Calculation of embodied GHG emissions in early building design stages using BIM and NLP-based semantic model healing

Kasimir Forth[a], Jimmy Abualdenien[a], André Borrmann[a]

[a] *Chair of Computational Modeling and Simulation, Technical University of Munich, Germany*

Abstract

To reach the goals of limiting global warming, the embodied greenhouse gas (GHG) emissions of new buildings need to be quantified and optimized in the very early design stages, during which design decisions significantly influence the success of projects in achieving their performance goals. Semantically rich building information models (BIM) enable to perform an automated quantity take-off of the relevant elements for calculating a whole building life cycle assessment (LCA). However, imprecise type and property information often found in today's BIM practice hinders a seamless processing for downstream applications. At the same time, the early design stages are characterized by high uncertainty due to the lack of information and knowledge, making a holistic and consistent LCA for supporting design decisions and optimizing performance challenging. In assessing this often vague information, it is essential to consider different levels of element and material information for matching BIM to LCA data. For example, the structural properties of concrete are not yet defined in early design stages and should instead be considered as a range of material options due to different compressive strength classes.

This paper presents a novel methodology for automatically matching the coarse information available in BIM models of the early design stages to the respective entries in LCA databases as a basis for a fully automated calculation process of the embodied GHG emissions of new buildings. This approach solves the existing gap in the automation process of manually enriching BIM models and adding information of LCA data and missing layers of vague models. In more detail, the proposed method is based on Natural Language Processing (NLP), using different strategies to increase performance in matching elements and materials from a BIM model to a knowledge database to enrich environmental indicators of commonly used elements' materials. The knowledge database contains all missing information for LCAs and has different levels of information for a range of several potential design options of elements and materials, including their dependencies. Accordingly, this paper investigates multiple NLP techniques and evaluates the performance of state-of-the-art deep learning models such as GermaNet, SpaCy, or BERT. Following this, the most performant NLP approach is used to provide an automatic workflow for matching Industry Foundation Classes (IFC) elements to the knowledge database, facilitating a seamless LCA in the early stages of design. For five different case studies, the performances of the proposed matching method are analyzed. Finally, one case study is selected to compare the embodied emissions results to those of the conventional process.

Keywords: BIM, NLP, Model Healing, Early Design Stage, LCA

1. Introduction

According to the United Nations, the construction industry, specifically through the production of materials for building construction, is responsible for 11% of the global energy-related carbon emissions [1]. In order to reach the international goals of the Paris Agreement and reduce the environmental impacts, Green House Gas (GHG) emissions of new buildings must be significantly reduced. To assess the Global Warming Potential (GWP) of buildings, life cycle assessment (LCA) is an established method for calculating environmental indicators along the whole life cycle. At its core, it is based on environmental impact datasets for individual materials, typically provided through dedicated databases. During the design phase, a careful LCA of the different design options is required in order to identify the main drivers and optimize the building design accordingly. However, in conventional projects in today's practice, the main focus is still on improving the economic performance of buildings, while environmental qualities are usually not prioritized or even considered.

Until recently, LCA has mainly been calculated manually, which is time-consuming, especially when it comes to quantifying the building elements and matching them to environmental datasets, which have a different classification system and ontology. BIM combines geometry and semantics and thus facilitates deriving consistent and automated quantity take-off of the relevant elements for calculating whole building LCA. Using and enriching the semantic information of e.g., materials has great potential to completely automate the calculation of whole building LCA [2].

In early design stages, significant decisions are taken that have a major impact on the carbon footprint of the building to be realized. This is a primary reason for conducting a holistic multi-criteria variant analysis in the early design stages. At the same time, the early design stages are characterized by a high degree of uncertainty due to the lack of information and not-yet-taken decisions, making a holistic and consistent LCA for supporting design decisions and optimizing performance challenging [3]. In more detail, in the rough BIM models of early design stages, materials are typically defined by material groups rather than specific types, which allows a wide range of possibilities for each material group. Furthermore, several materials or element layers might not yet be defined, which gives the opportunity to explore and compare different design options. While several approaches for BIM-LCA integration exist, they are limited in implementing a fully automated workflow with open BIM models, in particular when it comes to early design phases [4]. A major challenge lies in the fact that imprecise type and property information in BIM models hinder a seamless processing for LCA applications.

To overcome this issue of vague model information in early design phases resulting in labor-intensive processes with additional manual input, we introduce the concept of "semantic healing" for automatically calculating embodied greenhouse gas (GHG) emissions. In doing so, we propose a novel automated method of matching LCA and BIM data on the element level by using Natural Language Processing (NLP). This gap of a fully automated matching process has not been filled yet [2], while research on NLP has recently advanced significantly and has strong potential for solving problems in the AEC industry [5].

This paper focuses on supporting decision-making in the early design phases. To support the decision-making in these phases, decisions for more detailed phases are also anticipated and analyzed. Based on the current approaches in the literature, the findings are considered to further extend the approach in the sense of a holistic analysis that is adaptable for further sustainability criteria.

The main contribution of this paper to the previously described problem involves a novel approach for semantically healing conceptual BIM models to assist the calculation of a holistic LCA, informing design decisions to detail the design further. The model healing process is conducted by enriching all necessary information to the model by automatically matching elements from BIM models to a knowledge database (discussed in detail in section 4) using Natural Language Processing (NLP).

In summary, this paper aims to answer the following research question: *Is automated semantic healing of BIM models possible in a way that allows assigning correct element types and materials to the respective model elements such that a reliable LCA can be calculated?*

It is structured as follows: Section 2 provides the relevant background in the field of BIM, classification systems, NLP, and its application with BIM. Afterwards, Section 3 focuses on the state of the art of BIM-based LCA and discusses existing literature reviews, highlighting their limitations. Section 4 presents the methodology for enriching BIM models for LCA and proposes a new methodology for the semantic model healing process. The proposed methodology is then evaluated in Section 5 through different real-world case studies, where the potential, as well as limitations, are highlighted. Finally, Section 6 presents our conclusions and recommendations for future research.

2. Background

This Section describes multiple fundamental topics about BIM, level of development, classification systems, and Natural Language Processing (NLP), providing the necessary background for the following Sections.

2.1. Level of Development (LOD) and Building Development Level (BDL)

As building design is a progressive process in which initially vague information is further detailed, also BIM models gain more accuracy and reliability along the modeling process. level of development (LOD) represents the degree of completion, maturity, or elaboration [6]. While the BIMforum, the US chapter of buildingSMART International, has defined individual LOD [7], they have not been adopted as an international standard, yet. Defined in the European standardization effort EN 17412, level of information needs (LOIN) describes similar content like LOD, such as geometric and alphanumerical information [6], but specifies a particular use-case and milestone it is supposed to be applied for.

In Germany, LOD is known as the aggregation of level of geometry (LOG), specifying the geometric detailing, and level of information (LOI), representing the extent of alphanumerical information. Borrmann et al. discuss that

LOI is highly dependant on the project and client, so they can not be generalized. In BIM practice, LOI is often described with "Type-and-attribute tables" (TAT) specifying object types and attributes [8]. Additionally, buildingSMART International proposed the Information Delivery Specifications (IDS). "The main goal of IDS is to provide a simple yet comprehensive way to author and validate nongeometrical [Information Requirements]", for example specifying material or classifications [9].

Abualdenien and Borrmann developed a meta-model approach where multi-LOD data represent buildings at different design phases [10]. It is based on the BIMForum's LOD definitions and introduces a new concept, building development level (BDL). While LOD defines specific components, the BDL concept defines the maturity of the overall building with multiple LODs for each component.

LOIs and LOGs are of great importance for BIM-based LCA as they provide a means to specify the required information, or in turn, allow to take into account the vagueness and uncertainty of information provided in early design phases. Since less information is available in early design phases, generic datasets are used and missing material layers have to be assumed. During construction, on the other hand, product-specific data sets can be included in the calculation depending on the components used.

2.2. Open BIM and open formats

The design and construction of a building is a collaborative process that incorporates multiple disciplines. Each expert, such as the architect and structural engineer, uses different authoring tools and requires specific information to be present in the model to support a particular type of simulation and analysis. With the increasing specialization of the stakeholders, the building industry requires a high level of interoperability. The US National Institute of Standards and Technology (NIST) [11], as well as many researchers and case studies [12, 13, 14] have confirmed the difficulties and high annual costs resulting from the lack of interoperability between the AEC industry software systems.

The Industry Foundation Classes (IFC) schema [15] is an open data exchange format developed and maintained by buildingSMART with the goal of enabling interoperability across the AEC industry. It provides a common data model for lossless geometric as well as semantic data exchange. IFC is a free vendor-neutral standard and includes a large set of building information representations, including a variety of different geometry representations and a large set of semantic objects modeled in a strictly object-oriented manner.

Since 2009, the exchange format Green Building XML (gbXML) has been established as a public, non-profit schema [16] focusing on exchanging building information for operational energy simulations. Initially developed by Green Building Studio and later acquired by Autodesk, it is currently not maintained by an official standardization body. The extension markup language (XML) schema does not intend to describe a complete BIM model but represents the relevant building's environmental and geometric information. Often, the reduced BIM model is referred to as building energy model (BEM). The schema provides a container denoted as "campus" for one or several buildings, each of which has a closed building envelope described by surfaces. The surfaces have a type specification (e.g., "InteriorWall"), B-Rep geometry, references to adjacent spaces, which are referenced to zones, and assigned openings.

The gbXML format is used for LCA in early design stages, e.g., using CAALA software, considering both embodied and operational emissions. Nevertheless, the details about specific element layers and materials are not represented and therefore, not suitable for accurately matching environmental datasets on material level.

2.3. Classification systems

The classification of elements in BIM models enables the project-wide, uniform structuring of information in order to be read and used in an uniform and automatic manner. Applying "a classification system for component types in a digital building information model" enables all stakeholders "to have a common understanding of the information contained in the building model and, in conjunction with a system for model development, enables the realization of a high degree of automation for the processes to be operated by them" [17].

In the international context, the classification systems Omniclass and Uniclass are among the most widespread. In Germany, due to the lack of a full-scale classification system, the most common classification systems are DIN 276 for cost groups [18] and DIN 277 for room usage types[19]. According to German standards for calculating LCA, e.g., certification systems like DGNB or BNB, the classification system of the cost groups of DIN 276 is used [20, 21].

For the LCA context, DIN V 18599, focusing on the Energetic evaluation of buildings [1], has been recently established [22].

For LCA of buildings, a uniform classification of building elements defines the system boundary, especially for the manufacturing phase (A1-A3) as well as the end of life cycle (C3-C4) and module D. Thus, it is part of the "target and investigation framework" according to DIN EN ISO 14040 [23]. In German certification systems, according to Deutsches Gütesiegel Nachhaltiges Bauen (DGNB) and Bewertungssystem Nachhaltiges Bauen für Bundesgebäude (BNB), the classification of cost groups is carried out according to DIN 276 [18], taking into account the building elements for the cost groups KG 300 "Building - Structures" (see 8.3). The system boundary for the operational phase, in particular the energy consumption during operation (B6), on the other hand, refers to DIN 18960, which however is not relevant to this paper. For the classification of relevant areas, on the other hand, the net room area (NRF) according to DIN 277 is used [19].

2.4. Natural language processing (NLP)

Natural language processing allows computers to analyze and "understand" text created by human authors. At its core, natural text is transformed into a computer-readable representation through various techniques, including tokenization, lemmatization, and vectorization. Those techniques convert each word to its original/dictionary form and represent each word with a numerical value, describing the semantic similarity through their distance (e.g., the word *window* has a smaller distance to *door* than to a *tree*). Semantic similarity is a key feature of the matching process described in this paper.

As in other domains, artificial intelligence revolutionized its advancement. In this regard, long short-term memory (LSTM) and recurrent neural networks (RNN) dominated NLP as they learn bidirectional links between the vector representations of words and sentences to capture the overall meaning. Recently, those networks were outperformed by transformer-based models. One example of a pretrained deep bidirectional transformers is BERT by Google [24]. The structure of transformers consists of an encoder and a decoder, and transformer-based models themselves consist of multiple layers

[1] Full title: Energetic evaluation of buildings in the context of the energy consumption in the use phase (B6) relevant for the life cycle assessment in accordance with DIN EN 15643-2

of transformers [25]. This enables learning the contextual representations of input data.

