EXPLORING THE MISSING ROLE OF DATA DICTIONARIES

S. F. Sujan
Norwegian University of Science and Technology, Trondheim, Norway
Mott MacDonald, Sheffield, United Kingdom
E. Hjelseth
Norwegian University of Science and Technology, Trondheim, Norway
M. Munir
University of Salford, Salford, United Kingdom

ABSTRACT: Automation is one of the best ways to realise (achieve) the benefits of digitalization. This requires a flow of information between multiple stakeholders. In this respect, data dictionaries can enable governed and trustworthy data. The technology for Data Dictionaries are available, but integrated and automated services are missing. Through an exploratory study on circularity with 11 practitioners in Norway, the importance of machine interpretable governed language and the lack of awareness and understanding of data dictionaries first appeared to the authors. This paper provides an overview of data dictionaries and their role in integrating decentralized data. A holistic view of how to achieve integrated data by connecting concepts such as data dictionaries, data templates and data libraries is derived. Three use cases of Data Dictionaries have been abducted from a combination of academic literature review and interviews. A significant research gap in academic literature on use of data dictionaries to exchange and map to enable integration was also found. Findings indicate a shift from the principle of a centralised data distribution to a decentralised data distribution.

1. INTRODUCTION

It is paradoxical that the degree of automation of digital operations are low, despite the availability of technology. This study explores the limited degree of automation from a process-oriented perspective to reveal some of the missing elements.

According to ISO23386-2020, a data dictionary is a centralized repository of information about data such as meaning, relationships to other data, origin, usage and format. For decades, Data Dictionaries have been used to govern interoperability in digital processes. In the construction industry, where there is a need for integrated data and exchange between multiple stakeholders in short time collaborations, one would expect that use of data dictionaries would be common knowledge and practice. In a highly fragmented decentralised project-based industry, the ability to reuse data depends on the agility and reliability of the definition of data structures. The difference between machine readable and machine interpretable is highly relevant in this paper. Machine-readable is structured data that can be directly processed by a computer. Machine-interpretable data has meaning with context and relations (RDMkit, 2024).

Building Information Modelling/Management (BIM) was developed with the intention of improving the inter-organisational collaboration in the construction industry as part of the project lifecycle. Through the development of BIM, the Industry Foundation Classes (IFC) are well known to the stakeholders that are associated with the design part of the asset lifecycle. IFC is an open data standard which supports exchange between various BIM software to enable interoperability (Laakso and Kiviniemi, 2012). Whilst the IFC format is a solution for exchange, the definition of the terminology that is applied within the structure can limit its use. Furthermore, it is important to note that the model-based workflow does not typically dominate the methods of working for some actors e.g., logistics (Whitlock *et al.*, 2021) and product manufacturing (Kebede *et al.*, 2022). This paper aims to ***explain the need for governance to enable machine interpretable integration.***

The Research Aims (RA) are:

RA1 - to explore whether decentralisation of built environment data is increasingly being accepted as an underlying premise.

RA2 – to abduct and discuss the importance of defining governed machine interpretable language i.e. using data dictionaries as a basis for integrated data.

This paper intends to raise discussion on the inter-organisational utilisation of multiple data dictionaries from decentralized data sources to achieve integrated data. Section 2 explains the research

methods employed and the role of abduction, Section 3 presents perspectives from literature and theories on data dictionaries. Section 4 presents 3 abducted use cases which were used as criteria in the Section 5 literature review. Section 5 presents the findings; the first part is based on literature and the second part based on interviews. Section 6 provides a discussion based on the findings, research limitations and future considerations.

2. RESEARCH METHODOLOGY

To enable a reliable triangulation of limited knowledge in an emergent area of research, mixed sequential methods were utilized, building plausible hypotheses from stage to stage. Mixed methods involve the use of both qualitative and quantitative methods to develop understanding (Creswell and Creswell, 2018).

The purpose of this method is to combine research approaches to provide a more accurate and complete understanding of a complex problem. The first round of interviews did not study data dictionaries, data templates or any associated concepts directly, rather an explorative unstructured approach by convenience sampling to understand how digitalization could enhance sustainability outcomes in the built environment. A broad range of interviewees were selected, 11 in total. All interviewees had an active role in practice in Norway; 4 from Technology & Data providers, 3 from Public Clients, 1 from a Building Authority and 3 Design Consultants. From these participants, 3 out of 11 used data dictionaries actively in industry applications. In this stage of data collection, analysis was conducted using thematic analysis which showed that all participants described the need for governed language interpretation, of which 3 participants took this further in application of the underlying semantic concepts (data dictionary, templates or library). This led researchers to a plausible hypothesis that the role of language and more explicitly in some interviews, data dictionaries were important to close the circularity loop of information.

