kim, W., Kim, C., Kim, J., and Yu, J. (2018). "Integration of ifc objects and facility management work information using Semantic Web." *Automation in Construction*, 87, 173-187.

Krijnen, T., and Beetz, J. (2017). "An IFC schema extension and binary serialization format to efficiently integrate point cloud data into building models." *Advanced Engineering Informatics*, 33, 473-490.

Kwon, O. S., Park, C. S., and Lim, C. R. (2014). "A defect management system for reinforced concrete work utilizing BIM, image-matching and augmented reality." *Automation in Construction*, 46, 74-81.

Lee, Y.-C., Eastman, C. M., and Solihin, W. (2018). "Logic for ensuring the data exchange integrity of building information models." *Automation in Construction*, 85, 249-262.

Loshin, D. 2010. *The Practitioner's Guide to Data Quality Improvement*. Burlington, MA: Morgan Kaufmann.

Mirarchi, C., and Pavan, A. (2019). "Building information models are dirty." In *Proc., 2019 European Conference of Computing in Construction (2019 EC^3),*Chania, Greece.

Naumann, F., and Rolker, C. (2000). "Assessment Methods for Information Quality Criteria." In *Proc., Fifth Conference on Information Quality (IQ 2000),*Cambridge, US.

Oti, A. H., Tah, J. H. M., and Abanda, F. H. (2018). "Integration of Lessons Learned Knowledge in Building Information Modeling." *Journal of Construction Engineering and Management*, 144(9), 04018081.

Pauwels, P., De Meyer, R., Van Campenhout, J., Meyer, R. D., and Campenhout, J. V. (2010). "Interoperability for the Design and Construction Industry through Semantic Web Technology." *Lecture Notes in Computer Science*, 6725(1), 143-158.

Pauwels, P., Törmä, S., Beetz, J., Weise, M., and Liebich, T. (2015). "Linked Data in Architecture and Construction." *Automation in Construction*, 57, 175-177.

Pazlar, T., and Turk, Ž. (2008). "Interoperability in practice: Geometric data exchange using the IFC standard." *Journal of Information Technology in Construction*, 13(1), 362-380.

Pedro, A., Lee, D. Y., Hussain, R., and Park, C. S. "Linked Data System for Sharing Construction Safety Information." In *Proc., 34th International Symposium on Automation and Robotics in Construction (ISARC 2017)*, Taipei, Taiwan.

Preidel, C., Borrmann, A., Oberender, C., and Tretheway, M. (2016). "Seamless Integration of Common Data Environment Access into BIM Authoring Applications: the BIM Integration Framework."In *Proc., 11th European Conference on Product and Process Modelling (ECPPM 2016)*, Limassol, Cyprus, 119.

Quintero, D., Genovese, W., Kim, K., Li, M., Martins, F., Nainwal, A., Smolej, D., Tabinowski, M., and Tiwary, A. 2015. *IBM software defined environment*. IBM Redbooks.

Redmond, A., Hore, A., Alshawi, M., and West, R. (2012). "Exploring how information exchanges can be enhanced through Cloud BIM." *Automation in Construction*, 24, 175-183.

Sacks, R., Eastman, C., Lee, G., and Teicholz, P. 2018. *BIM handbook: a guide to building information modeling for owners, designers, engineers, contractors, and facility managers.* New Jersey, USA: John Wiley & Sons.

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), 04017062.

Sebastian, R. (2011). "Changing roles of the clients, architects and contractors through BIM." *Engineering, Construction and Architectural Management*, 18(2), 176-187.

Singh, V., Gu, N., and Wang, X. (2011). "A theoretical framework of a BIM-based multi-disciplinary collaboration platform." *Automation in Construction*, 20(2), 134-144.

Solihin, W., Eastman, C., and Lee, Y.-C. (2015). "Toward robust and quantifiable automated IFC quality validation." *Advanced Engineering Informatics*, 29(3), 739-756.

Taylor, J. E. (2007). "Antecedents of Successful Three-Dimensional Computer-Aided Design Implementation in Design and Construction Networks." *Journal of Construction Engineering and Management*, 133(12), 993-1002.

Wang, L. (2017). "Heterogeneous Data and Big Data Analytics." *Automatic Control and Information Sciences*, 3, 8-15.

Wang, R. Y., and Strong, D. M. (1996). "Beyond Accuracy: What Data Quality Means to Data Consumers." *Journal of Management Information Systems,*, 12(4), 5-33.