2.5. NLP application in AEC

Locatelli et al. investigated in their scientometric analysis the synergies between NLP and BIM [5]. Beside the field of Automatic Compliance Checking, they also identified Information Retrieval from BIM models and Information Enrichment of BIM objects as a further fields of relevant application. Wang et al. developed a query-answering (QA) system for BIM information extraction (IE) by using NLP and achieved high accuracy scores in their evaluation [26]. Xie et al. introduced a method for matching real-world facilities to BIM using NLP for word segmentation and keyword extraction by adopting the LTP word segmentation module [27]. For the matching method itself, matching matrices based on HiTree paths are evaluated using the highest degree of matching with the natural language feature vector. Reitschmidt proposed an matching method of IFC materials to the LCA database Ökobaudat based on tokenization of material names and a distinct matching or via Levenshtein distance [28]. Nevertheless, automated matching of LCA and IFC data on the element level using NLP has not been developed yet [2]. Finally, Zahedi et al. proposed an NLP approach for documenting design decisions by searching building codes and request for proposal documents [29].

3. State of the Art of BIM-based life cycle assessment (LCA)

This Section focuses on a literature review of the current approaches of BIM-based LCA. First, existing literature reviews are compared. Based on this, a structured literature analysis is conducted by analyzing each publication according to several topics. Finally, the findings and limits of conventional and current BIM-based LCA methodologies are shown.

3.1. Existing literature reviews

Before presenting the literature analysis, existing ones are analyzed to prevent repetition. The focus is primarily on embodied emissions and energy rather than operational emissions or energy. Nevertheless, the aspect of multi-criteria approaches will be investigated too, for example, a combination of embodied and operational energy with life-cycle costs (LCC). Analyzing eleven publications from 2013 to 2015, the literature review of the BIM-based LCA method by Soust-Verdaguer et al. differentiates between Data input (BIM-LOD, LCA goal & scope, stages, and inventory), Data analysis (BIM software, Energy Consumption Calculation, LCA tool) and Outputs and communication of results (Environmental impact indicators, sensitivity analysis, embodied and operational CO2 emissions) [30].

In 2019, Wastiels and Decuypere classified existing approaches and identified five different strategies for BIM-LCA integration [31]. Later literature reviews base their findings on these five strategies, which contain Bill of Quantities (BOW) export, IFC import of surfaces, BIM viewer for linking LCA profiles, LCA plugin for BIM software, and LCA-enriched BIM objects.

Potrč Obrecht et al. classified in their literature review all analyzed methods according to the five strategies by Wastiels and Decuypere [32]. In the second step, they differentiated between manual, semi-automated, and automated approaches. In 2020, several other literature reviews were published focusing on different aspects. Roberts et al. identified in their literature review about LCA in building design process three different trends: integration of LCA into BIM, combining LCA and LCC, and using parametric approaches [33].

Cavalliere et al. concentrate on the capabilities of the combination of BIM and parametric-based tools, analyzing 25 different publications between 2013 and 2018 [34]. Most of the analyzed methods focused on BIM and only a few had a parametric approach included. Hollberg and Ruth were the first ones to develop a parametric-based LCA (PLCA) in 2016, using visual programming language (VPL) but no BIM integration [35]. Llatas et al. focus in their systematic literature review on life cycle sustainability assessment (LCSA) and add, besides LCA and LCC, also sLCA in their investigation approach [36]. In total, they reviewed 36 papers about BIM-LCSA integration, but only six methods included LCA and LCC and none sLCA. Tam et al. analyzed in their critical review on BIM and LCA 61 articles by using content analysis method [37]. Furthermore, they identified several unaddressed issues, for example, the lack of a standardized structure between BIM and LCA data.

3.2. Literature analysis

Based on the findings of existing literature reviews in the field of BIM-based LCA, a systematic literature analysis was conducted. After reviewing more than 60 publications in this field, published in 2018-2022, 25 were selected and analyzed. In the following, the main findings are described. The main focus of several approaches is on detailed design stages such as [38, 39, 40]. However, optimization of the building design can be achieved in early design stages, when information is still uncertain. Therefore, Rezaei et al. are suggesting a workflow that is based on Autodesk Revit but doesn't include an optimization process [41]. Only a few methodologies implemented uncertainties in their approach [42, 39, 41].

As previously shown, Wastiels and Decuypere classified five different integration strategies. The two mainly implemented approaches of the analyzed publications are the one which uses authoring tools for getting the bill of quantities (BoQ), which was analyzed by [32]. The second primary strategy is using BIM objects enriched with property set (Pset)s [39, 43, 44, 45]. A new approach by Lee et al. suggests BIM templates for authoring tools to avoid data loss due to exchange formats [46]. Only a few of the analyzed publications use existing LCA Plugins for Autodesk Revit, such as Tally, eToolLC, or One Click LCA [47, 48, 49, 50, 51]. As most of the approaches use the BIM model only for downstreaming LCA-related information, only one includes a computer-readable feedback communication process of the calculated results back to the BIM model [52].

Most of the analyzed approaches used the open BIM format, mainly IFC, such as [53, 52, 43, 38, 40]. Nevertheless, another open BIM exchange format specialized in energy simulation is Green Building Extensible Markup Language (gbXML), which was used by [51]. Other approaches use the closed

BIM approach with software tools like Autodesk Revit [54, 53, 55] or additionally in combination with the VPL tool Autodesk Dynamo [56, 57, 58, 59, 60]. Another used VPL tool is McNeel's Rhino and Grasshopper, which was used by [61, 62, 63], which is not considered as a BIM tool just as little as Trimble's Sketchup, used by [64].

Although this publication focuses on LCA, the framework allows it to be extended to multiple criteria for design optimization. Only a few analyzed approaches show a few more criteria, which can be included in their workflows. While Kiamili et al. focus only on embodied energy of heating, ventilation, air conditioning (HVAC) systems [58], other approaches include both the embodied emissions of building construction and HVAC [42, 40]. In a next step, further publications even include operational energy besides embodied energy [65, 53, 51]. Besides LCA, Life Cycle Costs (LCC) and social Life Cycle Assessments (sLCA) are further relevant criteria to consider in the field of LCSA. A few approaches include both LCA and LCC Abu-Ghaida and Kamari, Eleftheriadis et al., Figl et al., Santos et al.. Llatas et al. propose the only approach, which considers all three criteria of LCSA, while the main focus of sLCA is on working hours [43]. Nevertheless, there is no methodology that integrates embodied emissions of building construction and HVAC, as well as operational emissions in early design phases.

As a functional unit of the approaches, most of the analyzed publications focus on the whole building. Global Warming Potential (GWP) was considered by all approaches, while other publications also considered further environmental impact categories such as acidification potential (AP), eutrophication potential (EP), ozone depletion potential (ODP), and photochemical creation potential (POCP) [47, 64, 66, 44, 45]. Depending on the country of the publication, several different international life cycle inventory (LCI) databases were used, such as German Ökobaudat, or ecoinvent and KBOB from Switzerland, and sometimes even product-specific Environmental Product Declaration (EPD)s. Palumbo et al. investigated the challenge of using EPDs in early design stages to obtain accurate LCA results [66].

As a result of the literature analysis, there is great potential for including LCA calculations in an optimization process in early design stages using open BIM models. Furthermore, most of the analyzed publications focused only on the criterium of LCA, extending the focus on multiple criteria such as LCC is also becoming more relevant. Nevertheless, the process of matching LCA and IFC data on element and material levels is still manual, and an automated approach is not developed or solved yet.

3.3. Limits of conventional BIM-based LCA calculation

As the findings of the literature review showed, there are still challenges and opportunities in the field of BIM-LCA integration. In this Section, the limits of conventional BIM-based LCA approaches will be critically investigated using a case study.

Safari and AzariJafari stress in their publication out that a major focus will be in early design stages, considering LODs and uncertainties in future approaches [2]. Zimmermann et al. showed in their investigation of industry practice and needs different challenges, such as manual workflows, matching model data with LCA data, quality in models, and many more [67].

Nevertheless, in conventional projects in practice, the main focus is still on the economic performance of buildings, while environmental qualities are not widely spread yet. This is the reason to approach the holistic multi-criteria variant analysis, in the early design stages, based on existing approaches of BIM-integration strategies for LCA. Current approaches still have limits of fully automated workflow with open BIM models [4]. Scherz et al. propose in their methodology of hierarchical reference-based know-why models design support for several sustainability criteria focusing on building envelopes [68]. Nevertheless, BIM integration is only envisioned in their future work.

The main scope of this paper focuses on the early design phases. To support the decision-making at these phases, detailing decisions from more detailed phases are additionally analyzed. Based on the current approaches in the literature analysis, the findings are considered to further extend the approach in the sense of a holistic analysis that is adaptable for further criteria, for example, LCC or similar.

4. Methodology for semantic model healing for early BIM models for LCA calculation

The aim of this paper is to develop a framework for calculating ranges of embodied emissions of building designs based on element-specific design variants to support decision-making in early design phases. The methodological approach includes open BIM data exchange in early design stages, environmental impacts of construction, operation, and End-of-Life phase of buildings), as well as an automated matching of relevant information from the model. Therefore a robust implementation should take different modeling approaches (model authors & software products) into consideration. Furthermore, the framework provides flexibility to add economic impact or individual cost benchmarks and the calculation of further criteria.

As shown in the previous Section 3.3, the BIM-integration of LCA lacks an approach for early design stages, which fully automatically matches all information from BIM models to LCA datasets and considers uncertainties and missing information in early design stages. Therefore, the proposed methodology focuses on the following key features:

- Semantic model healing by using an LCA knowledge database (LKdb)
- Automated matching of IFC elements to the elements of LKdb using pretrained NLP models
- Calculation of LCA result ranges according to the early design uncertainties

The details of the method are described in the following Sections. First, the general framework is introduced, followed by more detailed descriptions of each part, such as semantic model healing, LKdb, and the matching method.

4.1. Proposed methodology

To perform multi-criteria analyses using BIM, engineers need a set of information to be present in the BIM model. Usually, in early design stages, some of the required information is uncertain or even completely missing, which has a significant influence on analysis or simulation results. For this reason, the concept of a knowledge database is introduced, which provides all relevant information and default values in the case that relevant information is missing.

As this paper focuses on embodied GHG emissions, the database is filled with LCA-relevant information. Nevertheless, the database can be easily extended to cover other criteria as well. In case of missing or uncertain information, such as elements or properties, the LKdb provides a set of possible options or ranges of values. Furthermore, several design variants can be explored in these cases, and their performances can be evaluated according to the influence of the incorporated uncertainties on the environmental qualities.

In the proposed methodology, design decisions are made by selecting one of these variants. To implement the conducted selections in the design, these are communicated back to the BIM authoring software. The proposed methodology follows the open BIM approach to support a wide range of authoring tools. Therefore, it uses Industry Foundation Class (IFC) and BIM Collaboration Format (BCF) as exchange data formats.

4.1.1. General Framework

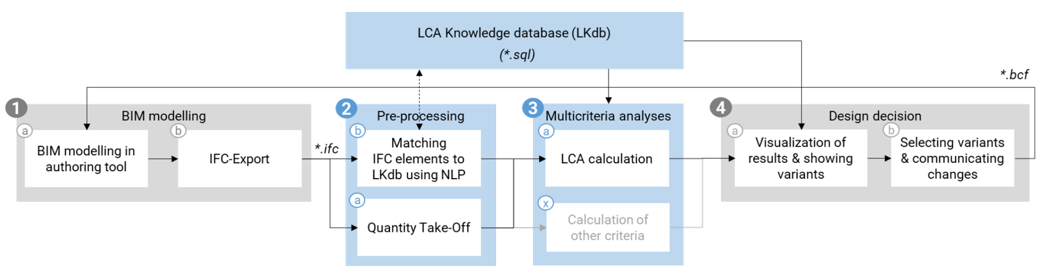


Figure 1: General Framework of NLP-based semantic model healing of early BIM models for LCA calculation

Figure 1 presents the different steps of the proposed methodology, which was briefly described above. In the first step (1), the BIM model is exported from the authoring tool as an IFC file. In the next step (2), the IFC data are pre-processed for the following analyses. This is split into the Quantity Take-Off (2.a) and the NLP-based matching method (2.b), which is explained in more detail in the upcoming Section 4.1.2. The Quantity Take-Off (QTo) contains information about the element type, the classification, the sum of all type-specific element areas, the area unit, the amount of type-specific elements, the element-specific materials, and the thicknesses of the material layers. In terms of the multi-criteria analyses to be performed in the next step, the focus is on the LCA calculation in this publication (3.a). The final step consists of visualizing the results (4.a), supporting the selection process, and communicating the design decisions and changes back to the BIM author (4.b). In this regard, this publication concentrates on the visualization of the LCA result ranges, including relevant benchmarks.