A scoping literature review was conducted and experts interviewed focusing on the use of data dictionaries. The expert interviews involved members of BuildingSMART, and a semi-structured approach was utilised. The first expert was involved in the day-to-day developments of Building Smart Data Dictionary. The second expert was a senior member of BuildingSMART with significant experience in digitalization for the last four decades. The literature review led researchers to find that knowledge of data dictionaries related to development of new services is limited within the domain. Therefore, authors took a step back to explore research areas underpinned by a decentralized data premise in the domain (Section 3.2). After finding evidence that indicates support to this claim and due to limited academic knowledge on the role of data dictionaries, a positivist philosophy was deemed unsuitable – there are not enough studies or experts to numerically (dis)confirm hypotheses.

Therefore, the philosophical stance taken by the research is best described as critical realism. Bhaskar (1975) formulated critical realism; it is an alternative to interpretive and positivist philosophies, to provide researchers with new ways of describing the way knowledge is developed (Wynn and Williams, 2012). In other words, it combines a general philosophy of science with a philosophy of social science to describe an interface between the natural and social worlds. The construction industry is highly fragmented, making the social dimension important in drawing conclusions or plausible hypotheses. Essentially, critical realists believe that ‘what is real’ is not reducible to our knowledge of reality (Fletcher, 2017).

The role of the researcher is to construct a narrative rather than discovering the truth (Cruickshank, 2003). It occupies a unique space between the polarizing perspectives of positivism and interpretivism, critical realism emerges as a particularly pertinent framework for investigating the intricate interplay between technology and human actors within the domain of information systems (Zachariadis, Scott and Barrett, 2010).

This paper predominantly employs abduction, a mode of inference characterized by its ability to generate plausible explanations for observed phenomena, even when certainty remains elusive (Merriam-Webster, 2022). Envisioned as akin to solving a captivating mystery, abduction utilizes available clues to identify the most likely explanation, yet acknowledges the inherent uncertainty (Bryant and Charmaz, 2007).

Abduction serves as the vital engine for knowledge expansion, even while readily admitting its inherent vulnerabilities. Consequently, abduction wields potent investigative power, yet demands cautious application to navigate the nuanced realm of the "potential explanation" (Meyer and Lunnay, 2013).

3. PERSPECTIVES IN DD RESEARCH

3.1 Theoretical Point of Departure

3.1.1 Underpinning use of language

Ogden’s Triangle, also known as the Triangle of Reference or Meaning in Figure 1, is a visual model developed by Ogden and Richards (1923). It represents the relationship between three key components of meaning in the context of human interpretation.

- Symbol: The word or linguistic representation.
- Reference: The mental concept evoked by symbol.

- Referent: The real-world object or entity the symbol refers to.

Ogden's Triangle illustrates the complex interplay between words, thoughts, and the real world. The relationship between the symbol and the referent is indirect. This means that words don't directly point to things in the world. Instead, they evoke mental concepts that are then associated with things. The relationship between the symbol and the thought or referent is also not fixed. The same symbol can evoke different thoughts in different people, depending on their individual experiences and knowledge. Ogden's triangle demonstrates how human interpretation can create inconsistencies in definition of language. This indicate why use of a Data Dictionary is needed as default.

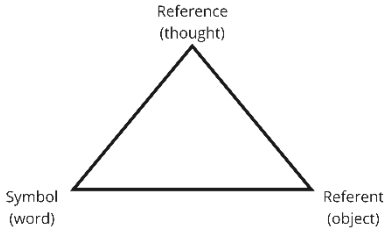

Figure 1. Ogden's Triangle (Ogden and Richards, 1923)

3.1.2 Interoperability Governance

The European Interoperability Framework (EIF, 2024) in Figure 2 is used to demonstrate the role of governance and the need to unify technical, semantic, organisational and legal interoperability to enable integrated services.

Semantic and technical interoperability has been dominating the focus of interoperability. Whilst in principle, this solves machine interoperability challenges, in practice, limited application in industry may be due to missing focus on legal interoperability (i.e. standards, regulations), and organizational interoperability (understanding in different disciplines and contexts). The joint understanding from EIF could enable machine interpretation. It is also important to note the cross-cutting element 'Integrated service governance' and the overall scope 'Interoperability governance'.

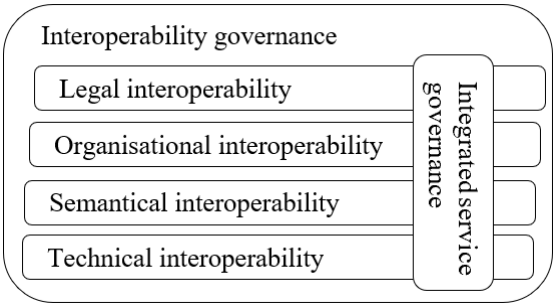

Figure 3.European Interoperability Framework (EIF, 2024)

3.2 Scoping Literature Review

Figure 3 shows the distribution of searches done in scoping topics that (dis)confirm the notion that the focus of the domain is evolving underpinned by a growing belief in accepting the decentralization of data in the sector.