Whyte, J. (2011). "Managing digital coordination of design: emerging hybrid practices in an institutionalized project setting." *Engineering Project Organization Journal*, 1(3), 159-168.

Whyte, J., Stasis, A., and Lindkvist, C. (2016). "Managing change in the delivery of complex projects: Configuration management, asset information and 'big data'." *International Journal of Project Management*, 34(2), 339-351.

Won, J., Lee, G., Dossick, C., and Messner, J. (2013). "Where to Focus for Successful Adoption of Building Information Modeling within Organization." *Journal of Construction Engineering and Management*, 139(11), 04013014.

Younas, M. (2019). "Research challenges of big data." *Service Oriented Computing and Applications*, 13(2), 105-107.

Zadeh, P. A., Wang, G., Cavka, H. B., Staub-French, S., and Pottinger, R. (2017). "Information Quality Assessment for Facility Management." *Advanced Engineering Informatics*, 33, 181-205.

Zhang, C., and Beetz, J. (2015). "Model Checking on the Semantic Web: IFC Validation Using Modularized and Distributed Constraints." In *Proc., 32nd CIB W78 Conference 2015, 27th-29th October 2015*, Eindhoven, The Netherlands, 819-827.

Zhang, J., Liu, Q., Hu, Z., Lin, J., and Yu, F. (2017). "A multi-server information-sharing environment for cross-party collaboration on a private cloud." *Automation in Construction*, 81, 180-195.

Zhang, S., Pan, F., Wang, C., Sun, Y., and Wang, H. (2017). "BIM-Based Collaboration Platform for the Management of EPC Projects in Hydropower Engineering." *Journal of Construction Engineering and Management*, 143(12), 04017087.

Fig. 1. Visual representation showing the effect of the low level of detail in process information

Table 1. Level of machine readability the data based on linked data principles set out by Berners-Lee (2006)

Table 2. Codification challenges across cases

Table 3. Information sharing between different teams

Table 4: Construction data sets categorized based on levels of machine readability

Table 5. Mapping between codification challenges and data quality dimensions

Table 1: Level of machine readability the data based on linked data principles set out by Berners-Lee (2006)

Quality of data	Principles for publishing a machine-readable data set
1-Star	Data is available on the web
2-Star	1-star data structured in a proprietary format
3-Star	1-star data structured in a non-proprietary format
4-Star	3-star data that is published using open standards
5-Star	4-star data with links to other 4 star datasets

Table 2: Codification challenges across the cases

Codification challenges	Observations	The student apartment	The metro project	The water project
Software usage	Interoperability	X	X	X
	Information loss during conversions	X		X
	Modelling technique			X
Information sharing	Unstructured information sharing	X	X	X
	Drawing and file-based sharing	X	X	X
	Document control bottleneck		X	X
	Lack of process change	X	X	X
Process information	Loss of constraints	X	X	X
	Low level of detail	X	X	X

Table 3: Information sharing between different teams

Media	Example evidence from the dataset
Common Data Environments (CDE1, CDE2 from different vendors)	*Formal submissions, drawing receipts, design and temporary works*: "if it's formal document submission, we do it through [CDE1] [...] When I receive drawings from [design consultant], I get them through [CDE2]. And quite a lot of the designers use [CDE2] and[...] Well, we try get all the design functions, including temporary works to use [CDE2]." (Project engineer, C3I1)
Reports	*Spreadsheets and documents:* "There's lots of reporting on the project[...] And then that gets out into various outputs, so that could be just a schedule in Excel. Lots of Excel outputs as well, huge amount of Excel outputs. And, if it's a commercial discussion there may need to be some narrative around it, so using Microsoft Word to develop a narrative." (Project Planner lead, C3I2)
Meetings	*Design review* meetings: "You could do a design meeting, review something and then say, write comment on that[...]to understand what information, they're going to require at a particular stage. So that may consist of meetings; that might consist of face-to-face conversations, emails, etc." (Digital engineering lead, C3I3) *Buildability meetings* "attendance to buildability meetings and trying to get out of them what sort of temporary works may be needed to build something" (Project engineer, C3I1) *Client meetings*: "I will be going off-site to attend meetings with the client" (Information manager, C3I4) *Design for Manufacture and Assembly (DfMA) input*: "I'm trying to attend meetings and troubleshoot and try to help and provide technical input into the design and assisting the designer and support team" (Technical manager/DfMA coordinator, C3I5)
Email	*Highly used*: "So, obviously we do use emails a lot." (Project Planner lead, C3I2) *A normal type of communication*: "Then emails, meetings, usually types of communication." (Project engineer, C3I1) *Provides remote precision*: "If I'm communicating over longer distances or if I think to myself, I'd better make a precise request, then it'll be emails. We don't use a communicator-type facility in [CDE1]". (Information manager, C3I4) "Stakeholders, yes, it's certainly meetings and emails. Most of our stakeholders don't want to use [CDE1], because of the admin that comes with that" (Principal engineer, C3I8)
Remote conversations	*Online meetings*: "Generally, like a Skype, conference calls, linked meetings" (Senior digital engineer, C3I6) *Online communications and records of design logs, discussion points, online forums*: "Yes, so all the design data is held within a common data environment, which was [CDE2]. And so, I managed that area and access to that area. Then all communications were stored on SharePoint on Microsoft online, so everyone had access to registers or design logs or discussion points, almost used as an online forum where anyone could ask questions" (Senior digital engineer, C3I6) *Telephone calls with design consultants*: "So, I'll start with between us and design consultants: there's meetings, emails, and phone calls. I prefer meetings and phone calls" (Principal engineer, C3I8)

