Concept for digitally AI-readable buildings for urban mining based on digital building logbooks using large language models

Marcel Heiß[1], Benedikt Kandler[1] and Uwe Rüppel[1]
[1]Institute of Numerical Methods and Informatics in Civil Engineering, Darmstadt, Germany
E-mail(s): heiss@iib.tu-darmstadt.de

Abstract: The construction sector holds great potential for contributing to climate neutrality through circular economy strategies such as Urban Mining. But reuse of building materials often is hindered by a lack of structured, machine readable information on the existing building stock. This paper presents a concept for enhancing Digital Building Logbooks (DBL's) through the combined use of Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG). By semantically processing unstructured documents such as reports and invoices, the proposed approach enables AI-readable DBL's that facilitate data-driven decision-making for Urban Mining. The study analyses DBL's and current barriers like lack of accessible and interoperable data sources and static nature of DBL and counters them with introducing a system architecture that transforms unstructured building documentation into a dynamic DBL. This AI-enabled DBL approach significantly enhances data accessibility, supports material inventories and fosters reuse scenarios in the built environment.

Keywords: Large Language Models, Urban Mining, Circular economy, Digital Building Logbook

DOI: 10.18154/RWTH-CONV-254913. Published in the conference proceedings of the 36. Forum Bauinformatik 2025, Aachen, Germany.

1 Introduction

The construction sector is responsible for a considerable portion of resource consumption and waste generation, and also emits significant levels of greenhouse gases [1], [2]. Concurrently, the built environment constitutes society's largest resource stock. Urban Mining offers a pathway to reduce the high consumption and dependency on virgin materials, by recovering and reusing resources, as well as integrating them into a circular economy (CE). A major barrier is the lack of information required. For CE and urban mining, creating a material inventory of individual buildings and representing it digitally at city level through a GIS-based material cadastre is the first step towards knowledge-based decision-making, followed by digital building material passports and digital product passports, which quantify the sustainability of buildings, components and materials. Currently, extensive on-site recording and manual information extraction from documents are necessary for existing buildings to obtain this information and create detailed digital building material passports. Consequently, the idea of a whole-life-cycle repository of building information in the form of DBL's is becoming increasingly

important in order to gather and consolidate building data. Especially documentation of existing buildings often consists of unstructured text documents. These documents are part of the current state of DBL development, but are not analysed further. Concurrently, the performance of LLM's is witnessing a significant enhancement by handling unstructured text data. This research proposes a framework for using LLM in RAG to handle unstructered text data in DBL environment.
This work comprises a conceptual study beginning with the current state of development of DBL and LLM. The following consolidation analyses the shortcomings of current DBL implementations and examines their approaches to integrating unstructured, heterogeneous data sources to build DBL's, highlighting the research gap and its potential. Building on this, a DBL framework is developed that applies RAG and LLMs to extract structured data from unstructured documents and transfers gathered data into a vector that represents the "building construction DNA". After that the approach is analysed in a discussion with regard to solving the identified deficits and its potential for urban mining.

2 Related Work

2.1 Digital Building Logbook

A DBL is defined as a centralised and common repository of building-related information in digital form that can be used for effective decision making throughout the life cycle of a building [3]. It dynamically collects, stores and makes available information on a variety of aspects of a building, including design, construction, quality and environmental and performance criteria. The continuous collection of data provides a basis of information on the current status as well as the history and thus offers transparency and accessibility to this information for different stakeholders compared to traditional paper-based building records. Synonyms include 'property logbook', 'building passport' and 'electronic building file', but it is distinct from the terms 'digital building material' and 'building resource passport', which only contain essential circularity information for each individual building. The research projects European Union Building SuperHub (EUB SuperHub) defined a DBL data structure and thus laid the foundation for the general use of DBL's [4]. To determine the characteristics of a building, the main categories of DBL containing individual properties are defined. Examples of central main categories are general and administrative information, building performance and building operation and usage and can be extended by main categories like Smart Readiness and Finance [4]. Current Research has focused primarily on the application of DBL's in the context of renovation planning and energy efficiency to achieve sustainability [4], [5]. To achieve climate neutrality, the focus is also shifting to CE in the construction industry and is the subject of research in the Demo-Blog project [5]. While current DBL approaches are based on machine readable data, which are gained from existing building data sources or user inputs and follow a determined data flow [6], a large part of the information is currently still recorded in the form of documents and is therefore not available for current approaches. Recent developments in LLMs provide new methods to extract and utilise information from unstructured data.