From the Figure 3, the following can be drawn:

1. In the early 1900s, there was increasing research in information management, toppled by the hype of a new topic, it took a dip until research disseminated and awareness grew. Arisen further in the early 2000s with significant increase in attention. In 2011, the UK BIM mandate was in alignment indicative by earlier mass dissemination.

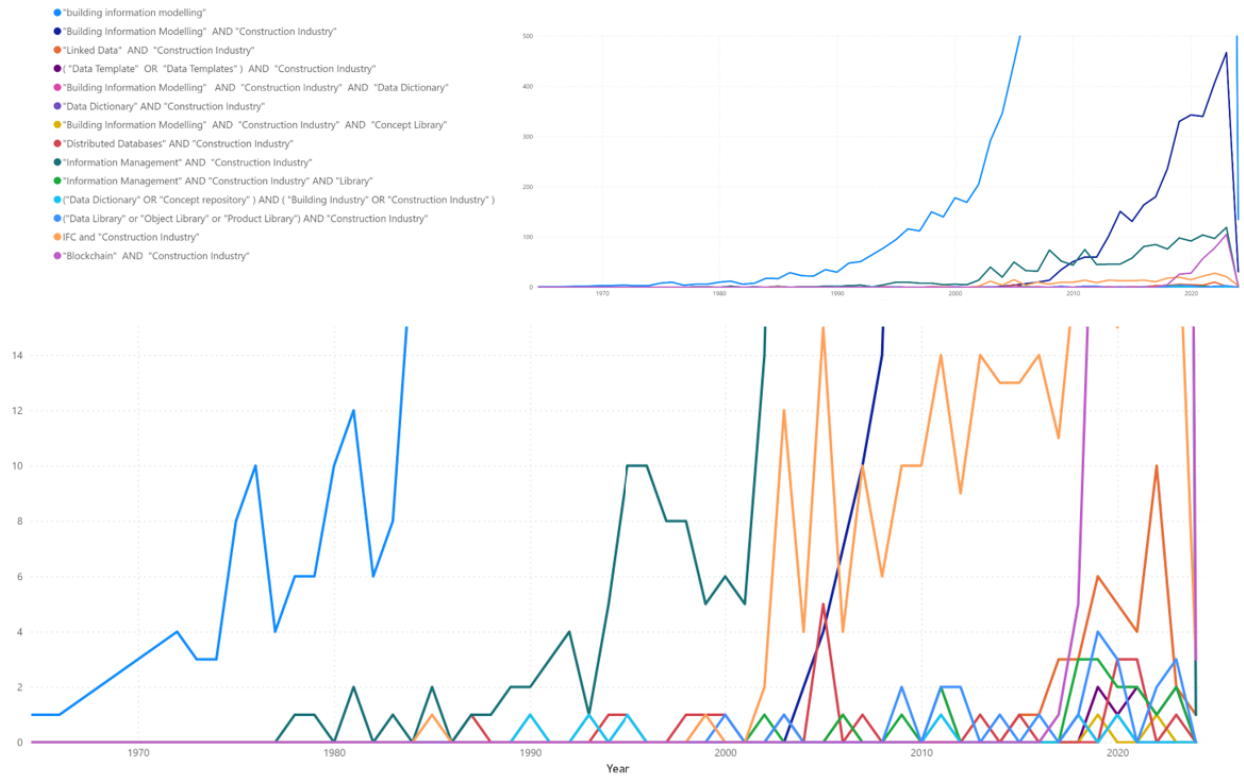

Figure 2. Paper distribution in literature review

2. The rise of linked data research is on the premise that decentralized data is to be connected (Pauwels *et al.*, 2018).
3. The rise of blockchain's applicability to the sector also fundamentally bases its application on connecting decentralized data in a reliable and trustworthy manner (Shishehgarkhaneh, Moehler and Moradinia, 2023).
4. Increased research activity in data templates, data dictionary, data library were found.

3.3 Decentralised vs Centralised.

In this section, two extremes are drawn on a centralized and decentralized data distribution. On the centralized end, the focal point is the Project Information Model (PIM) or the Asset Information Model (AIM) – an abstraction described in ISO19650 which could be applied between these extremes in interpretation. Although all the data in the PIM or the AIM is not owned by one stakeholder, in this fragmented environment where information hoarding is a practice, it is often thought of that way i.e. manufacturer information about a product is often copied into the PIM rather than being live linked to data hosted by the manufacturer.

On the other hand, the decentralized distribution is all about connecting data held separately but defined uniformly using governed machine interpretable language; data can be owned by multiple stakeholders, but the meaning within a single context is owned by a single stakeholder at a given time. This means that data can serve multiple contexts by adding layers of meaning, one of these contexts is the AIM or the PIM.

Beetz (2014) discusses the challenge of standardizing content in the building and construction industry by categorising system organization of distributed, interconnected Data libraries. Unlike traditional centralized approaches, this network aims to improve collaboration, sustainability, and adaptability while adhering to established standards.