Table 4: Construction data sets categorized based on levels of machine readability

Quality of data	Principles for publishing a machine-readable data set	Construction data sets
1-Star	Data is available on the web	Files and Models uploaded in the common data environment
2-Star	1-star data structured in a proprietary format	BIM files in proprietary formats (Revit files, Microstation files, etc.), project management information (Asta power project, Primavera P6, Microsoft project), design rationale and associated information (in Microsoft Excel), etc.
3-Star	1-star data structured in a non-proprietary format	BIM files in IFC format, CSV data etc.
4-Star	3-star data that is published using open standards	BIM files published using open standards such as ifcOWL, BOT ontology etc.
5-Star	4-star data with links to other 4 star datasets	BIM files published using open standards linked to other such files (BIM files, GIS data etc).

Table 5. The mapping between findings and data quality dimensions

Data quality dimension	Codes and data
Accuracy	*Multiple modelling techniques:* "use the wrong tool to model something […] I can't just say there's a slab now, that's just a piece of geometry" (Digital engineer, C3I3); "when you try to extract 2D drawings from 3D BIM models, those drawings are not as correct and as detailed as they used to be" (Technical manager, C3I5) *Document control bottlenecks:* "it's no longer the most current version anymore by the time I'm reviewing it" (Project engineer, C3I1) "he keeps on updating but he hasn't he hasn't put it on the [CDE1]." (Technical manager, C3I5) "I just want to know where I can get my latest drawing" (Digital engineer, C3I3) "I think someone within the doc management system had obviously circumnavigated it somehow, to get the drawings out. And then when we were trying to get the said revisions for our set out, the system wouldn't allow it because directory hadn't been properly created." (Technical manager, C3I5)
Completeness	*Interoperability:* "transferring things […], you lose data" (Technical manager, C3I5) *Information loss during conversion:* "when you upload a PDF." (Information manager, C3I4) *Multiple modelling techniques:* "use the wrong tool to model something […] I can't just say there's a slab now, that's just a piece of geometry" (Digital engineer, C3I3); "when you try to extract 2D drawings from 3D BIM models, those drawings are not as correct and as detailed as they used to be" (Technical manager, C3I5) *Lack of process change:* "We're going to print it out, we're going to staple it together […] get three signatures, scan it back in, put it back into [CDE1] and submit it."(Project engineer, C3I1) "it's not actually speeding everything up, it's sort of making everything a lot slower; which I find very frustrating" (Principal engineer, C3I8) "can be very confusing when we have two platforms" (Technical manager, C3I5) "Everything had to be taken out of one data environment and pushed into another. One of the issues with that is the consistency or the compliance or knowing the latest versions of information" (Digital engineer, C3I6) "As the contractor, then we have to deliver it to a completely separate, disconnected CDE […] we're double-handling" (Digital engineer, C3I3) *Loss of constraint information*: "we physically need that information to know what we're building and what the constraints in building it are." (Project planner, C3I2). "Access chamber works will conflict with access road for pile work, piling work package has to be moved back 2 weeks." (Progress review meeting, C3M3); *Low level of detail:* "Work package for three spans were linked to a work order. Model showed the deck for a span was completed before the pier supporting it was completed because the work package for the first span was reported as completed." (Field notes- BIM Consultant 2, C2I5)
Timeliness	*Unstructured information sharing:* "when I have finished everything - by the way, we have this spread sheet" (Technical manager, C3I5) *Document control bottlenecks:* "it's no longer the most current version anymore by the time I'm reviewing it" (Project engineer, C3I1) "he keeps on updating but he hasn't he hasn't put it on the [CDE1]." (Technical manager, C3I5) "I just want to know where I can get my latest drawing" (Digital engineer, C3I3) "I think someone within the doc management system had obviously circumnavigated it somehow, to get the drawings out. And then when we were trying to get the said revisions for our set out, the system wouldn't allow it because directory hadn't been properly created." (Technical manager, C3I5) *Lack of process change:* "We're going to print it out, we're going to staple it together […] get three signatures, scan it back in, put it back into [CDE1] and submit it."(Project engineer, C3I1) "it's not actually speeding everything up, it's sort of making everything a lot slower; which I find very frustrating" (Principal engineer, C3I8) "can be very confusing when we have two

