Technical University of Munich

TUM Department of Civil, Geo and Environmental Engineering

Chair of Computational Modeling and Simulation

Automated Methods of Mapping LCA Data to BIM Models

Master's Thesis

of the Master of Science program Civil Engineering

Author:	Yue Xia
Matriculation Number:	[redacted]
1. Supervisor:	Prof. Dr.-Ing. André Borrmann
2. Supervisor:	M.Sc. Kasimir Forth
Date of Issue:	July 01, 2021
Date of Submission:	December 31, 2021

Preface

This master's thesis is an original, unpublished, independent work of the author Yue Xia under the guidance of the supervisor Kasimir Forth. It has been written to fulfill the graduation requirement of the Master of Science program at the Department of Civil, Geo and Environmental Engineering of Technical University of Munich (TUM).

The theme of the thesis and the research questions stemmed from the discussion with the supervisor in April 2021. Taking into account that the automation and digitization level of the Civil Engineering is considerably low compared with other engineering and industrial fields, and with great interest in the integration of Building Information Modeling (BIM) and building Life Cycle Assessment (LCA), this research explored the feasibility of introducing Natural Language Processing (NLP) into the Industry Foundation Classes (IFC) data mapping task. Having conducted a systematic literature review and developed an automated method of mapping LCA data to BIM models, this thesis is expected to make a small contribution to the automation and digitization of the modern building sector, especially to the building data exchange and processing process.

In addition, I would like to thank my supervisor for his excellent guidance since our first meeting in April, 2021. Your ideas and suggestions have always inspired me and benefited me a lot. Without your support I would not have been able to accomplish this thesis. Furthermore, I would also like to thank TUM for providing me with the educational access to all the literature databases, software and digital tools.

Finally, my friends and parents deserve a particular note of thanks: Your kind words and encouragement have, as always, kept me motivated.

Yue Xia

December 31, 2021

Abstract

How we design, construct and operate buildings and other infrastructures has a profound influence on the economy and ecology of the society and the earth. Building Information Modeling (BIM) based Life Cycle Assessment (LCA) is an effective method to quantify and alleviate the ecological effect of buildings while considering the economic aspect. As the international BIM standard, Industry Foundation Classes (IFC) format plays an irreplaceable role in the BIM-LCA integration notably in the data exchange. In addition, with the rapid development of computer science and artificial intelligence, the application of which in the architecture, engineering and construction (AEC) industry arouses more and more interest in recent years. Thereinto, the technology Natural Language Processing (NLP) is regarded as a powerful helping hand for dealing with data exchange and processing tasks particularly the mapping of BIM (IFC) data to the predefined LCA profiles.

On the basis of a systematic literature review, this thesis summarizes state-of-the-art utilization of LCA, BIM, and NLP in the AEC industry, and compares three kinds of BIM-LCA integration strategies: manual, semi-automated, and automated approaches. Moreover, an automated method of mapping LCA data to BIM models which introduces NLP technology is proposed, including the methodology and a prototypical workflow. To verify the feasibility of this method, a case study is implemented, where an IFC file is at first extracted from a BIM model, followed by the processing of the LCA-related data contained in the IFC file. Then the processed information (material names) is mapped to the LCA database Ökobaudat respectively through manual attempt and three different NLP tools – Gensim, spaCy, and BERT. In the end, after quantitatively analyzing and comparing the contributions of the three NLP technologies to the accuracy of the mapping task, performing the automated method based on the pre-trained BERT model in conjunction with manual checking and adjustment is concluded to be the most efficient and recommendable.

Zusammenfassung

Wie wir Gebäude und andere Infrastrukturen planen, bauen und betreiben, hat einen tiefgreifenden Einfluss auf die Ökonomie und Ökologie der Gesellschaft und der Erde. Die Ökobilanzierung (LCA) basierend auf Building Information Modeling (BIM) ist eine effektive Methode, um die ökologischen Auswirkungen von Gebäuden unter Berücksichtigung des ökonomischen Aspekts zu quantifizieren und zu mindern. Als internationaler BIM-Standard spielt das Format Industry Foundation Classes (IFC) eine unersetzliche Rolle bei der BIM-LCA-Integration, insbesondere beim Datenaustausch. Darüber hinaus hat mit der rasanten Entwicklung der Informatik und der künstlichen Intelligenz ihre Anwendung in der Bau-, Ingenieur- und Bauindustrie (AEC) in den letzten Jahren immer mehr Aufmerksamkeit auf sich gezogen. Die Technologie Natural Language Processing (NLP) gilt als leistungsstarkes Hilfswerkzeug für die Bewältigung von Datenaustausch- und Verarbeitungsaufgaben, insbesondere der Abbildung von BIM (IFC)-Daten auf vordefinierte LCA-Profile.