4.1.2. Semantic model healing

The semantic model healing process is performed to add all relevant but missing information for the model-based LCA calculation. The first step in this process is to collect all available and relevant information from the IFC model. Based on this information, the second step focuses on how existing techniques of NLP help to match IFC elements to those of a knowledge database. Different strategies are used for the NLP-based healing process to increase the performance of the matching element from an "imperfect" BIM model to this knowledge database. In the last step, all missing element information is added by those of the matched knowledge database. The knowledge database contains all missing information for LCA and has different levels of detail for a range of several potential design variants of elements and materials, including their dependencies. The semantic model healing process is performed when the incomplete IFC element data are matched to the most similar element in the LCA knowledge database (LKdb) and afterwards enriched by all missing element information provided by the LKdb.

4.2. LCA knowledge database (LKdb)

The LCA knowledge database, based on elements, layers, and materials, will contain all information that is relevant for the holistic calculation of different criteria and is typically not provided in the IFC model. This database is similar to the recently published "EarlyData knowledge base" by Schneider-Marin et al., which has a similar purpose of calculating reliable LCA results in early design stages considering uncertainties [71]. Nevertheless, the focus of their database was not focusing on using it for testing a robust matching approach. This LCA knowledge database is linked with different external databases, for example, databases for environmental criteria, such as Oekobaudat [72] (Figure 2). The main aim of the LKdb is to provide all necessary input information for a holistic and correct LCA calculation and analysis, which is typically missing in early design phases. Another aim is to combine several external databases with different input information on different levels of information. International databases can be

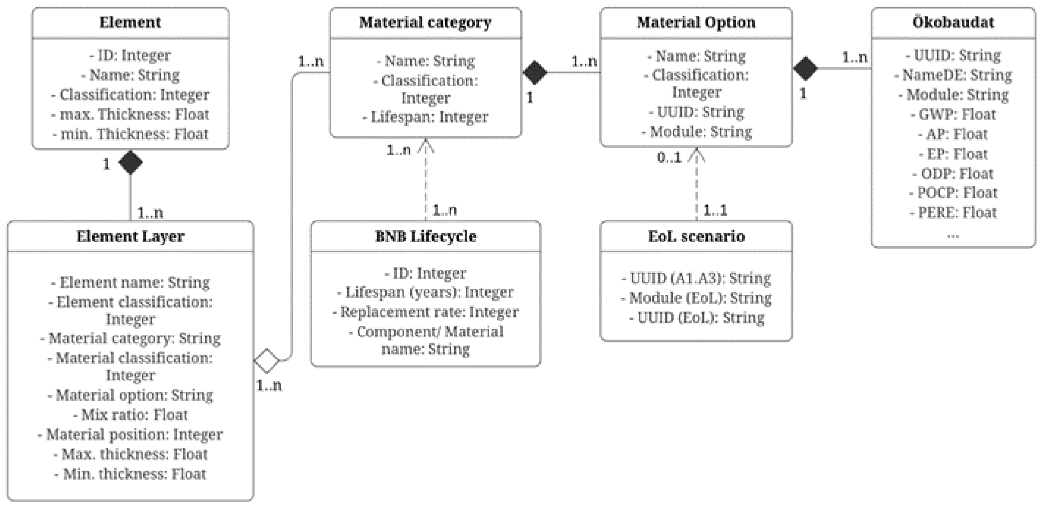


Figure 2: UML diagram of the proposed LCA Knowledge database with different hierarchies, such as elements, material categories and material options, and external databases such as BNB life cycle [69] or Ökobaudat [70]

added to the database using this methodology. Material and product-specific Environmental Product Declarations (EPD) can be linked, too.

The database provides additional information on different levels, which are needed for a sufficient LCA calculation, such as the lifespan of an element, End-of-Life scenarios if missing in the original external dataset, or densities. Due to the German LCA classification standard according to cost groups, the database itself is structured similarly to the classification system of DIN 276 on the third level but provides a material-specific level of different element layers. Other criteria information like cost values or U-values (if missing in the model) for calculating operational energy can be stored in the database as well but are out of scope in this publication. This ensures that a change in the variants leads to a change in all criteria calculations and shows the complex dependencies of the multi-criteria design decision process. A first extension, including LCC, was tested recently [73].

The general structure of the proposed LCA LKdb consists of three different levels: element, material category, and material option. As the LCI database, Ökobaudat was chosen [70]. The main reason for this decision is that the selected case studies are located in Germany. Thus the BIM models use German terminology for components and properties. Furthermore, Ökobaudat is the official LCI database for German certification systems and consists of more than 1400 datasets specifically for building products.

Every single dataset of Ökobaudat has as keys the Universally Unique Identifier (UUID) and the relevant life cycle modules (A1-A3, C3, C4, D). All datasets from Ökobaudat consist of several environmental impact categories, such as Global Warming Potential (GWP), Acidification Potential (AP), Eutrophication Potential (EP), Ozone Depletion Potential (ODP), Photochemical Creation Potential (POCP), Primary Energy Renewable (PERE) and many more. Nevertheless, the quality of some datasets is not sufficient for a holistic LCA, as there are some End-of-Life scenarios missing. Therefore, generic End-of-Life (EoL) datasets from Ökobaudat have to be manually matched to those datasets, which are lacking this information. For this reason, up to two UUIDs are linked to the material option dataset of the LKdb: one for the production phases and, if necessary, one for the End-of-Life scenario. Stenzel conducted in her master thesis this manual mapping as well as a classification of all UUID according to German cost groups using DIN 276 [74]. This information is used for the prototypical implementation of the LKdb. All material options have a name and classification as their keys, which is derived from the German name in Ökobaudat. Further entries are the classification, UUID, included Modules, and the encoded NLP vectors of the name (spans and tokens), which are stored because of calculation performance reasons.

According to the structure of Ökobaudat, every material is classified according to specific material categories. As there are three different levels of categories, only the most specific one is used for material classification in the LKdb. Every material category is mapped to potential cost groups of the German classification system [18]. This is necessary to map the service life of building components on this level, according to [69]. This external input is named "BNB life cycle" and contains an ID, the lifespan in years, the replacement rate according to 50-year buildings life, and an element or material name according to its own classification. The key for material categories is the name and the classification. Additional information is the encoded NLP vectors of the material category name (spans and tokens) due to calculation performance reasons.

For setting up element layers, material options and categories are used in the next level. Elements themselves can consist of one or multiple element layers. Both elements and element layers have a default maximum and minimum thickness. The material layer corresponds to the third level of the German cost group system [18]. As the material layer can consist of composite materials, different mix ratios need to be defined. For monolithic layers, the ratio is 100%. As an example of composite materials in one element layer, reinforced concrete consists of different materials, such as concrete and reinforcement steel. Every element layer has a unique material position, which describes the order of the material in the specific element. For the element levels, every entry gets a unique ID as a key. Due to calculation performance reasons and also for the elements, the encoded NLP vectors of the element names (spans & tokens) are stored in the LKdb.

All entries for elements are inserted due to common domain knowledge. The most typical construction types were considered and modeled using the proposed schema. Due to the versatility of constructions, the database is continuously updated and has no claim to be ever completed.

4.3. Matching method

In later design stages, conventional methods rely on manually matching each IFC material to a UUID of external databases and store this information as a Pset attribute in "PSetEnvironmentalImpactIndicators" and "PSetEnvironmentalImpactValues" according to [75] or self-defined Psets, such as "Plca_Lca" according to [40]. To avoid the laborious manual work of matching elements and materials of the BIM models to the related ones in the LKdb, an automated matching method is proposed in this paper. Another approach by Reitschmidt also follows automated matching on material level [28]. In contrast, in early design stages, information about the materials is missing or incomplete. For this reason, the proposed method is matching on an element-level, so this vague or missing information about material layers can be added using the LKdb.

The main challenge of this method is to automatically and correctly match IFC and LKdb elements and materials so that calculation and analysis results are also reliable. In early design stages, materials are often defined in a more general way and not as specific as in LCA databases, e.g., "concrete" rather than "concrete C20/25". Sometimes, for some elements, material information is completely missing, while in the element naming, some material information is included, for example "brick wall". Furthermore, the proposed methodology aims to be a robust approach, which also considers poor model quality due to multiple ways of modeling BIM models and exporting them as IFC files. As the structure and nomenclature of elements and materials in IFC and the used LCI database Ökobaudat differ, the goal is to find the semantically most similar pairs on material and element level.

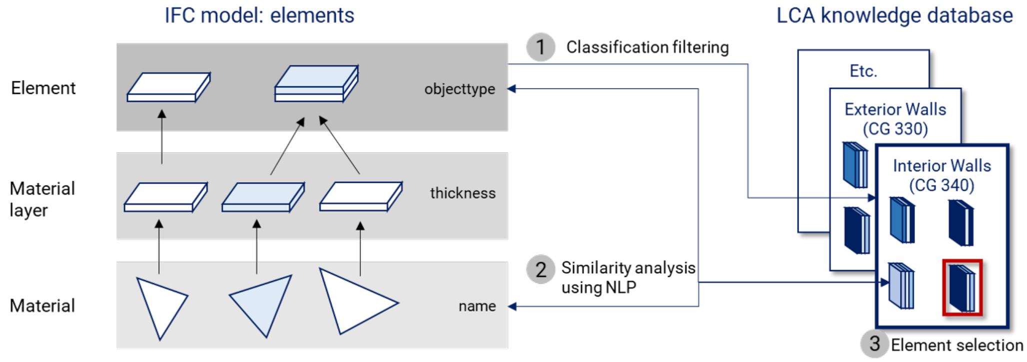


Figure 3: General steps for an automated method of matching elements of IFC models to those of the LCA knowledge database (LKdb)

Figure 3 shows the proposed matching method, which is divided into three steps:

1. Filtering of element classification
2. Similarity analysis using NLP
3. Element selection

First, IFC elements are filtered according to their classification type. This classification, according to the German cost group schema [18], is an exchange requirement (ER) and is stored as "IfcRelAssociatesClassification". If the element does not comply with the ER and no classification is available, the method can also classify the IFC element using its "IfcProduct" class types (e.g., IfcWall, IfcColumn, IfcSlab, etc.) and properties (e.g., IsExternal, IsLoadBearing, etc.) according to [76].

In the second step, every IFC element and its properties are analyzed and semantically compared according to its similarity with the filtered element variants in the LKdb. Not only the element expressions but also the material expression is analyzed according to the NLP technique used. In order to measure semantic similarity, every expression needs to be converted from text to a vector representation. In this case, a vector is a list of numerical values, and the combination of them represents the overall meaning [77]. Afterwards, the similarity between two different vectors A and B can be calculated using the cosine similarity, while n is the dimension of the vector:

$$cosine - similarity := cos(\theta) = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \tag{1}$$

In the following Sections, these three main steps of the matching method are explained in more detail, as shown in Figure 4. The choice of NLP technique will be investigated in Section 5.2.4.