	platforms" (Technical manager, C3I5) "Everything had to be taken out of one data environment and pushed into another. One of the issues with that is the consistency or the compliance or knowing the latest versions of information" (Digital engineer, C3I6) "As the contractor, then we have to deliver it to a completely separate, disconnected CDE […] we're double-handling" (Digital engineer, C3I3)
Consistency	*Document control bottlenecks:* "it's no longer the most current version anymore by the time I'm reviewing it" (Project engineer, C3I1) "he keeps on updating but he hasn't he hasn't put it on the [CDE1]." (Technical manager, C3I5) "I just want to know where I can get my latest drawing" (Digital engineer, C3I3) "I think someone within the doc management system had obviously circumnavigated it somehow, to get the drawings out. And then when we were trying to get the said revisions for our set out, the system wouldn't allow it because directory hadn't been properly created." (Technical manager, C3I5)
Accessibility	*Interoperability:* "transferring things […], you lose data" (Technical manager, C3I5); *Unstructured information sharing:* "when I have finished everything - by the way, we have this spread sheet" (Technical manager, C3I5); *Drawings and file-based sharing:* "I use all the navigator tools that we've got here. But I prefer to use AutoCAD because I find it a lot easier" (Project engineer, C3I1) "It's not the best way because we haven't got the technology. I haven't got a big screen" (Project engineer, C3I1) *Lack of process change:* "We're going to print it out, we're going to staple it together […] get three signatures, scan it back in, put it back into [CDE1] and submit it."(Project engineer, C3I1) "it's not actually speeding everything up, it's sort of making everything a lot slower; which I find very frustrating" (Principal engineer, C3I8) "can be very confusing when we have two platforms" (Technical manager, C3I5) "Everything had to be taken out of one data environment and pushed into another. One of the issues with that is the consistency or the compliance or knowing the latest versions of information" (Digital engineer, C3I6) "As the contractor, then we have to deliver it to a completely separate, disconnected CDE […] we're double-handling" (Digital engineer, C3I3) *Loss of constraint information*: "we physically need that information to know what we're building and what the constraints in building it are." (Project planner, C3I2). "Access chamber works will conflict with access road for pile work, piling work package has to be moved back 2 weeks." (Progress review meeting, C3M3); *Low level of detail:* "Work package for three spans were linked to a work order. Model showed the deck for a span was completed before the pier supporting it was completed because the work package for the first span was reported as completed." (Field notes- BIM Consultant 2, C2I5)
Data Provenance	*Information loss during conversion:* "when you upload a PDF." (Information manager, C3I4); *Drawings and file-based sharing:* "I use all the navigator tools that we've got here. But I prefer to use AutoCAD because I find it a lot easier" (Project engineer, C3I1) "It's not the best way because we haven't got the technology. I haven't got a big screen" (Project engineer, C3I1) *Lack of process change:* "We're going to print it out, we're going to staple it together […] get three signatures, scan it back in, put it back into [CDE1] and submit it."(Project engineer, C3I1) "it's not actually speeding everything up, it's sort of making everything a lot slower; which I find very frustrating" (Principal engineer, C3I8) "can be very confusing when we have two platforms" (Technical manager, C3I5) "Everything had to be taken out of one data environment and pushed into another. One of the issues with that is the consistency or the compliance or knowing the latest versions of information" (Digital engineer, C3I6) "As the contractor, then we have to deliver it to a completely separate, disconnected CDE […] we're double-handling" (Digital engineer, C3I3)