The decentralised system comprises a network of data libraries structured in tiers: international, local, and organizational. This tiered structure allows concepts to be inherited and refined according to specific contexts and regulations, ensuring both standardization and customization. Standard building and construction concepts reside at the international level, providing a foundational layer for all lower tiers. Local and organizational libraries then build upon these concepts, tailoring them to their specific needs while referencing and adhering to the established international standards. According to Beetz (2014), this decentralized approach offers several advantages for content standardization including increased collaboration, improved adaptability and enhanced trust and transparency.

4.1 HOLISTIC OVERVIEW

In Figure 4 a depiction of the links between the Data Dictionaries (DD) to Data Templates (DT) to Data Libraries (DL) are shown. Data dictionaries provide governed meaning and machine-readable unique identifiers to the language used to define data structures in data templates. Data templates are a collection of standardized sets of language relating to a specific context e.g., Product data templates, Environmental Product Declarations (Mêda, Sousa and Hjelseth, 2020). Data libraries are the closest to the context, they are also known as BIM libraries, object libraries or product libraries, the PIM, or the AIM. Data libraries apply combinations of data from templates based on the contextual needs from non-digital standards (e.g., regulation, local standards & practices, client requirements, ISO standards).

"… to <map data to> model you need a dictionary as a starting point. And then you need templates or groups or properties actually the data in the template standard allows you … say what is the information on an object … the data template has the capability to filter information … <including> the LOIN standard …<to filter>… purpose and actor …" Technology & Data Provider

The novelty that data dictionaries bring is that the language that is used to define parts of data structures that exist in various databases is standardized and made machine interpretable through a governance process ensuring consistency. This means, human interpretation of a standard which opens many possibilities (Ogden's triangle) is negated by having a governed machine-interpretable list of definitions which are given their own unique identifiers.

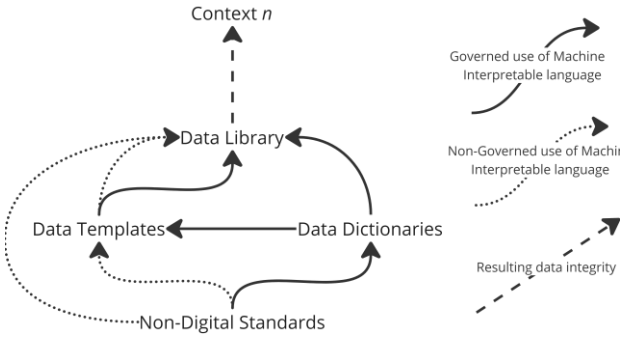

Figure 4. The Semantic Trio - Connecting DD, DT, DL

4.2 USE CASES OF DATA DICTIONARIES

This section presents the abducted use cases which were search criteria in the literature review.

4.2.1 Referential – to find knowledge

From the most popular of use cases especially in the context of bSDD. Several authors used bSDD to understand the IFC structure. This use case relies on the interpretation of the user and the data dictionary merely provides governed definitions. The users in this use case then interpret the data from the dictionary and do not create links between machine interpretable information.

"…just an example here. We want to talk about a window. And burglar resistance. So, the question is, how do we map burglar resistance to IFC 2.4 … this is more like a functionality, because all the govern-

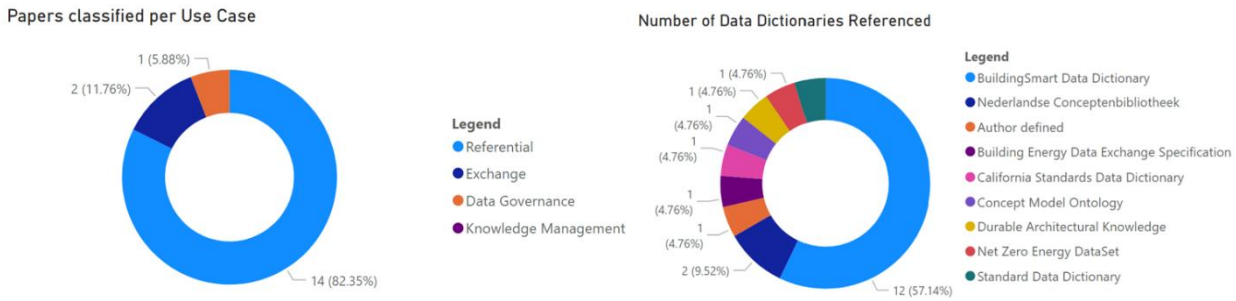

Figure 5. Distribution of Papers within Acceptance Criteria n=17

ments are talking about IFC. So, let's make a relation between the object world and the data template like mapping solution. So, what we saw is that when we extracted from the web page of BuildingSMART… everyone was <interpreting it> differently…" Technology & Data Provider

4.2.2 Mapping and Exchange for Integration

The use of data dictionaries to translate between data structures depending on geographical location and languages. Böger et al., (2018) describes a solution concept with exchange between databases that are governed by standardised definitions as intended by data dictionaries, the exchange of data is not about exchanging between identical data structures.