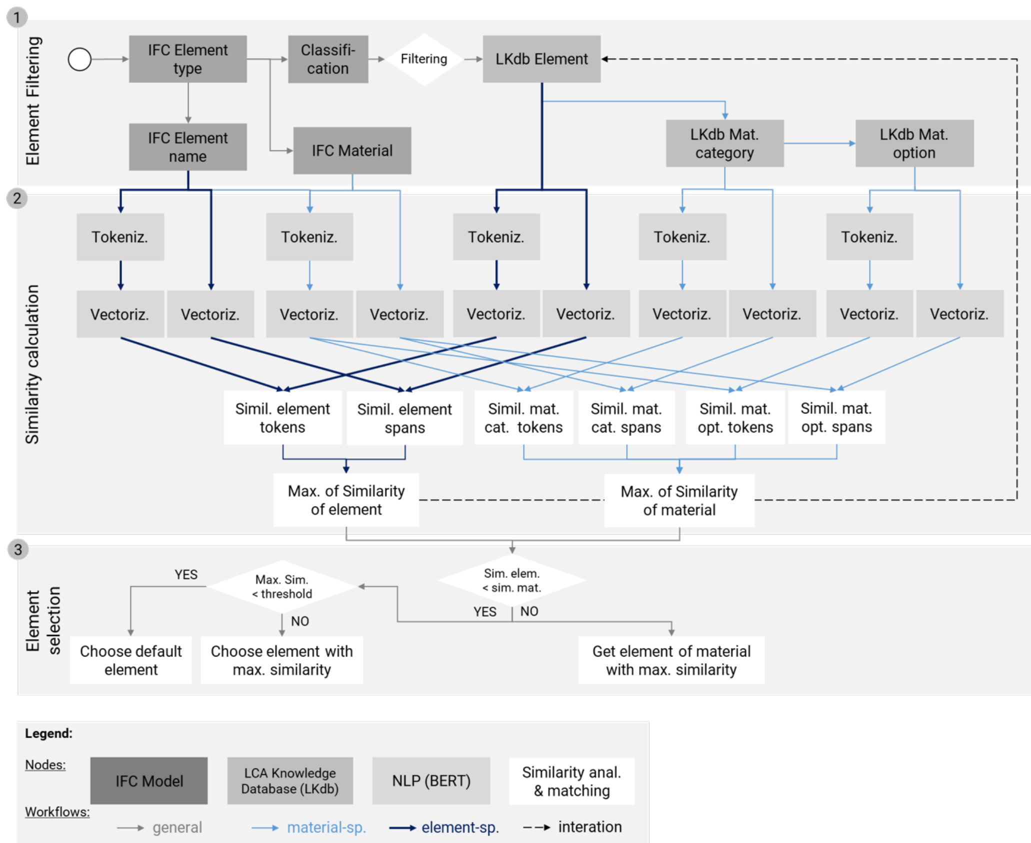


Figure 4: Detailed workflow for matching IFC elements to LKdb elements using Natural Language Processing (using BERT language model) and cosine similarity on different levels of information (element, material category, and material option)

4.3.1. Element filtering

The starting point is iterating through each element type from the IFC model. Each element type consists of an element name, its classification according to DIN 276, and its material name. Based on the classification, a list of LKdb elements is filtered to compare similarities with the IFC element. For performing a robust matching method, the elements are compared on material and on element levels. Therefore, the IFC element name is compared to the filtered list of LKdb elements. And furthermore, the IFC material is compared to the material categories and material options which are contained in the filtered element list. The differentiation between material category and material options is required due to the fact that the matching method considers different LOIs for the naming of materials (see Section 4.3.2).

4.3.2. Similarity calculation

In the calculation of semantic similarities, three couples of IFC and LKdb are considered: on element level, material level comparing with the material category, and comparing with the material option. Each of these three couples is split into calculating the whole span and all tokens. To this end, the word encoding or vectorization is conducted for twelve different words per every iteration step, while the tokens themselves are also iterated. For each token set, only the maximum token is considered in the following selection process. The calculation of the cosine similarity is conducted six times per iteration step and is stored in a list for the following selection process:

- element tokens
- element spans
- material category tokens
- material category spans
- material option tokens
- material option spans

After the calculation of all cosine similarities, the most similar element and material are identified. The maximum similarity of all element tokens and element spans are compared for the most similar element. Accordingly, the maximum similarity of all material category tokens and spans, as well as material option tokens and spans, are derived for the most similar material.

4.3.3. Element selection

In the next step, the final element selection is performed based on the previously derived most similar element and material. Therefore, the two cosine similarities of the most similar element and most similar material are compared. If the similarity of the material is higher, the corresponding element of the material is searched and selected. In case the similarity of the element outperforms the one of the material, this element is selected if its cosine similarity is higher than a threshold. As a threshold, 80% was set, according to the material similarity analyses using the BERT model in Section 5.2.3. If this threshold is not reached, the default element of the classification group is chosen, as the identified element similarity is too low to ensure the quality of this matching method. For IFC elements with multiple material layers, the steps of the previously explained workflow are derived for every material layer. Nevertheless, in the end, the different results have to identify only one selected element. For this, the different elements of each layer are counted, and their cosine similarities are summed up. Finally, the element with the highest summed-up cosine similarity is selected as the overall multi-layer matched element.

4.4. LCA calculation of LKdb elements

This paper focuses exclusively on embodied emissions. For this reason, for the LCA calculation, the operational part B6 is omitted. This study does not focus on different environmental impact potentials but on the reliability of the calculation process. The system boundaries of the LCA include the life cycle phases production (A1-A3), maintenance and replacement (B4), and End-of-Life (C3, C4, D).

Generally, the Environmental Impact Potential (EIP_{c_o}) of the construction phase (c) for each element (e) is the sum of the production phase (P_e), recovery and disposal phase (D_e), and the maintenance and replacement (M_e) in a reference period (t_D). As in the LKdb, different material options for one material layer exist. The element-specific environmental impact potential can consist of a range of results rather than a single value. In the following, the different steps are described for calculating the Environmental Impact Potential of one specific option set (o). The final LCA result ranges are derived by the different options and can be clustered on element or cost group level or determined for the whole building.

$$EIP_{c_o} = \sum_{e_o}^{n} \frac{P_{e_o} + D_{e_o} + M_{e_o}}{t_D} \tag{2}$$

The maintenance and replacement M_{e_o} of each element are calculated by the frequency of replacement ($n_{replacement,e}$) and the sum of the production P_{e_o} and recovery and disposal phase D_{e_o}, while the frequency of replacement depends on the ratio of the reference period t_D and the service life of the element (t_e).

$$M_{e_o} = n_{replacement,e} * (P_{e_o} + D_{e_o}) \tag{3}$$

$$n_{replacement,e} = roundup(\frac{t_D}{t_e}) - 1 \tag{4}$$

The production P_e of each element is the sum of the product of each layer-specific dataset for the production phase ($EIP_{e_{o,i}}^{A1-A3}$) and element-specific quantities ($f_{e_{o,i},x}$) over each element layer (i) of the element-specific maximum amount of layers (m_{e_o}). The recovery and disposal D_{e_o} is, accordingly, just taking the datasets for different life cycle phases into account (C3-C4, D).

$$P_{e_o} = \sum_{i=0}^{m_{e_o}} EIP_{e_{o,i}}^{A1-A3} * f_{e_i,x} \tag{5}$$

$$D_{e_o} = \sum_{i=0}^{m_{e_o}} EIP_{e_{o,i}}^{C3-C4,D} * f_{e_i,x} \tag{6}$$

The datasets $EIP_{e_i}^{A1-A3}$ or $EIP_{e_i}^{C3-C4,D}$ are stored in the LKdb. Depending on the functional unit (x), the quantity of each dataset can either be area a_e, length l_e, volume depending on the layer-specific thickness $d_{e_{o,i}}$, mass depending on the material-specific density $\rho_{o,i}$, or amount s_e.

$$f_{e_{o,i},a} = a_e \tag{7}$$

$$f_{e_{o,i},l} = l_e \tag{8}$$

$$f_{k_{o,i},v} = a_e * d_{e_{o,i}} \tag{9}$$

$$f_{e_{o,i},m} = a_e * d_{e_{o,i}} * \rho_{o,i} \tag{10}$$

$$f_{e_{o,i},s} = s_e \tag{11}$$

Depending on the level of the matching and available attributes of the IFC elements, different quantities can be used for this calculation step. The total area, length, and amount of all IFC elements of one specific object type are always derived by the Quantity Takeoff. If no material information is available in the IFC element and the matching is performed on the element level, the default quantities, such as thicknesses and densities, from the LKdb are used. If the matched element is based on most similar materials, the material layer information of the IFC element is used for the LCA calculation. This is also valid if, for a multi-layer element, only a few materials were identified in the matched element. For these matched materials, the material layer thicknesses of the IFC element are used, while for the missing ones, the default values are used according to LKdb. This selection ensures that all available and relevant information of the IFC model is used for LCA calculation. The LKdb provides all geometric and semantic information of the material layers, which are not modeled in the IFC model but are crucial for a holistic LCA calculation.

Figure 5: Selected case studies for validating the proposed matching method (Picture of case study 1: [78], case study 3 [79]

5. Evaluation and results

In this Section, we first briefly introduce five case studies, which are used to evaluate the proposed methodology. In the first evaluation, the best-performing language model is identified by testing three different models (GermaNet, spaCy, BERT) using the manually matched couples (IFC-LKdb) of case study 1. In the following Subsection, the whole element matching workflow is evaluated on all five case studies. Case study 2 is used for evaluating the whole procedure, including the LCA calculation using Global Warming Potential (GWP) as environmental impact category. Finally, we discuss the limitations of the proposed methodology based on the evaluations.

5.1. Case studies

To validate the proposed matching method, five case studies from real-world projects were selected, as shown in Figure 5 and Table 1. They are all office buildings, so the performance of the proposed approach is comparable but from different modelers and designers. Nevertheless, the quality of material and element naming, as well as the modeled BDL and classification, differ in all five case studies and need to be taken into account in the following analysis.

In Figure 6, the element distributions of the 2nd and 3rd levels of the German classification system according to DIN 276 are shown. Case studies

Case study number	Net floor area (sqm)	Total amount of elements	Total surface area of all elements (sqm)
1	ca. 11.870	2.110	68.949,39
2	ca. 1.950	307	5.823,82
3	ca. 35.300	13.966	85.193,77
4	ca. 11.390	7.144	118.155,97
5	ca. 8.710	5.822	117.562,25

Table 1: Information about the five case studies considering net floor area, total amount of elements, and total surface area of all elements

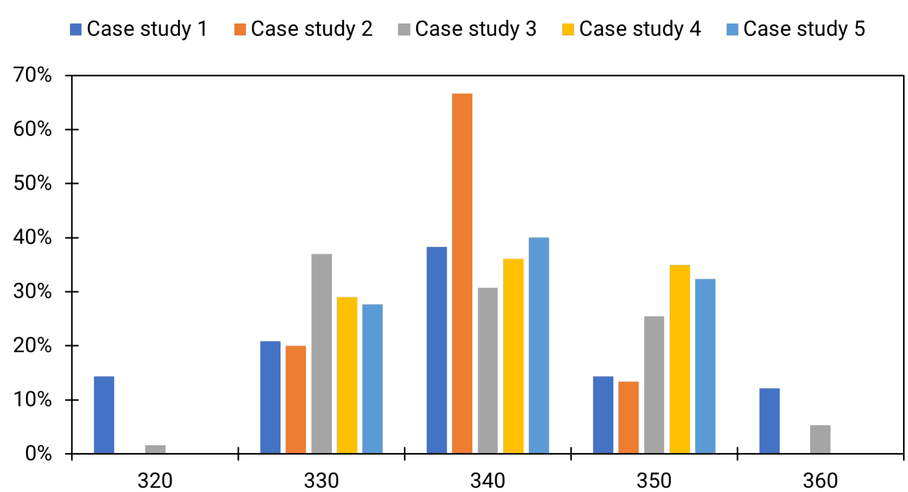


Figure 6: Overview of elements' classification distribution of the five case studies

2, 4 and 5 do not have elements in classes 320 (foundations) and 360 (roofs).

5.2. Evaluation of different NLP techniques for material matching

Following this, this publication investigates multiple NLP techniques and evaluates the performance of state-of-the-art deep learning models such as GermaNet, SpaCy, or BERT. They will be introduced in the following Sections and are the basis for the previously introduced matching method. The best-performing NLP technique is applied for the prototypical implementation and validation.