2.2 Large Language Models

LLMs represent a paradigm shift in natural language processing (NLP) driven by the transformer architecture, increased computational capabilities and the availability of large-scale training data [7]. It represents a type of artificial intelligence model trained on massive amounts of text data to perform

various natural language processing tasks. The performance of LLMs correlates strongly with the number of parameters, the amount of data and the computing power [8]. Performance for specific applications can be increased through pretraining and fine-tuning. While pretraining is usually carried out in autoregressive mode on huge amounts of text, fine-tuning can be general or domain-specific. Despite these methods, LLMs face limitations in their reasoning abilities. In so-called hallucinations, models can generate plausible-sounding but false statements [9]. Limiting the number of tokens that can be processed results in a limited context scope and often the training data is not up-to-date so that information after the training period is unknown. This is why approaches such as RAG were developed. RAG combines LLMs with external knowledge sources through embedding search and dynamic information integration [10]. First, the text documents and the query are vectorised using an embedding encoder and stored in a vector database. In the initial retrieval step, relevant documents are extracted from a knowledge database via vector search. In the augmented step, the content found is transferred to the LLM together with the original query. The LLM then generates a response based on the updated, fact-based information in selected documents.

3 Consolidation of DBL and combination with AI

Previous approaches of DBL are mainly concerned with the content and structure of DBL's but don't define a procedure to automatically gather, store and share information about the lifecycle of a building [11]. Literature identifies national or regional databases, BIM models, building monitoring-related sources and European initiatives as the four main categories potential DBL data sources. Building documents often part of a DBL in the form of a document repository [12]. Current governmental data relevant to DBLs is often incomplete, not interoperable, and not openly accessible [6]. Core sources such as cadastral data provide basic building attributes but seldom contain information relevant to urban mining, such as material composition [6] . However, this information can be derived from detailed descriptions of structures contained in building performance and thermal insulation certifications, as well as from plans from which quantities can be extracted. In conclusion, more information needs to be collected on the actual condition of the buildings [6]. For new buildings, BIM facilitates the creation of machine-readable data, but such information frequently remains siloed within specialist disciplines and is rarely transferred in a structured form into the operational phase. [11]. For existing buildings, full digital capture (e.g., as-built BIM) is uncommon, further limiting DBL digitisation levels. The evaluation of different initiatives with regard to their online accessibility and level of digitisation showed that the majority of these only represent paper-based documents and only a few approaches allow a digital twin integration and the fewest allow an automatic input of data for example from BIM model [13].
The first part of the analysis shows current deficits of DBL in the form of a lack of data sources and their lack of accessibility, interoperability and low level of digitisation. This results in barriers to using DBL, including the static nature of DBL, as information often has to be updated manually, and the limited accessibility for various stakeholders [13]. Unstructured and heterogeneous documents are not characterised as data sources. Some initiatives, such as OpenDBL and DigiBuild, address heterogeneous data integration through ontologies, ETL processes, and microservice architectures [14], [15]. Hybrid approaches like the EUB SuperHub link distributed data sources via building UUIDs,

while frameworks like Demo-BLog enrich stored files with metadata to support future processing [4], [16]. Included CLEA system uses ETL to extract data from national repositories and a microservice architecture to integrate sensor data [17]. However, across all approaches, unstructured documents are generally stored but not analysed. The absence of methods for extracting meaningful, structured information from such data represents a critical research gap, which recent advances in LLMs could address [12].

4 Adapted DBL framework for Urban mining using LLM

The analysed research shows different approaches to interoperate different heterogeneous data sources to create a database for DBL using different digital technologies. Based on [6], [18], the system model in figure 1 summarises the main findings in a structure of using different data sources for DBL with the help of digital methods. In addition to the inputs of the DBL and their preprocessing to ensure their use in the DBL, a processing and output sequence is added, which represents the use of DBL data for various use cases.