"…a project to build a bridge between Norway and Finland, and you want to hire construction workers from Poland.. the same thing in Norway is called differently in Finland and they can see that... Even though they have different agreements... At some point they need to have common models." Expert 1

4.2.3 Governance and Trustworthiness

This use case enables trusted structures of information by ensuring compliance with relevant standards, legislation, requirements etc. For example, if a data dictionary is developed for product data, the meaning of the language utilized is sourced from standards or requirements relating to Construction Product Regulations and is then traceable through the data dictionary. This means that if information requirements are defined by linking to data dictionary content (machine interpretable), this reduces the subjectivity in between standard and application.

5. FINDINGS

5.1 Data Dictionary Scoping Literature Review

Academic literature was reviewed to gain an understanding of the use and application of data dictionaries. The following keywords used on Scopus:

TITLE-ABS-KEY ("data dictionary" OR "concept library") AND ("construction industry" OR AEC OR building))

This returned a total of 46 papers between 2014 and 2023. From the 46 papers, 17 papers were relevant. The papers were excluded based on relevant to the construction industry (e.g., 25% of papers were healthcare related) and the papers should mention data dictionary or concept library in connection to the use of data in the industry. Eight different data dictionaries were found with this search criteria. The reference to bSDD dominated the search result as over half of the papers referenced it. Abduction of use cases presented in Section 5 were utilized to categorize the use cases as shown in Figure 5. From this figure the data dictionaries are being used predominantly to refer to knowledge (82%). Whereas 12% of papers were about exchange of data but were conceptual with limited practical validation. Only one paper focused on governance. Furthermore, about 65% of papers were 3D modelling focused with about 25% of papers being both 3D modelling and physical product focused.

5.2 Interviews of digitalisation experts

This section provides a narrative with evidence from the interviews. As the interviews were exploratory and limited in predisposed structure, it is not reliable to draw statistical data. For abductions made in this paper, the narrative is crucial.

The history of digitalization paints a picture of moving from a highly controlled centralized setting to one that is centralizing within a decentralized environment due to changes in hardware.

"We went from large mainframes systems, centralised a degree to being completely decentralised… Its really the history of the hardware <that changed>…" Expert 2

It was also clear that other industries have used the data dictionary concept, however, the construction industry has been blinded by a project specific focus – limiting reuse of data in other frames of reference.

"15 years ago, the idea already existed in a different form in oil & gas, ISO 15926… but for the construction sector its <gone in the direction of> an organisation making a reference data library for a project... Many organisations have been in the habit of writing, their own CAD <and> … BIM standards, and part of that, its got used to writing their own dictionary <although they do not call it that>" Expert 2

When speaking of reuse of data, an expert attempted to ask stakeholders to fill in data templates to no avail. This resulted in the realization that data dictionaries are needed to map between the contexts of stakeholders.

"PDT <Product Data Templates> … the idea is to build up data templates… I just tried to give these templates to the market… the manufacturer feeds in the numbers and so on… the funding is also there to do this… But they were actually struggling with this… now they understand without the data dictionary, this is not possible… we are not able to digitalize so much without having this dictionary." Building Authority

The buildingSMART Data Dictionary (bSDD) is an aggregator of data dictionaries. Alignment was seen in the use of bSDD, and the literature reviewed in this paper, where majority of bSDD users use the DD to refer to data structures as compared to exchange and governance.

"<bSDD> is not one dictionary…99% <of users are referential use case>, I would say, and 1% are those that do the mapping <exchange use case>…" Expert 1

The exchange use case was evident from the description of the differences of use of language to describe common elements – the differences in understanding of the standards that exist is limiting ability to communicate effectively.

"…Source A; diameter distance between the inner walls of a pipe source the pipe diameter, Source B; diameter distance between the outer walls of a pipe… we are working with manufacturers of pipes that are <frustrated> because every time they get something, its called Diameter… and they have no clue what they're talking about. So is the source credible?.. Is it defined correctly?" Technology & Data provider

Furthermore, the importance of governance; the quality and trustworthiness of data is enhanced by linking machine interpretable information that is underpinned by data dictionary. Here with European Technical Information Model (ETIM, 2024) as example.