For comparing the three different NLP techniques and the performance of their workflows as well as calculating the whole building LCA, case study 1 was chosen, which was presented in Section 5.1. This real-life project guarantees that the material naming is not optimized but according to current industry standards so that the matching performances are tested under realistic conditions. In total, the IFC model of case study 1 consists of 2110 individual elements, which are summed up to 133 unique elements from the same families. Those consists of 59 unique IFC materials, which were manually matched to LCA material options and categories.

5.2.1. GermaNet

GermaNET is a Lexical-Semantic Net for the German language and is also known as the German version of the Princeton WordNet [80, 81]. GermaNet relates German nouns, verbs, and adjectives semantically by grouping lexical units that express the same concept into synsets and by defining semantic relations between these synsets (sets of synonyms). It can be represented as a graph whose nodes are synsets and its edges its semantic relations [82]. Therefore, the similarity is not measured using cosine similarity but graph-related shortest path similarity, which is equal to the inverse of the shortest path length between two synsets. There are other path-related similarity analyses, such as Wu-Palmer similarity [83] or Leacock-Chodorow similarity Leacock and Chodorow [84], which are not considered in this paper.

As the workflow of the GermaNet differs partially from the other two NLP techniques, the identification rate of the material token's synsets needs to be analyzed before analyzing the shortest path similarity. After the tokenization of the IFC material names, material options, and their related material categories of the LKdb, synsets are identified to calculate the shortest path similarity. Nevertheless, not for every token set, synsets could be identified. As shown in Figure 7, only for 20.3% of the material category tokens and

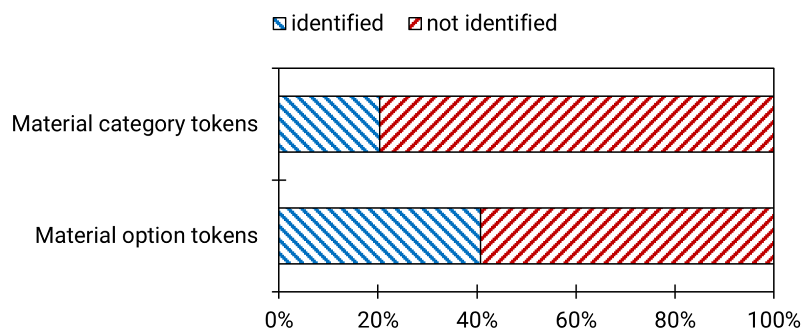


Figure 7: Identification rate of material token synsets using GermaNet for case study 1

40.7% of the material option tokens, a pair of synsets with the IFC material could be identified.

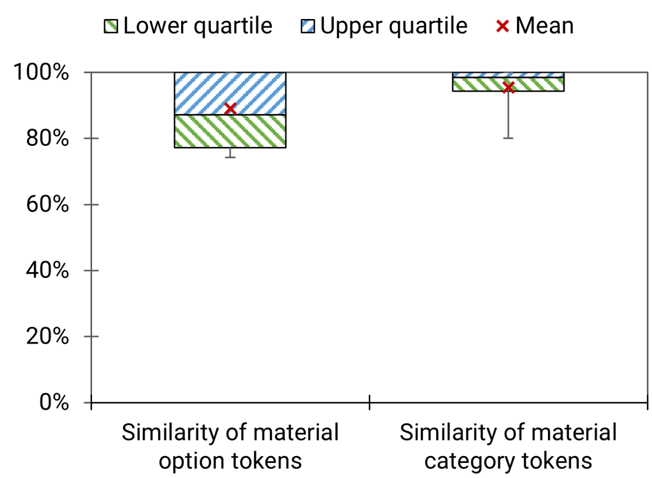


Figure 8: Shortest path similarity of identified, pre-matched material couples (IFC-LKdb) using GermaNet for case study 1

Nevertheless, the shortest path similarities of the identified pairs of synsets show promising results (Figure 8). The median of the similarity of material option tokens is 87.1%, and of the material category tokens, even 98.6%, both with little deviation. However, including the low synset identification rate of both material options and material categories from the LKdb, the total similarity are very low and not sufficient for being used in the proposed matching methodology.

5.2.2. spaCy

SpaCy is a pretrained neural network model and a promising implementation of the state of the art in the field of NLP [85]. Its large German model ("de_core_news_lg") includes 500k unique vectors in its corpus and represents every word or expression with a vector of 300 dimensions. As sources for training data, existing corpi were used, such as e.g., TiGer Corpus [86].

For the results of spaCy and BERT, the vectorization of both tokens and whole spans of the material options and material categories are compared.

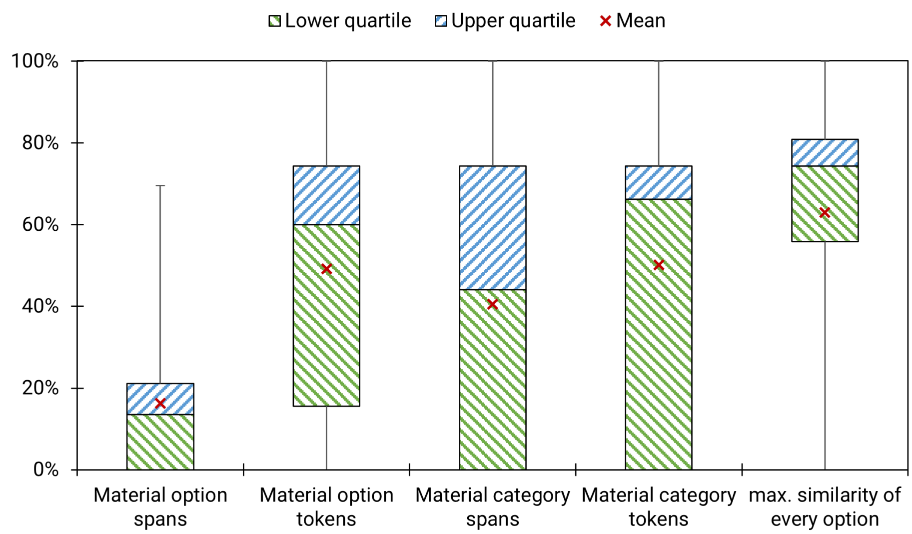


Figure 9: Cosine similarity of pre-matched material couples (IFC-LKdb) using spaCy for case study 1

As shown in Figure 9, the ranges of the cosine similarity of all different comparisons, according to Section 4.3.2, differ a lot. Generally, the similarities of IFC materials to the material option spans have the worst performance, with the median being 13.6%. The tokenization improves the performance of matching the material performances up to a median of 60.0%. Also, the spans of the material categories are much better (median at 44.4%). The tokenization of the material categories improves the performance results by up to 60.3%. As an additional performance result, the maximum similarity of all comparisons (material option spans and tokens, as well as material category spans and tokens) is calculated. Its median is 74.4%, but also the quartile ranges improved compared to all other ranges. In general, the results are not sufficient for further usage in the proposed framework but show a promising strategy for getting the maximum similarity of every option.

5.2.3. BERT

BERT stands for Bidirectional Encoder Representations from Transformers and was released by Google in 2018 [24]. Transformers-based pretrained models are currently state of the art and are capable of solving a wide range of tasks as they "can represent the characteristics of word usage such as syntax and how words are used in various contexts" [5]. BERT represents each word or expression with a vector of 768 dimensions, which is significantly higher compared to spaCy and makes the similarity calculation more time-consuming.

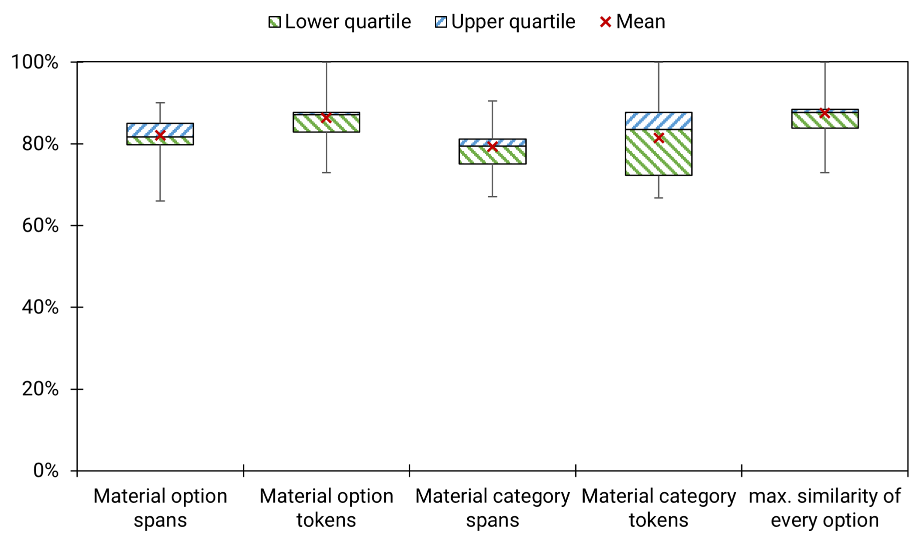


Figure 10: Cosine similarity of pre-matched material couples (IFC-LKdb) using BERT for case study 1

For the NLP technique BERT, the same similarity comparisons using cosine similarity are calculated as previously shown with spaCy. Figure 10 is showing the results as ranges of the material option spans and tokens and material category spans and tokens according to the workflow described in Section 4.3.2.

Generally, all result ranges differ much less compared to the results using spaCy. Additionally, all medians are between 79.2% (material category spans) and 87.2% (material option tokens). Also, the strategy of getting the maximum similarity of every option is improving the promising general results (median 87.7%). In addition, the minimum values of each result range show that BERT generally performs much better than spaCy.

5.2.4. Conclusions regarding NLP-based matching performance

It was possible to apply all three NLP techniques to the case study, although their language body was not specifically trained for material expressions in the construction industry. While GermaNET shows promising results in the ranges of shortest path similarity, the identification rate of synsets is too low. Therefore, using GermaNET for the proposed matching methodology is not pursued further.

The NLP library spaCy shows that different strategies of calculating the cosine similarity of material option spans and material category spans are improving the results. Furthermore, the tokenization of both material options and material categories, as well as choosing the maximum similarity of every calculated option, improve the result ranges significantly. However, the ranges are deviating too much and are generally too low, so further consideration for implementation is not planned.

The NLP technique BERT showed the most promising results. Low deviations of the result ranges and high cosine similarity of all strategies lead to applying it for the matching approach. Nevertheless, due to its large vectors with 786 dimensions, the calculation time is significantly higher than with spaCy and needs to be considered for further optimization.

5.3. Evaluation of element matching method

In this Section, the proposed matching method is tested with real-world case studies. In the first step, five office buildings were chosen, consisting of the required model information, such as element classification according to DIN 276 and materials. In the next step, the performance of the previously proposed matching method on element level using the best-performing NLP model, BERT, is analyzed for all case studies. In the last step, the ratio of correctly matched versus complete set is evaluated for each case study depending on their specific model quality.

According to the proposed matching method, as shown in Section 4.3, all elements and their materials are filtered and encoded, the similarities are calculated, and finally, the most similar element is selected. To evaluate the performance of the proposed matching method, all matched elements are evaluated according to correctness. If not matched correctly, the reason for wrong matching is recorded. For validation, a manual element matching is set as ground truth, also using the same LKdb.

Besides correct and wrong element matching, there are other reasons why correct matching was not possible. As the LKdb is just taking the most common elements into account, it is not covering all potential element structures. Therefore, one of the reasons for incorrect matching is the insufficient amount of available elements. Another reason for incorrect matching is that there is no valid cost group classification according to the German classification system DIN 276 available for the element to be matched. As a result, the algorithm cannot filter the relevant list of elements in LKdb, and no default element can be selected. Furthermore, also wrong classifications of the model's elements can lead to incorrect matching. This reason will be described in more detail in the following Sections. Finally, incorrect matching can also occur if the element's name and material's name are too generic or not existing. In this case, the default element of the classification group is matched according to the proposed matching method. In total, there are five different error clusters:

a) correctly matched
b) no correct matching element available in LKdb
c) wrong element classification
d) no valid element classification
e) too little information/ details
f) wrong matching

Figure 11 shows the matching performance of all case studies summed up, once weighted by the amount of individual elements (left) and, on the other hand, weighted by the element areas (right). The area-weighted result shows the influence of wrong matching according to the LCA relevant quantities, while the element-weighted results show the performance compared to the manual matching step.