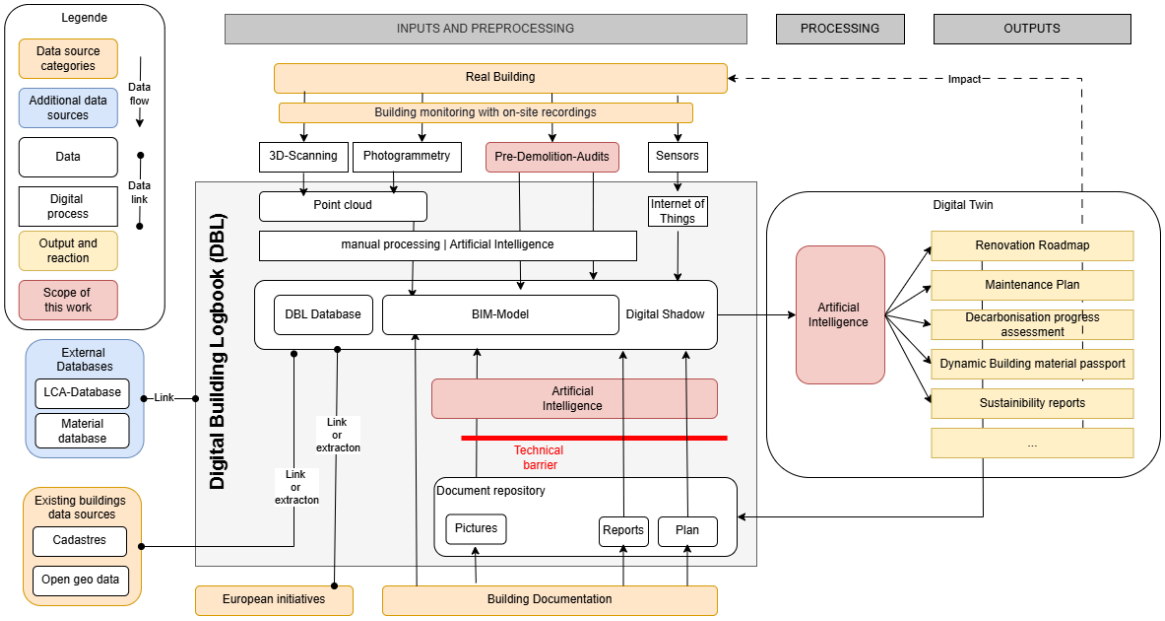

Figure 1: Scope of work and problem definition in context of DBL

"Inputs and preprocessing" is based on four categories of data sources: existing buildings, European initiatives, building documentation and building monitoring. Existing databases, such as cadastres, land registries or open geo data, can be included in DBL as an initial data set by extracting them, for example, using ETL, or by making them interoperable via links and ontologies [6], [15]. In addition to existing databases, new initiatives currently being developed, such as the Digital Product Passport, can be made accessible through extraction and linking. Another data source is building monitoring with on-site recordings. The most established approaches are those for the geometric recording of buildings using 3D scanning and photogrammetry, which can be further processed manually or by AI to generate a BIM model. Furthermore, the use of sensors and Internet of Things (IoT) technology is widespread during the use phase of buildings to capture various time series data for supervising and

controlling a building by coupling it with a BIM model to create a digital shadow of the building. For Urban Mining and the circular use of materials and components pre-demolitions audits (PDA) are of crucial importance to create an information basis.

While building documentation is often captured mainly in reports and plans, supplemented by images, and remains within a document repository due to the absence of technical methods, BIM models can be used directly as a data source in DBL. The lack of integration is a major issue and represents a crucial research gap. The utilisation of the various data sources and their pre-processing result in the creation of a digital shadow of the building, which provides data in DBL database, links, BIM model and other forms. The ultimate goal is to use the data obtained as an information basis for various questions in a processing step, which often generates outputs in the form of reports. At building level, for example, building material passports and renovations roadmaps play a decisive role in assessing sustainability and characterising measures to improve it. In addition to integrating the outputs into the document repository of the DBL, the impact in the form of the measures to be implemented on the real building plays a key role.