"You know that if you are delivering models at the handover, you need to have ETIM codes in it… you can put that data manually… but first of all, its very prone to errors. You can have typos… much easier if your software provides you with a drop-down list of only possible values and… what bSDD is providing is where you can draw that information" Expert 1

6. DISCUSSION

6.1 The need for governed language interpretation

Data dictionaries have traditionally been used to underpin systems within a single organisation in the backend of databases (Rashid *et al.*, 2020). However, there is a difference in how it is being used in the construction industry from an organisational perspective as its content is applied and developed between organisations/actors due to the nature of the industry and its product. The built environment focus is therefore on aggregation of data dictionaries from various governed interpretations of language from e.g., standards. The recent evolution and new understanding of bSDD exemplifies the need to view bSDD as a data dictionary aggregator (service) to enable the language to underpin integrated machine interpretable requirements (BuildingSMART, 2023). This is also the reason that bSDD refers to itself as a 'service' and not a 'standard' as many still interpret. This is in comparison to the original centralised understanding of bSDD as being a single data dictionary to serve the industry. The construction industry is highly fragmented project-based inter-organizational network that has data dependencies between various organisations and actors. This means that to integrate data between multiple parties, there is need for definition of machine interpretable common language that is accepted in the industry through governance processes (Cobuilder, 2021).

From early interviews on digitalisation and sustainability, the ability to integrate, collaborate and use data from various sources became a core premise that all interviewees mentioned within the realms of their interaction within the topic of digitalisation. This resulted in questions on how data integration from various sources could be enabled. Bringing about, agreeability in standardisation as being an important part of this process from all participants. However, when further probed, the gaps in interviewee understanding started to show its limits. On the premise all interviewees agreed that there needed to be agreed methods of defining and sharing data. However, the technical semantic language was predominantly found in approx. 25 % of the interviewees with some variety. Only two stage 1 interviewees could explain data templates, data dictionary and how it links to create a vision that integrates information in various contexts for various purposes e.g., circularity**.** This showed us the lack of understanding in the 'how' of achieving integrated data in different frames of reference, e.g., single asset, multi asset, multi owner, city level, national level.

A high degree of focus has been put on standardising data structure definition e.g., the definition of IFC, EPD, PDT. However, keeping in mind that these are defined not completely in a machine-interpretable format. There is still a degree of ambiguity as these concepts from various standards are applied in reality. Data dictionaries open the possibility to narrow the gap in ambiguity in how standards of various kinds are interpreted to their definition in a machine-interpretable format (Cobuilder, 2021).

Regarding sustainability, there is need to integrate data for new purposes as business models adapt for reuse workflows. The Built Environment is responsible for 39 % of gross carbon emissions (Global Alliance for Buildings and Construction, 2019). Digitalisation represents an enabler for change due to the need to integrate information flow for circular workflows (Tomczak *et al.*, 2023). In many countries, new workflows are expected to be introduced, driven by legislation for reuse of building products and disclosure of both operational and embodied carbon emissions (Debacker and Manshoven, 2016). In this respect, Environmental Product Declaration (EPD) data is expected to be used by various actors to fulfil regulatory requirements. Recently, Construction Product Regulations (CPR) have been updated to include product environmental sustainability performance data through Digital Product Passports (DPPs) (GS1 2023). CPR represents a driver for flow of information outside the

model-based environment, enabled by DPPs. The question remains on how to integrate this data into models that serve a variety of purposes. By having EPD data in databases (EPDNorge, 2023), this data can be linked to the objects in the models by using unique identifiers – the data does not need to be hosted in the model(s). Furthermore, its important to note that proprietary tools are not made to integrate data in a decentralized manner, there is a current lack in capability to map and link data.

"…exchange of data using IFC…let's mix data and model…Not a good idea, this is what we've been doing…<If> we need <to model> in Revit, put data in, but it should come from a dictionary, because that's actually where we can say that this comes from that <standard/requirement with interpretation governed by definition in a data dictionary>" Technology & Data provider

RA1 focused on finding evidence that the industry is in an evolution of creating capability to centralise information (within a single context) in a decentralised data and organisational environment. From the literature review, there is more research exploring topics with the underlying premise of integrating trustworthy data from decentralised sources. This implies that, when it comes to digitalisation in the construction sector, decentralised technologies are available, but are not utilised as intended. Even though the sector is decentralised, it still relies predominantly on centralising information but not typically by linking data in a decentralised environment. However, the experts and practitioners interviewed indicated that this was the future of the industry because of the need to close the information circularity loop and meet regulatory requirements i.e. Construction Product Regulations. Overcoming the norm of hoarding information into a single federated environment to meet requirements defined by a client in isolation without machine interpretable definition of data presents a current challenge. The federated environment is not the problem, the definition of information within that federated environment if to be reused in different frames of reference, requires definition based on governed and accepted machine-interpretable language. This would enable ability to scale up datasets from single asset to multi-asset. This leads to discussions on the need for use of governed interpretation of language through data dictionaries.