The total element-weighted matching performance results show a correct matching of 78.1% for all five case studies. The biggest drivers of incorrect matching are due to too little information/ details (8.62%), no correct matching element available in LKdb (5.65%), and wrong element classification (5.50%). Nevertheless, the different ratios between element-weighted

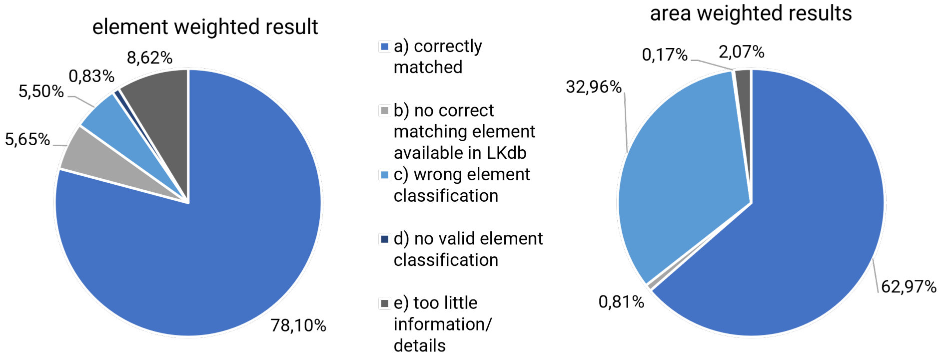
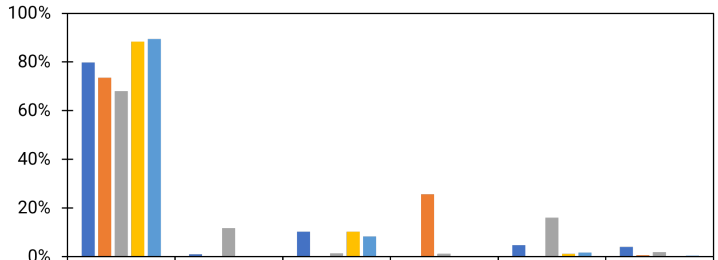


Figure 11: Total element matching performance of all case studies according to correct matches or matching error cluster, weighted by the amount of elements (left) and area of elements (right)

and area-weighted matching performances differ so widely that wrong element classification is 32.96%, and only 62.97% of the elements are correctly matched. Therefore, the results need to be analyzed in more detail and case-study-specific in the following.

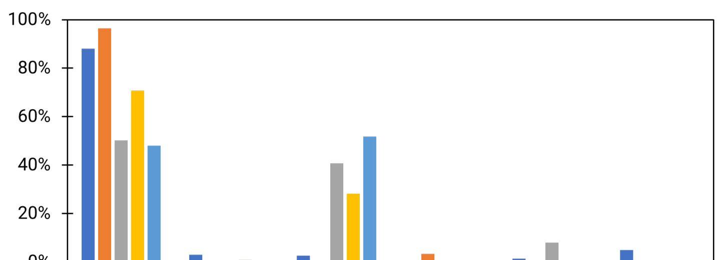
As shown in Figures 12 and 13, there are major differences in the error clusters between the different case studies and the weighting scenario. When looking at the element-weighted incorrectly matched elements of case study 2, the main error is no valid element classification with more than 25.0%, which is mainly due to a different classification nomenclature for windows (”B20” instead of ”334”). For weighting the scenario using the areas of the elements, the error is only 3.42%, and the correctly matched elements show the best performance of all case studies. Similar differences can be seen for case study 3, where the main error is due to clusters b) (11.68%) and e) (16.04%) in element weighting. In the area-weighted performance, these two clusters seem less significant compared to cluster c) (40.6%). This is due to the fact that the amount of elements is a different weighting factor. Nevertheless, as in case studies 4 and 5 are more columns modeled, which do not have the quantity of area but only length, the area-weighted performance results become significantly worse, although the element-weighted performance seems satisfying.

Generally, the matching performance shows satisfying results as, in total, 86,72% of the elements were correctly matched, or due to too little information, the default element was matched. 11,15% of the total elements were



	a) correctly matched	b) no correct matching element available in LKdb	c) wrong element classification	d) no valid element classification	e) too little information / details	f) wrong matched
Case study 1	79,76%	1,04%	10,24%	0,14%	4,83%	3,98%
Case study 2	73,62%	0,00%	0,00%	25,73%	0,00%	0,65%
Case study 3	67,98%	11,68%	1,30%	1,13%	16,04%	1,87%
Case study 4	88,37%	0,07%	10,23%	0,03%	1,26%	0,04%
Case study 5	89,40%	0,00%	8,33%	0,02%	1,70%	0,55%

Figure 12: Case-study-specific element matching performance according to correct matches or matching error clusters, weighted by the amount of element



	a) correctly matched	b) no correct matching element available in LKdb	c) wrong element classification	d) no valid element classification	e) too little information / details	f) wrong matched
Case study 1	88,05%	2,94%	2,54%	0,08%	1,42%	4,98%
Case study 2	96,36%	0,00%	0,00%	3,42%	0,00%	0,22%
Case study 3	50,23%	0,12%	40,60%	0,44%	7,90%	0,70%
Case study 4	70,69%	0,90%	28,15%	0,03%	0,21%	0,03%
Case study 5	48,07%	0,00%	51,72%	0,00%	0,19%	0,02%

Figure 13: Case-study-specific element matching performance according to correct matches or matching error clusters, weighted by the area of elements

wrongly matched as there are not sufficient classifications available. For only 0,83% of the total elements, the matching method results in wrong matches. The performance results differ due to model complexity and the quality of correct element classification according to DIN 276 of each real-world case study. The quality of LOD, sufficient amount of elements in LKdb, and wrong matching due to the proposed methodology and chosen NLP model seem to have a minor influence on the matching performance. There can be different matching performances depending if the total amount of matched elements or their areas are considered, which is mainly driven by influences of columns without area quantity sets. Considering the fact that tested IFC models were not optimized for this use case, the performance results prove the proposed matching method for real-world projects. The performance can be further increased by checking the model requirements of the elements' classification.

5.4. Evaluation of LCA result range calculation

Next, we chose one case study to validate the whole semantic healing process by evaluating the calculation of the embodied GHG emissions. As case study 2 shows in the area-weighted performance the best results, we select it for calculating the LCA results. The results will then be compared to a manual calculation, focusing on GWP as the main impact indicator. For the conventional LCA calculation, we chose the German LCA calculation tool eLCA [72]. Furthermore, only the total sum of all life cycle phases (A1-A3, B4, C3-C4, D) is considered to directly compare the final results of the examples. The reference period for this office building is 50 years, according to DGNB and BNB standards. The main goal of this evaluation is to show the results of the entire semantic healing workflow and its advantages compared to conventional processes. The optimization of element-specific LCA results itself is not the focus of this Section.

Figure 14 shows the GWP results clustered by cost groups (KG) and the total sum of the case study. Generally, the results show that the specific values of the conventional calculation following the manual, conventional workflow using eLCA are in the same range as the result ranges using the proposed methodology, including the matching method and the LKdb. The total manual result of 3,04 kg CO2-eq./ sqm*a calculated with eLCA is slightly lower than the range calculated by the proposed methodology and LKdb (Minimum 2,56, Median 3,25, Maximum 3,89 kg CO2-eq./sqm*a). To evaluate the difference in more detail, the element-specific results have to be analyzed.

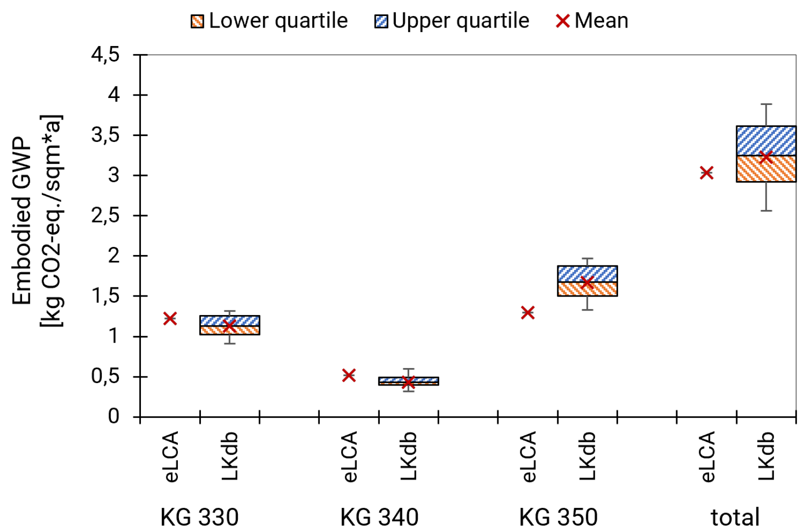


Figure 14: Total and cost group-specific results of Global Warming Potential (GWP) of case study 2 in [kg CO2-eq./ sqm*a]

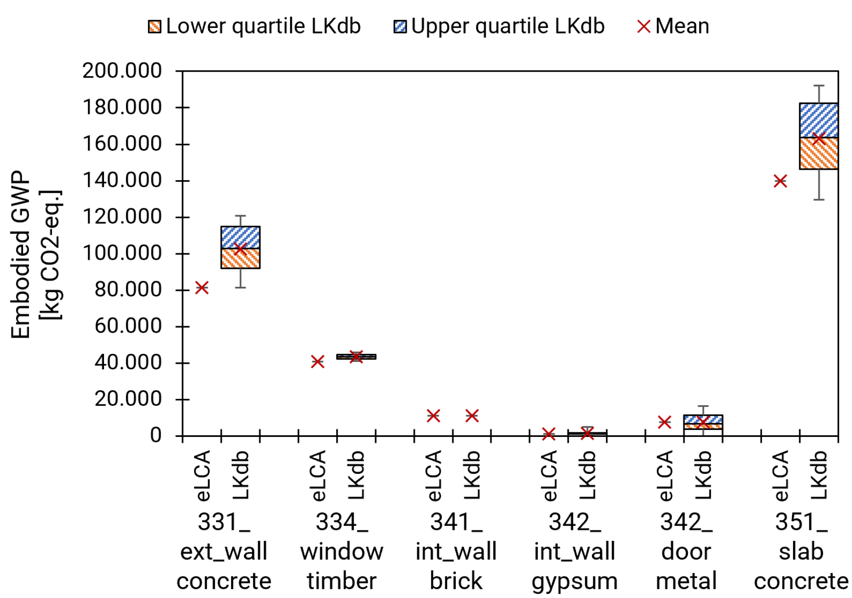


Figure 15: Element-specific results of Global Warming Potential (GWP) for selected elements of each classification group and different materials of case study 2 in [kg CO2-eq.]

Figure 15 shows the GWP results of the most relevant elements for each class according to the total sum of GWP over all life cycle phases. For each of the five chosen elements, on the left side, the results of the manual calculation using eLCA are shown, and on the right side, the automated calculated results using the matching method and LKdb are shown. The shown IFC elements consist of different element types, such as single- and multi-layer solid elements, windows and doors, or elements with composite materials. For the element with the cost group 331 and 351, reinforced concrete was matched, which consists of the materials reinforcement steel and concrete. While for the reinforcement steel, only one material option is available, for the concrete, there are several according to the specific compressive strength, which hasn't been specified in this early design phase yet. These different material options lead to a range of results for the total GWP.

In comparison, for the element of the cost group 341, the monolithic brick wall was chosen, while only one material option of brick is available in this case. For this reason, both results of eLCA and LKdb are identical and do not differ. For the selected door (KG 344), different EPDs are used in the LKdb, while for the manual selection, only one EPD was chosen. Usually, the LCA calculation of windows needs different quantity inputs as solid elements, as the functional units for the window frame are the length of the perimeter and the area for the transparent glass. The only varying material for the implemented LKdb windows is the frame material, which is, in this matched case, wood. In the LKdb, glass was implemented as only one material option per element, either single, double, or triple pane, and is therefore not varying. The total GWP range is not varying a lot due to a few different wood-based frame options, but also close to the manual calculation results.