The research question explores how the use of LLMs can overcome this technical barrier and render this unstructured information usable in the context of DBL and further using AI in processing step.

4.1 Adapted RAG dataflow structure

The proposed approach involves adapting the technical architecture model for RAG to create a dynamic DBL for semantic indexing and utilisation of unstructured documents. The aim is to extract specific DBL information from different unstructured documents and transfer it into an overview. Further information is queried from the user via prompts and answered according to the RAG principle, with reference to the relevant sources.

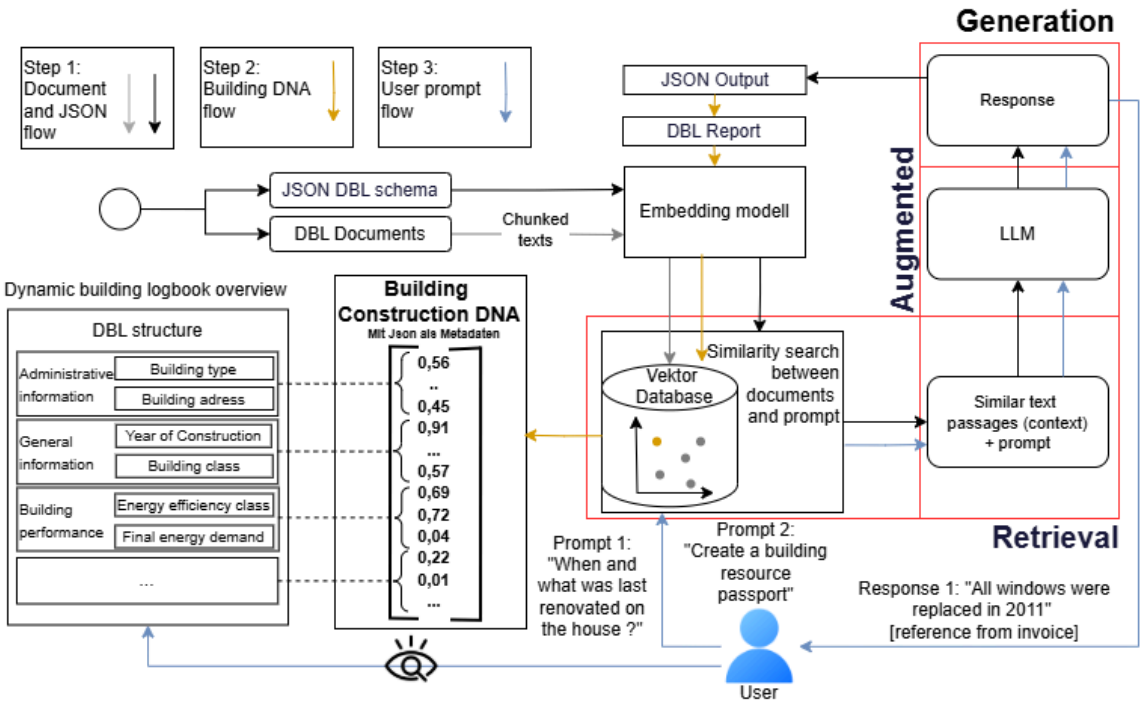

Figure 2: Adapted RAG dataflow structure

The starting point is the unstructured documents within the document repository and a JSON Schema, which defines the expected data to be included in the DBL and extracted from the documents. The workflow begins with document pre-processing. All building-related documents—reports, plans, certificates—are split into semantically coherent text chunks. Each chunk is converted into a high-dimensional vector using an embedding model. Alongside the vector, metadata such as the source document, section position, and chunk length is stored. These embeddings are kept in a vector database optimised for similarity search.
In the retrieval process of RAG vector database is searched for semantically closest chunks to the JSON schema .The relevant excerpts from the external knowledge base in form of DBL documents are used in the Augmented step for a context-enriched prompt that serves as input for the LLM, which generates a response based on it. This response is then transferred to the JSON output format, which structures the information according to the JSON DBL schema.
From this structured output the second step then begins, aiming to create a vector-based building construction DNA. Technically, this DNA is a second-level embedding that encodes the complete set of structured building attributes into a single high-dimensional vector. This representation functions as a compressed digital fingerprint of the building. First, the JSON output is converted into a text-based report by an LLM and then converted into a vector representation by the embedding model and stored in a vector database. Referring to human DNA, the construction building DNA represents the encoded blueprint for the building and contains all the information on how the building functions. This is then used to generate a dynamic building logbook overview in the form of a structured report, which bundles all the information from the DBL previously defined in the JSON schema. The third step deals with user interaction. Users can view the dynamic building logbook overview to gain quick insight into the building's key figures. Further information can be entered via a chatbot function in the form of prompts, based on the RAG principle and the document vector representations.
In practice, the Building Construction DNA serves as an indexable and comparable representation of a building's technical state within the DBL system. By storing both chunk-level embeddings (fine-grained retrieval) and DNA-level embeddings (holistic building profile), the framework supports both targeted queries and high-level building comparisons.

5 Discussion