RA2 was achieved by abducting three use cases from empirical data to assess the application in academic literature. It is clear from the findings that academic application of the data dictionary concept is mainly referential meaning that researchers primarily use data dictionaries to reference a data standard. The second use case was in exchange and mapping data between standards. The third use case was in enabling improved data quality and trustworthiness by basing data on governed interpretation of language. According to the literature review, there is a research gap in understanding how data dictionaries can enable a future where data is linked between systems, models, databases etc. with governed interpretations of language.. The practical gap is in the lack of awareness of data dictionaries i.e. only approx. 25% of practitioners interviewed could explain the semantic trio in Figure 4. Therefore, missing software capability and services to integrate information from data dictionaries is not surprising as the customers are currently unable to define their needs appropriately. This may be due to a combination of lack of awareness, and difficulty in understanding other levels of the European Interoperability Framework (Figure 2); the underlying semantic trio. Furthermore, a plausible root cause may be that most processes in the industry do not actually need to be fully integrated and automated, the semi-automated approach fulfils current needs.

6.2 Research limitations

The paper aimed to construct a narrative based on empirical data and literature. Abduction as a mode of inference has limitations in ability to justify the narrative produced. Therefore, the findings in this paper should be viewed as plausible hypotheses that require further validation. The scoping literature review was heavily dependent on the key words utilised, whilst every effort was made to capture papers as widely as possible, there may have been keyword combinations that were not considered which could have had an impact on findings.

6.3 Future Research

This paper raises a number of questions that need to be considered when developing future research;

1. Further evidence is required to test the hypothesis that the construction industry is centralising data in a decentralised environment rather than accepting the environment and using technology to integrate data from decentralized sources. This means that a mindset shift is necessary to enable adoption of integrated services. There is need for further validation.
2. Use cases for exchange between multiple data structures underpinned by data dictionaries needs to be developed to demonstrate applicability based on real projects, to disseminate value realised.
3. There appears to be two different perspectives, product and model-based workflows. To enable integration between these two perspectives, it is important to make data in the model-based workflow more accessible to actors working with non-geometrical represented databases. Not every actor needs to view the information as a geometric model(s). Practice needs methods to link information from model-based workflows into non-geometrically represented databases to enable increased accessibility.
4. Governance of the interpretation of standards, requirements etc. that enable trustworthy data sets in data dictionaries have a major role to play. There was limited research in connecting data governance to the role of data dictionaries in real projects. When doing research on governance, the EIF is a suitable point of departure to ensure that governance goes beyond technical interoperability, considering organisational and legal needs in definition of language.

7. CONCLUDING REMARKS

The paper explored the notion of evolving belief in the industry of the acceptance of decentralized data because this is underlying premise that leads to the need for data dictionaries. From the scoping review on decentralization, trends to support the claim were found to be the increase in research interest in linked data and blockchain. Furthermore, increase in research focus over the last 5-6 years in topics such as data templates, data dictionaries and data libraries further support the claim that the underlying beliefs are ones that support the decentralized nature of data in the sector. Underpinned by the evidence from literature on the growing theme of data decentralization, the use of language becomes important in sharing knowledge, information, and data between different parts of the industry. This is increasingly important in the built environment as there is need to share between the different parts of the industry that contribute to the circular flow of physical products underpinned by information flow, requiring new services.

The data dictionary plays a vital role in ensuring that machines read and interpret information in a reliable manner, this means that humans need to agree on how information is interpreted (from standards, requirements, legislation) prior to using it to describe a physical product, property, object etc. Three use cases were abducted from literature and interviews. Data dictionaries were shown to be primarily used in referencing knowledge/data structures. Rarely was an example of using data dictionary to translate between data structures found in either literature or interviews.

Regarding interoperability, most focus has been on the semantic and technical solutions enabling machine readability. Machine interpretability can be enabled by giving priority to legal and organizational interoperability. Use of a (single) Data Dictionary in a digital solution (software) motivates a focus on technology. However, as per the nature of the industry and various origins of data, utilisation of Data Dictionaries (multiple) is about integration from decentralized data sources. The Semantic Trio presented in this paper draws the underlying connections between semantic concepts to facilitate exchange of data. The Authors recommend that change starts from information requirements using governed machine-interpretable language from data dictionaries. By enabling new trustworthy integrated services, fully automated processes can be developed to realise the benefits of digitalisation.

8 REFERENCES

Beetz, J. (2014) 'A scalable network of concept libraries using distributed graph databases', in *Computing in Civil and Building Engineering - Proceedings of the 2014 International Conference on Computing in Civil and Building Engineering*, pp. 569–576. doi: 10.1061/9780784413616.071.

Bhaskar, R. (1975) *A realist theory of science*. doi: https://doi.org/10.4324/9780203090732.

Böger, L., Huhnt, W. and Wernik, S. (2018) 'Standardized names for object types and attributes as basis for cooperation between planning and building', *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10864 LNCS, pp. 307–324. doi: 10.1007/978-3-319-91638-5_17.

Bryant, A. and Charmaz, K. (2007) *The Evolving Nature of Grounded Theory Method: The Case of the Information Systems Discipline*. 1st edn, *The SAGE Handbook of Grounded Theory*. 1st edn. Los Angeles: SAGE Publications. doi: 10.4135/9781848607941.n16.