Finally, the interior wall (KG 342) consists of a multi-layer element of plasterboard and mineral wool. In the IFC model, the element consists of four different layers of plasterboard, while in the LKdb, there are only two. Therefore, the different thicknesses were summed up so that the total thickness for plasterboard layers is the same. Nevertheless, also in this case, there are 26 different material options for plasterboard, which leads to a range for the total GWP results.

In general, the evaluation of the whole process shows reliable GWP results compared to manual calculation using eLCA. The results depend on the different element types and the level of information, which was decisive for the matching. Another aspect is that with the manual workflow in early design stages, the total GWP results of this case study seem to be lower than the average of the result range derived from the proposed methodology. This underlines the need for a semantic healing process to enable more realistic LCA result ranges based on this uncertain information.

5.5. Limitations

The authors had to make a couple of assumptions to validate the proposed methodology, which led to certain limitations. For implementing the LKdb and its embodied emissions values, the German database Ökobaudat was used, as all the applied case studies are located in Germany, and German material naming was used. An extension using other databases and mapping them to elements and material options can be easily realized and has been prototypically tested [73]. Nevertheless, the implemented elements in the LKdb only cover the most common element structures. Specific element structures for special cases need to be included in future work. So far, neither operational energy simulation nor life-cycle cost calculation is included in the database, as the focus of this publication is solely on embodied GHG emissions. Although we only discussed GWP results for evaluating the LCA calculation, other environmental impact metrics have been calculated, too, such as AP, EP, POCP, and ODP, as well as energy-related impact metrics.

The results of the element matching of five case studies presented in Section 5.3 show that a correct classification is crucial to match the IFC element to realistic LKdb elements. However, the German classification system DIN 276 was used, which cannot be directly transferred to other countries' classification systems. If IFC models have no or a lot of incorrectly classified cost group elements, the LCA results will differ significantly and are not meaningful.

Furthermore, the NLP model BERT employed here was not specifically trained for the application in the AEC context. Nevertheless, the results from the material and element matching showed that this circumstance does not affect the results due to the robust selection process of the matching method. Nonetheless, the bidirectional trained model leads to a high amount of vector dimensions for each expression and, as a result, a time-intensive computation process. A specific trained model could decrease the computational effort while providing similarly satisfying results as with BERT. For training such a model, a high amount of real-world data from different companies and designers is needed, which is difficult to collect due to privacy issues.

6. Conclusions and future research

To enable the calculation of embodied emissions of buildings in early design phases, automated workflows based on BIM models can be used to compare different design alternatives and find those solutions that have a minimal environmental impact. However, the uncertainties in these stages are unavoidable and missing information can lead to erroneous LCA results. Therefore, enriching vague models is crucial for calculating meaningful results, which are usually a range of results rather than single values. Among the most challenging boundary conditions is the fact that early-stage BIM models often lack precise specifications of object types and material properties. Instead, a wide range of mixed terminology is used, and some information remains completely unprovided. With this unstructured data, however, finding correct LCA information from the respective databases is almost impossible.

To overcome this issue of manual material matching and vague model information, in this paper, we propose a novel approach for automated semantic healing of BIM models. The proposed method allows assigning correct LCA information of element types and materials to the respective model element such that a reliable and holistic LCA can be calculated in early design stages. For the semantic healing process, an NLP-based method is used to enrich the model by automatically matching elements of an LCA Knowledge database (LKdb) to close the missing gap of the automation process of enriching LCA datasets to IFC materials and elements, and adding missing layer information of imprecise model elements. This LKdb contains all relevant information for the LCA calculation process, including LCA datasets on material level and different design alternatives, such as element variants of the same classification group or different material options of each element layer. Missing element layers are added to ensure reliable and consistent LCA results. The elements are matched by the most similar material or element names using the cosine similarity of the pre-trained NLP model vectors.

In an initial evaluation, different NLP models were compared by aligning the results of pre-matched materials of a case study. BERT was identified as having the best-performing results and proved to be suitable for the element-matching method. In a second evaluation, the proposed matching method was tested using five real-world BIM models, and their performances were analyzed. Generally, the proposed matching method proved to be satisfactory, correctly matching the majority of the IFC elements (86,72% success rate in total) to the corresponding LKdb elements. Nevertheless, the importance of correct classification of the IFC models is a relevant requirement for correct element matching. The success rate depends on the semantic model quality, mostly on correct and valid element classification for the initial filtering step. In a third evaluation, one of the five case studies was selected to calculate the embodied emissions focusing on global warming potential of each element and summing the resulting ranges up for the whole building. These results were compared to a manually calculated LCA using the tool eLCA, showing that the manual results are in the range of the results using the proposed method.

Finally, answering the research question raised, it can be confirmed that the proposed automated semantic healing methodology is sufficient for calculating embodied emissions based on early design BIM models. The main limitations are the processing time of the prototypical implementation using large NLP vector dimensions and the correct element classification, which can be error-prone in a manual workflow.

In our ongoing research, we plan to investigate the visualization of the results and selection process of element variants or specific material options. Using the geometric BIM model as an interactive representation and mapping the LCA results as color ranges has great potential for the visualization and selection process. Furthermore, the developed methodology and the LCA Knowledge database will be extended according to other element groups, such as HVAC, as well as further criteria, such as for operational energy simulation, LCC calculation, or circularity aspects. These criteria will also be included in the visualization and selection process.

7. Acknowledgments

The authors would like to gratefully acknowledge the support of Siemens Real Estate (SRE) for providing financial support for this research and giving access to the analyzed case studies.

8. Appendix

8.1. Acronyms

AP acidification potential

BDL building development level

BIM building information modeling

BNB Bewertungssystem Nachhaltiges Bauen für Bundesgebäude

BoQ bill of quantities

DGNB Deutsches Gütesiegel Nachhaltiges Bauen

EoL end of life

EP eutrophication potential

EPD Environmental Product Declaration

gbXML Green Building Extensible Markup Language

GHG greenhouse gas

GWP global warming potential

HVAC heating, ventilation, air conditioning

IFC Industry Foundation Classes

LCA life cycle assessment

LCC life cycle costs

LCI life cycle inventory

LCSA life cycle sustainability assessment

LKdb LCA knowledge database

LOD level of development

LOG level of geometry

LOI level of information

LOIN level of information needs

MEP mechanical electric plumbing

NLP natural language processing

ODP ozone depletion potential

POCP photochemical creation potential

Pset property set

RNN recurrent neural networks

sLCA social life Cycle Assessment

UUID universally unique identifier

VPL visual programming language

8.2. Nomenclature for equations

a_e Area of element (e)

D_{e_o} Recovery and disposal phase (C3-C4, D) of each element's (e) material option (o)

$d_{e_{o,i}}$ Thickness of element's (e) material option's (o) layer (i)

e Element

EIP_{c_o} Environmental Impact Potential of the construction phase (c) for each element (e)

e_o Element's material option

$f_{e_{o,i,x}}$ Quantities of each element's (e) material option's (o) layer (i) according to its dataset's functional unit (x)

l_e Length of element (e)

m_{e_o} Maximum amount of element's (e) material option's (o) layers

M_{e_o} Maintenance and replacement phase (B4) of each element's (e) material option (o)

n Maximum amount of element's (e) material options (o)

$n_{replacement,e}$ Frequency of replacement of each element (e)

o Material option

P_{e_o} Production phase (A1-A3) of each element's (e) material option (o)

$\rho_{o,i}$ Density of material option's (o) layer (i)

s_e Amount of element (e)

t_D Reference period of the whole building [years]

t_e Service life of the element (e)

x Functional unit of a dataset, either area (a), length (l), volume (v), mass (m), or amount (s)

8.3. Classification according to DIN 276 cost groups

Nr.	**2nd Level**	**Nr.**	**3rd Level**
320	Foundations		
330	External walls		
		331	Load-bearing external walls
		332	Non-load-bearing external walls
		333	External columns
		334	External doors and windows
		335	External cladding units
		336	Internal wall linings (of external walls)
		337	Prefabricated facade units
340	External walls		
		341	Load-bearing interior walls
		342	Non-load-bearing interior walls
		343	Interior columns
		344	Interior doors and windows
		345	Interior cladding units
		346	Elemental interior wall constructions
350	External walls		
		351	Ceiling constructions
		352	Ceiling openings
		353	Ceiling coatings
		354	Ceiling claddings
		355	Elemental ceiling structures
360	Roofs		

Table 2: Classification of LCA relevant cost group 300 (Structure - construction works) according to DIN 276 cost groups

References

[1] T. Abergel, B. Dean, J. Dulac, Global status report 2017, 2017.

[2] K. Safari, H. AzariJafari, Challenges and opportunities for integrating bim and lca: Methodological choices and framework development, Sustainable Cities and Society 67 (2021) 102728.

[3] P. Schneider-Marin, W. Lang, Environmental costs of buildings: monetary valuation of ecological indicators for the building industry, The International Journal of Life Cycle Assessment 25 (2020) 1637–1659.

[4] K. Forth, A. Braun, A. Borrmann, Bim-integrated lca - model analysis and implementation for practice, IOP Conference Series: Earth and Environmental Science 323 (2019) 012100.

[5] M. Locatelli, E. Seghezzi, L. Pellegrini, L. C. Tagliabue, G. M. Di Giuda, Exploring natural language processing in construction and integration with building information modeling: A scientometric analysis, Buildings 11 (2021) 583.

[6] J. Abualdenien, A. Borrmann, M. König, Ausarbeitungsgrade von BIM-Modellen, Springer Fachmedien Wiesbaden, Wiesbaden, pp. 165–191.

[7] BIM Forum, Level of development (lod) specification part i & commentary: For building information models and data (2020).

[8] A. Borrmann, M. König, C. Koch, J. Beetz (Eds.), Building Information Modeling: Technologische Grundlagen und industrielle Praxis, VDI-Buch, Springer Vieweg, Wiesbaden, 2021.

[9] A. Tomczak, L. van Berlo, T. Krijnen, A. Borrmann, M. Bolpagni, A review of methods to specify information requirements in digital construction projects, in: Prof. of CIB World Congress 2022.

[10] J. Abualdenien, A. Borrmann, A meta-model approach for formal specification and consistent management of multi-lod building models, Advanced Engineering Informatics 40 (2019) 135–153.

[11] N. GCR, Cost analysis of inadequate interoperability in the us capital facilities industry, National Institute of Standards and Technology (NIST) (2004).

[12] A. Cemesova, C. J. Hopfe, R. S. Mcleod, Passivbim: Enhancing interoperability between bim and low energy design software, Automation in Construction 57 (2015) 17–32.

[13] H. Lai, X. Deng, Interoperability analysis of ifc-based data exchange between heterogeneous bim software, Journal of Civil Engineering and Management 24 (2018) 537–555.

[14] J. L. Hernández, P. M. Lerones, P. Bonsma, A. v. Delft, R. Deighton, J.-D. Braun, An ifc interoperability framework for self-inspection process in buildings, Buildings 8 (2018) 32.

[15] T. Liebich, IFC4—the new buildingSMART standard, in: IC Meeting, bSI Publications Helsinki, Finland.

[16] G. B. Foundation, gbxml - an industry supported standard for storing and sharing building properties between 3d architectural and engineering analysis software, 28.05.2021.

[17] VDI 2552 Blatt 9, Building information modeling - classification systems, März 2022.

[18] DIN 276, Din 276:2018-12, kosten im bauwesen, 2018.

[19] DIN 277, Din 277:2021-08, grundflächen und rauminhalte im hochbau, 2021.

[20] DGNB GmbH, Dgnb system - new construction, buildings - criteria set: Version 2020 international (2020).

[21] BMI, Bewertungssystem nachhaltiges bauen (bnb) neubau büro- und verwaltungsgebäude: Bilanzierungsregeln für die erstellung von ökobilanzen, 2015.

[22] DIN EN 15643-2, Din en 15643-2:2021-12, nachhaltigkeit von bauwerken - bewertung der nachhaltigkeit von gebäuden – teil 2: Rahmenbedingungen für die bewertung der umweltbezogenen qualität fassung en_15643_2:2011, 2021.

[23] DIN EN ISO 14040, Din en iso 14040 : 2021-02, umweltmanagement_-ökobilanz_- grundsätze und rahmenbedingungen (iso_14040:2006_+ amd_1:2020); deutsche fassung en_iso_14040:2006_+ a1:2020, 2006.