The presented approach addresses the limited number of DBL data sources by making unstructured documents accessible as data sources. Despite the low level of digitisation of the data sources, the approach achieves a high level of digitisation. This is achieved by universally storing the data in a vector database and accessing it via a chatbot and LLMs. The continuous integration of new documents results in a dynamic DBL, thus releasing a static structure and the need for manual updates of information. The approach is limited by the use of non-open data, so the voluntary participation of those involved is needed. Technical issues also need to be clarified, such as how to handle contradictory information in documents on the same topic, and how to manage document versions. Further research is needed to establish the interoperability of this approach with other data sources. The approach has further potential with other unstructured data sources, such as plans and images,

as well as structured data sources, such as databases. This approach offers two significant potential benefits for urban mining. Firstly, the inclusion of PDA enables the integration of specialised data, such as the higher-value utilisation paths of components, in an unstructured form. Secondly, the use of this and all other data to generate material and component inventories in the dynamic DBL provides a decisive basis for urban mining. In addition, the vector database created can also serve as a basis for building material cadastres on city level.

6 Conclusion

The presented concept demonstrates how LLMs can overcome the current limitations of DBL, such as the lack of data sources and the inability to handle unstructured data. By embedding document content and integrating similarity-based retrieval with natural language generation, the system can transform static document repositories into dynamic knowledge bases. However, other challenges remain, such as data privacy and validation, particularly when handling contradictory information. Additionally, the approach relies on the voluntary participation of building owners who possess the relevant documents. However, this approach enables the digitisation of DBLs, particularly for existing buildings, and ensures that data relevant to urban mining and the circular economy can be automatically extracted, structured and utilised.

References

[1] G. A. f. B. United Nations Environment Programme and Construction, *Not just another brick in the wall: The solutions exist - Scaling them will build on progress and cut emissions fast. Global Status Report for Buildings and Construction 2024/2025*, Mar. 2025.

[2] M. de Wit, J. Hoogzaad, S. Rumjumar, H. Friedl, and A. Douma, “The Circularity Gap Report”, Circle Economy, Amsterdam, 2019.

[3] E. Commission, E. A. for Small, M.-s. Enterprises, *et al.*, *Definition of the Digital Building Logbook – Report 1 of the Study on the Development of a European Union Framework for Buildings' Digital Logbook*. Publications Office of the European Union, 2020. DOI: doi/10.2826/480977.

[4] M. Malinovec Puček, A. Khoja, E. Bazzan, and P. Gyuris, “A Data Structure for Digital Building Logbooks: Achieving Energy Efficiency, Sustainability, and Smartness in Buildings across the EU”, *Buildings*, vol. 13, no. 4, p. 1082, Apr. 20, 2023, ISSN: 2075-5309. DOI: 10.3390/buildings13041082.