BuildingSMART (2023) *BSDD Documentation - Read me, BSDD Official Repository*. Available at: https://github.com/buildingSMART/bSDD/tree/master?tab=readme-ov-file#readme.

Cobuilder (2021) *Standardising Construction Object Data for Digital Use*. Lysaker. Available at: https://cobuilder.com/en/content-offers/standardising-construction-data-for-digital-use/.

Cruickshank, J. E. (2003) *Critical Realism: The Difference It Makes, Journal of Critical Realism*. Routledge (Routledge studies in critical realism: 6).

Debacker, W. and Manshoven, S. (2016) 'D1 Synthesis of the state-of-the-art', *Key Barriers and Opportunities for Material Passports and Reversible Design in the Current System (H2020 BAMB project)*.

EIF (2024). European Interoperability Framework (EIF) Toolbox https://joinup.ec.europa.eu/interoperable-europe

EPDNorge (2023) *EPD-Norway Digi, EPD-Norway Digi*. Available at: https://digi.epd-norge.no/.

ETIM (2024). European Technical Information Model, https://www.etim-international.com/

Fletcher, A. J. (2017) 'Applying critical realism in qualitative research: methodology meets method', *International Journal of Social Research Methodology*. Taylor & Francis, 20(2), pp. 181–194.

Global Alliance for Buildings and Construction, I. E. A. and the U. N. E. P. (2019) '2019 Global Status Report for Buildings and Construction: Towards a Zero-Emission, Efficient and Resilient Buildings and Construction Sector'. United Nations Environment Programme Nairobi, Kenya.

GS1 2023. Digital Product Passport (DPP) and the adoption of GS1 standards for the identification of construction products.

ISO 23386:2020, Building information modelling and other digital processes used in construction - Methodology to describe, author and maintain properties in interconnected data dictionaries, https://www.iso.org/standard/75401.html accessed 3 July

2022.Kebede, R. *et al.* (2022) 'Integration of manufacturers' product data in BIM platforms using semantic web technologies', *Automation in Construction*. Elsevier, 144, p. 104630.

Laakso, M. and Kiviniemi, A. O. (2012) 'The IFC standard: A review of history, development, and standardization, information technology', *ITcon*, 17(9).

Mêda, P., Sousa, H. and Hjelseth, E. (2020) 'Data Templates—Product Information Management Across Project Life-Cycle', in J. M. P. Q. Delgado (ed.) *Sustainable Materials in Building Construction, Building Pathology and Rehabilitation 11*. Springer. Cham: Springer International Publishing, p. 168. doi: 10.1007/978-3-030-46800-2.

Merriam-Webster (2022) *'Deduction' vs. 'Induction' vs. 'Abduction', Merriam-Webster*.

Meyer, S. B. and Lunnay, B. (2013) 'The application of abductive and retroductive inference for the design and analysis of theory-driven sociological research', *Sociological Research Online*, 18(1), pp. 86–96. doi: 10.5153/sro.2819.

Ogden, C. K. and Richards, I. A. (1923) 'The meaning of meaning: A study of the influence of thought and of the science of symbolism'. Harcourt, Brace & World, Inc.

Pauwels, P. *et al.* (2018) 'Linked Data BT - Building Information Modeling: Technology Foundations and Industry Practice', in Borrmann, A. et al. (eds). Cham: Springer International Publishing, pp. 181–197. doi: 10.1007/978-3-319-92862-3_10.

Rashid, S. M. *et al.* (2020) 'The Semantic Data Dictionary – An Approach for Describing and Annotating Data', *Data Intelligence*, 2(4), pp. 443–486. doi: 10.1162/dint_a_00058.

RDMkit (2024) *Machine actionability*. Available at: https://rdmkit.elixir-europe.org/machine_actionability.html#:~:text=Machine-interpretable%3A machines can put,and the Linked Data concept.

Shishehgarkhaneh, M. B., Moehler, R. C. and Moradinia, S. F. (2023) 'Blockchain in the Construction Industry between 2016 and 2022: A Review, Bibliometric, and Network Analysis', *Smart Cities*, pp. 819–845. doi: 10.3390/smartcities6020040.

Tomczak, A. *et al.* (2023) 'Challenges and experiences with the reuse of products in building design', in *ECPPM 2022-eWork and eBusiness in Architecture, Engineering and Construction 2022*. CRC Press, pp. 53–59.

Whitlock, K. *et al.* (2021) '4D BIM for construction logistics management', *CivilEng*. MDPI, 2(2), pp. 325–348.

Wynn, D. and Williams, C. K. (2012) 'Principles for conducting critical realist case study research in information systems1', *MIS Quarterly: Management Information Systems*. JSTOR, 36(3), pp. 787–810.

Zachariadis, M., Scott, S. and Barrett, M. (2010) 'Designing mixed-method research inspired by a critical realism philosophy: A tale from the field of IS innovation'.