[24] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2017.

[26] N. Wang, R. R. A. Issa, C. J. Anumba, Nlp-based query-answering system for information extraction from building information models, Journal of Computing in Civil Engineering 36 (2022).

[27] Q. Xie, X. Zhou, J. Wang, X. Gao, X. Chen, L. Chun, Matching real-world facilities to building information modeling data using natural language processing, IEEE Access 7 (2019) 119465–119475.

[28] G. Reitschmidt, Ökobilanzierung auf Basis von Building Information Modeling: Entwicklung eines Instruments zur automatisierten Ökobilanzierung der Herstellungsphase von Bauwerken unter Nutzung der Ökobau.dat und Building Information Modeling, Masterthesis, Technische Hochschule Mittelhessen, Gießen, 2015.

[29] A. Zahedi, J. Abualdenien, F. Petzold, A. Borrmann, Bim-based design decisions documentation using design episodes, explanation tags, and constraints, Journal of Information Technology in Construction 27 (2022) 756–780.

[30] B. Soust-Verdaguer, C. Llatas, A. García-Martínez, Critical review of bim-based lca method to buildings, Energy and Buildings 136 (2017) 110–120.

[31] L. Wastiels, R. Decuypere, Identification and comparison of lca-bim integration strategies, IOP Conference Series: Earth and Environmental Science 323 (2019) 012101.

[32] T. Potrč Obrecht, M. Röck, E. Hoxha, A. Passer, Bim and lca integration: A systematic literature review, Sustainability 12 (2020) 5534.

[33] M. Roberts, S. Allen, D. Coley, Life cycle assessment in the building design process – a systematic literature review, Building and Environment 185 (2020) 107274.

[34] C. Cavalliere, L. Brescia, G. Maiorano, T. Dalla Mora, G. R. Dell'Osso, E. Naboni, Towards an accessible life cycle assessment: A literature based review of current bim and parametric based tools capabilities, in: V. Corrado, E. Fabrizio, A. Gasparella, F. Patuzzi (Eds.), Proceedings of Building Simulation 2019: 16th Conference of IBPSA, Building Simulation Conference proceedings, IBPSA, 2020, pp. 159–166.

[35] A. Hollberg, J. Ruth, Lca in architectural design—a parametric approach, The International Journal of Life Cycle Assessment 21 (2016) 943–960.

[36] C. Llatas, B. Soust-Verdaguer, A. Passer, Implementing life cycle sustainability assessment during design stages in building information modelling: From systematic literature review to a methodological approach, Building and Environment 182 (2020) 107164.

[37] V. W. Tam, Y. Zhou, C. Illankoon, K. N. Le, A critical review on bim and lca integration using the iso 14040 framework, Building and Environment (2022) 108865.

[38] R. Santos, A. A. Costa, J. D. Silvestre, L. Pyl, Integration of lca and lcc analysis within a bim-based environment, Automation in Construction 103 (2019) 127–149.

[39] S. Eleftheriadis, P. Duffour, D. Mumovic, Bim-embedded life cycle carbon assessment of rc buildings using optimised structural design alternatives, Energy and Buildings 173 (2018) 587–600.

[40] S. Theißen, J. Höper, J. Drzymalla, R. Wimmer, S. Markova, A. Meins-Becker, M. Lambertz, Using open bim and ifc to enable a comprehensive consideration of building services within a whole-building lca, Sustainability 12 (2020) 5644.

[41] F. Rezaei, C. Bulle, P. Lesage, Integrating building information modeling and life cycle assessment in the early and detailed building design stages, Building and Environment 153 (2019) 158–167.

[42] C. Cavalliere, A. Hollberg, G. R. Dell'Osso, G. Habert, Consistent bim-led lca during the entire building design process, IOP Conference Series: Earth and Environmental Science 323 (2019) 012099.

[43] C. Llatas, B. Soust-Verdaguer, A. Hollberg, E. Palumbo, R. Quiñones, Bim-based lcsa application in early design stages using ifc, Automation in Construction 138 (2022) 104259.

[44] R. Santos, A. Aguiar Costa, J. D. Silvestre, L. Pyl, Development of a bim-based environmental and economic life cycle assessment tool, Journal of Cleaner Production 265 (2020) 121705.

[45] S. Theißen, J. Drzymalla, J. Höper, E. Liermann, R. Wimmer, A. Meins-Becker, A. Henne, N. Kloster, M. Lambertz, Digitalization of user-oriented demand planning through building information modeling (bim), IOP Conference Series: Earth and Environmental Science 588 (2020) 032004.

[46] S. Lee, S. Tae, H. Jang, C. U. Chae, Y. Bok, Development of building information modeling template for environmental impact assessment, Sustainability 13 (2021) 3092.

[47] S. Atik, T. D. Aparisi, R. Raslan, Investigating the effectiveness and robustness of performing the bim-based cradle-to-cradle lca at early-design stages: a case study in the uk, Proceedings of BSO-V - 5th Building Simulation and Optimization Virtual Conference (2020).

[48] J. P. Carvalho, I. Alecrim, L. Bragança, R. Mateus, Integrating bim-based lca and building sustainability assessment, Sustainability 12 (2020) 7468.

[49] M. Nilsen, R. A. Bohne, Evaluation of bim based lca in early design phase (low lod) of buildings, IOP Conference Series: Earth and Environmental Science 323 (2019) 012119.

[50] J. Veselka, M. Nehasilová, K. Dvořáková, P. Ryklová, M. Volf, J. Růžička, A. Lupíšek, Recommendations for developing a bim for the purpose of lca in green building certifications, Sustainability 12 (2020) 6151.

[51] X. Yang, M. Hu, J. Wu, B. Zhao, Building-information-modeling enabled life cycle assessment, a case study on carbon footprint accounting for a residential building in china, Journal of Cleaner Production 183 (2018) 729–743.

[52] R. Horn, S. Ebertshäuser, R. Di Bari, O. Jorgji, R. Traunspurger, P. von Both, The bim2lca approach: An industry foundation classes (ifc)-based interface to integrate life cycle assessment in integral planning, Sustainability 12 (2020) 6558.

[53] H. Figl, M. Ilg, K. Battisti, 6d bim–terminal: Missing link for the design of co 2 -neutral buildings, IOP Conference Series: Earth and Environmental Science 323 (2019) 012104.

[54] H. Abu-Ghaida, A. Kamari, An alternative approach to material and epd mapping in the development of bim-based lca and lcc tools, Proc. of the Conference CIB W78 2021, Luxembourg (2021).

[55] R. S. Nizam, C. Zhang, L. Tian, A bim based tool for assessing embodied energy for buildings, Energy and Buildings 170 (2018) 1–14.

[56] C. Bueno, L. M. Pereira, M. M. Fabricio, Life cycle assessment and environmental-based choices at the early design stages: an application using building information modelling, Architectural Engineering and Design Management 14 (2018) 332–346.

[57] A. Hollberg, D. Kaushal, S. Basic, A. Galimshina, G. Habert, A data-driven parametric tool for under-specified lca in the design phase, Bauphysik 588 (2020) 052018.

[58] C. Kiamili, A. Hollberg, G. Habert, Detailed assessment of embodied carbon of hvac systems for a new office building based on bim, Sustainability 12 (2020) 3372.

[59] A. Naneva, M. Bonanomi, A. Hollberg, G. Habert, D. Hall, Integrated bim-based lca for the entire building process using an existing structure for cost estimation in the swiss context, Sustainability 12 (2020) 3748.

[60] M. Röck, A. Hollberg, G. Habert, A. Passer, Lca and bim: Integrated assessment and visualization of building elements' embodied impacts for design guidance in early stages, Procedia CIRP 69 (2018) 218–223.

[61] C. Cavalliere, G. Habert, G. R. Dell'Osso, A. Hollberg, Continuous bim-based assessment of embodied environmental impacts throughout the design process, Journal of Cleaner Production 211 (2019) 941–952.

[62] A. Hollberg, A parametric method for building design optimization based on Life Cycle Assessment, Phd thesis, Bauhaus-Universität Weimar, 2016.

[63] G. Lobaccaro, A. H. Wiberg, G. Ceci, M. Manni, N. Lolli, U. Berardi, Parametric design to minimize the embodied ghg emissions in a zeb, Energy and Buildings 167 (2018) 106–123.

[64] E. Meex, A. Hollberg, E. Knapen, L. Hildebrand, G. Verbeeck, Requirements for applying lca-based environmental impact assessment tools in the early stages of building design, Building and Environment 133 (2018) 228–236.

[65] R. Di Bari, O. Jorgji, R. Horn, J. Gantner, S. Ebertshäuser, Step-by-step implementation of bim-lca: A case study analysis associating defined construction phases with their respective environmental impacts, IOP Conference Series: Earth and Environmental Science 323 (2019) 012105.

[66] E. Palumbo, B. Soust-Verdaguer, C. Llatas, M. Traverso, How to obtain accurate environmental impacts at early design stages in bim when using environmental product declaration. a method to support decision-making, Sustainability 12 (2020) 6927.

[67] R. K. Zimmermann, S. Bruhn, H. Birgisdóttir, Bim-based life cycle assessment of buildings—an investigation of industry practice and needs, Sustainability 13 (2021) 5455.

[68] M. Scherz, E. Hoxha, H. Kreiner, A. Passer, A. Vafadarnikjoo, A hierarchical reference-based know-why model for design support of sustainable building envelopes, Automation in Construction 139 (2022) 104276.

[69] BBSR, Nutzungsdauern von bauteilen - informationsportal nachhaltiges bauen, 2017.

[70] BBSR, Ökobaudat, 2021.

[71] P. Schneider-Marin, T. Stocker, O. Abele, M. Margesin, J. Staudt, J. Abualdenien, W. Lang, Earlydata knowledge base for material decisions in building design, Advanced Engineering Informatics 54 (2022) 101769.

[72] BBSR, elca, 2022.

[73] B. Lammers, K. Forth, Ifc-based variant analysis considering multicriterial sustainability analysis of buildings, Proc. of 33. Forum Bauinformatik, München (2022).

[74] V. Stenzel, Wissensdatenbank für Graue Energie und Treibhauspotenzial von Baustoffen, Masterthesis, München, Technische Universität München, 2020.

[75] BuildingSMART International Limited, Industry foundation classes: Version 4-addendum 2: Psetenvironmentalimpactindicators., 06.02.2020.

[76] C. Richter, S. Liedtke, BKI IFC-Bildkommentar: Ausgewählte IFC 4 Begriffe für die BIM-Planungsarbeit gegliedert nach DIN 276, BKI Kostenplanung, Rudolf Müller, Köln, 2021.

[77] W. J. Wilbur, K. Sirotkin, The automatic identification of stop words, Journal of Information Science 18 (1992) 45–55.

[78] Baustart für neues siemens technology center in garching, 19.10.2022.

[79] Siemens Deutschland, Frankfurts innovativstes büro, 19.10.2022.

[80] B. Hamp, H. Feldweg, Germanet – a lexical–semantic net for german, in: In Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications, pp. 9–15.

[81] V. Henrich, E. Hinirchs, Gernedit – the germanet editing tool, Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010). Valletta, Malta (May 2010) pp. 2228–2235.

[82] R. Navigli, F. Martelli, An overview of word and sense similarity, Natural Language Engineering 25 (2019) 693–714.

[83] Z. Wu, M. Palmer, Verbs semantics and lexical selection, in: J. Pustejovsky (Ed.), Proceedings of the 32nd annual meeting on Association for Computational Linguistics -, Association for Computational Linguistics, Morristown, NJ, USA, 1994, pp. 133–138.

[84] C. Leacock, M. Chodorow, Combining local context and wordnet similarity for word sense identification, in: Fellbaum, C., ed., WordNet: An electronic lexical, volume MIT Press, 1998, pp. 265–283.

[85] M. Honnibal, I. Montani, spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing, To appear 7 (2017) 411–420.

[86] S. Brants, S. Dipper, P. Eisenberg, S. Hansen-Schirra, E. König, W. Lezius, C. Rohrer, G. Smith, H. Uszkoreit, Tiger: Linguistic interpretation of a german corpus, Research on Language and Computation 2 (2004) 597–620.