[5] J. D. S. Gonçalves, W. C. Lam, and M. Ritzen, “The Role of Digital Building Logbooks for a Circular Built Environment”, in *A Circular Built Environment in the Digital Age*, C. De Wolf, S. Çetin, and N. M. P. Bocken, editors, Cham: Springer International Publishing, 2024, pp. 229–243, ISBN: 978-3-031-39674-8 978-3-031-39675-5. DOI: 10.1007/978-3-031-39675-5_13.

[6] M. Gómez-Gil, M. M. Sesana, G. Salvalai, A. Espinosa-Fernández, and B. López-Mesa, “The Digital Building Logbook as a gateway linked to existing national data sources: The cases of Spain and Italy”, *Journal of Building Engineering*, vol. 63, p. 105461, Jan. 2023, ISSN: 23527102. DOI: 10.1016/j.jobe.2022.105461.

[7] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, “Attention Is All You Need”, in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017. DOI: 10.48550/ARXIV.1706.03762.

[8] J. Kaplan, S. McCandlish, T. Henighan, *et al.* “Scaling Laws for Neural Language Models”. arXiv: 2001.08361 [cs]. (Jan. 23, 2020), [Online]. Available: http://arxiv.org/abs/2001.08361 (visited on 04/16/2025), pre-published.

[9] Z. Ji, N. Lee, R. Frieske, *et al.*, “Survey of Hallucination in Natural Language Generation”, version 7, *ACM Computing Surveys*, vol. Volume 55, 2022. DOI: 10.1145/3571730.

[10] P. Lewis, E. Perez, A. Piktus, *et al.* “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks”. version 4. (2020), [Online]. Available: https://arxiv.org/abs/2005.11401 (visited on 04/16/2025), pre-published.

[11] M. Gómez-Gil, A. Espinosa-Fernández, and B. López-Mesa, “Review and Analysis of Models for a European Digital Building Logbook”, *Energies*, vol. 15, no. 6, p. 1994, Mar. 9, 2022, ISSN: 1996-1073. DOI: 10.3390/en15061994.

[12] R. Alonso, R. Olivadese, A. Ibba, and D. Reforgiato Recupero, “Towards the definition of a European Digital Building Logbook: A survey”, *Heliyon*, vol. 9, no. 9, e19285, Sep. 2023, ISSN: 24058440. DOI: 10.1016/j.heliyon.2023.e19285.

[13] Executive Agency for Small and Medium sized Enterprises., *Building Logbook State of Play: Report 2 of the Study on the Development of a European Union Framework for Buildings’ Digital Logbook.* LU: Publications Office, 2020. [Online]. Available: https://data.europa.eu/doi/10.2826/519144 (visited on 06/16/2025).

[14] “openDBL Project. D1.3 System Specifications and AI-based mapping concept”. (2023), [Online]. Available: https://ec.europa.eu/research/participants/documents/downloadPublic?documentIds=080166e501f4f81d&appId=PPGMS.

[15] J. L. Hernández, D. Arévalo, S. Martín, *et al.*, “Using Extraction, Transformation and Loading Procedures for Digitalisation of Buildings”, in *2024 9th International Conference on Smart and Sustainable Technologies (SpliTech)*, Bol and Split, Croatia: IEEE, Jun. 25, 2024, pp. 1–6, ISBN: 978-953-290-135-1. DOI: 10.23919/SpliTech61897.2024.10612635.

[16] A. Ibba, S. A. Hwang, D. R. Recupero, R. Alonso, and R. Olivadese, “Defining Valuable Data and Stakeholder Engagement in Digital Building Logbooks: A Framework for Effective Data Storage”, presented at the 2024 European Conference on Computing in Construction, Jul. 14, 2024. DOI: 10.35490/EC3.2024.207.

[17] Redmond, Alan Martin, “The conceptual architecture requirements for french digital building logbook”, in *IARIA Congress 2024*, Porto, Portugal, 2024, pp. 81–87, ISBN: 978-1-68558-180-0.

[18] M. Gómez-Gil, A. Espinosa-Fernández, and B. López-Mesa, “Contribution of New Digital Technologies to the Digital Building Logbook”, *Buildings*, vol. 12, no. 12, p. 2129, Dec. 4, 2022, ISSN: 2075-5309. DOI: 10.3390/buildings12122129.