Auf der Grundlage einer systematischen Literaturrecherche fasst diese Masterarbeit den aktuellen Stand der Nutzung von LCA, BIM und NLP in der AEC-Branche zusammen und vergleicht drei Arten von BIM-LCA-Integrationsstrategien: manuelle, halbautomatische und automatisierte Methoden. Hinzu kommt wird eine automatisierte Methode von Abbildung der Ökobilanzdaten zu BIM-Modellen vorgeschlagen, die die NLP-Technik einführt, einschließlich der Methodik und eines prototypischen Workflows. Um die Machbarkeit dieser Methode zu überprüfen, wird eine Fallstudie durchgeführt, bei der zunächst eine IFC-Datei aus einem BIM-Modell extrahiert wird, gefolgt von der Verarbeitung der in der IFC-Datei enthaltenen LCA-bezogenen Daten. Anschließend werden die verarbeiteten Informationen (Materialnamen) durch manuellen Versuch und drei verschiedene NLP-Tools – Gensim, spaCy und BERT – zu der Ökobilanzdatenbank Ökobaudat abgebildet. Am Ende, nach quantitativer Analyse und Vergleich der Beiträge der drei NLP-Technologien zur Genauigkeit der Abbildungsaufgabe, ist die Durchführung der automatisierten Methode auf Basis des vortrainierten BERT-Modell in Verbindung mit manueller Überprüfung und Anpassung am effizientesten und empfehlenswertesten.

Contents

List of Figures

List of Tables

List of Abbreviations

2D	Two-dimensional
3D	Three-dimensional
4D	Four-dimensional
5D	Five-dimensional
6D	Six-dimensional
7D	Seven-dimensional
ADP	Abiotic Depletion Potential
AEC	Architecture, Engineering and Construction
AI	Artificial Intelligence
AP	Acidification Potential
API	Application Programming Interface
BCF	BIM Collaboration Format
BIM	Building Information Modeling
BOQ	Bill of Quantities
CS	Computer Science
ELCD	European Life Cycle Database
EP	Eutrophication Potential
GHG	Greenhouse Gas
GUID	Global Unique Identifier
GWP	Global warming potential
IDM	Information Delivery Manual
IFC	Industry Foundation Classes
LCA	Life Cycle Assessment

LCC	Life Cycle Cost
LCEA	Life Cycle Energy Assessment
LCI	Life Cycle Inventory
LCSA	Life Cycle Sustainability Assessment
LOD	Level of Development
MVD	Model View Definition
NIBS	National Institute of Building Sciences
NL	Natural Language
NLP	Natural Language Processing
ODP	Ozone Depletion Potential
PE	Primary Energy
PENR	Non-renewable Primary Energy
PER	Renewable Primary Energy
POCP	Photochemical Ozone Creation Potential
QTO	Quantity Takeoff
SLR	Systematic Literature Review
STEP	STandard for the Exchange of Product model data
WorldGBC	World Green Building Council
XML	Extensible Markup Language

1 Introduction

The architecture, engineering and construction (AEC) industry yearly consumes approximately 3 billion tons of raw materials, being responsible for up to 40% of solid waste and 25-40% of global energy use (Valle, 2021). Furthermore, it accounts for a significant portion (about 40%) of global greenhouse gas (GHG) emissions (WorldGBC, 2019). When moving up to cities, the impacts become even larger, being obligated to 70% of GHG emissions and more than 66% of electricity consumption (Hauschild, Rosenbaum & Olsen, 2018). From another perspective, the building sector has great potential in reducing environmental impacts and mitigating global warming.

Life Cycle Assessment (LCA) is the established methodology for quantifying the environmental impacts of product systems during their creation, operation, and end-of-life phases (Klöpffer, 2014). Having gone through years of continuous development, enrichment and refinement in both methodology and applications, LCA is currently applied to dozens of fields including the AEC industry, where it is utilized to evaluate the environmental performance of buildings and other infrastructures (Hauschild, Rosenbaum & Olsen, 2018).

With the rapid development of computer science (CS) in the last few decades, the possibility of its application in other fields is constantly being explored, as well as in the AEC industry, which is exposed to new input and more information regarding the digitization and computerization of this sector (BibLus, 2021). Building Information Modeling (BIM*) is one of the most notable applications of computer technology in the AEC industry, being supported by a large number of professional platforms and software. BIM allows to cloud-create, -access and -modify models, consequently improving the collaboration and communication between project teams and stakeholders. Meanwhile, BIM efficiently integrates various interdisciplinary functions such as clash detection, cost estimation and project scheduling based on a central model, which reduces the risk, saves the cost, and strengthens the handover (Hall, 2021). Moreover, BIM facilitates the method advance and can support the application

* BIM is an abbreviation that can represent three terms: Building Information Modeling, Building Information Model, and Building Information Management. In this thesis, BIM stands for Building Information Modeling, unless otherwise specified.

of LCA/ Life Cycle Cost (LCC) in the AEC industry to a large extent through BIM-based LCA analysis and BIM-LCA integration (Wastiels & Decuypere, 2019).

Since BIM is a product of CS, and a lot of LCA calculations and analyse are dependent on computer tools, there is still great potential of optimizing BIM-based LCA with CS techniques such as Artificial Intelligence (AI). As an advanced technology in the field of AI, Natural Language Processing (NLP) is known as a powerful instrument to cope with semantic problems, which can be also applied to processing semantic information during the BIM-LCA integration.

1.1 Motivation and Objectives

Up to now, LCA is generally implemented as a post-design evaluation in later design stages, where there is sufficient building information and the design plan has been determined. It is not used to support or optimize design decisions during early design stages (Potrč Obrecht et al., 2020; Meex et al., 2018).

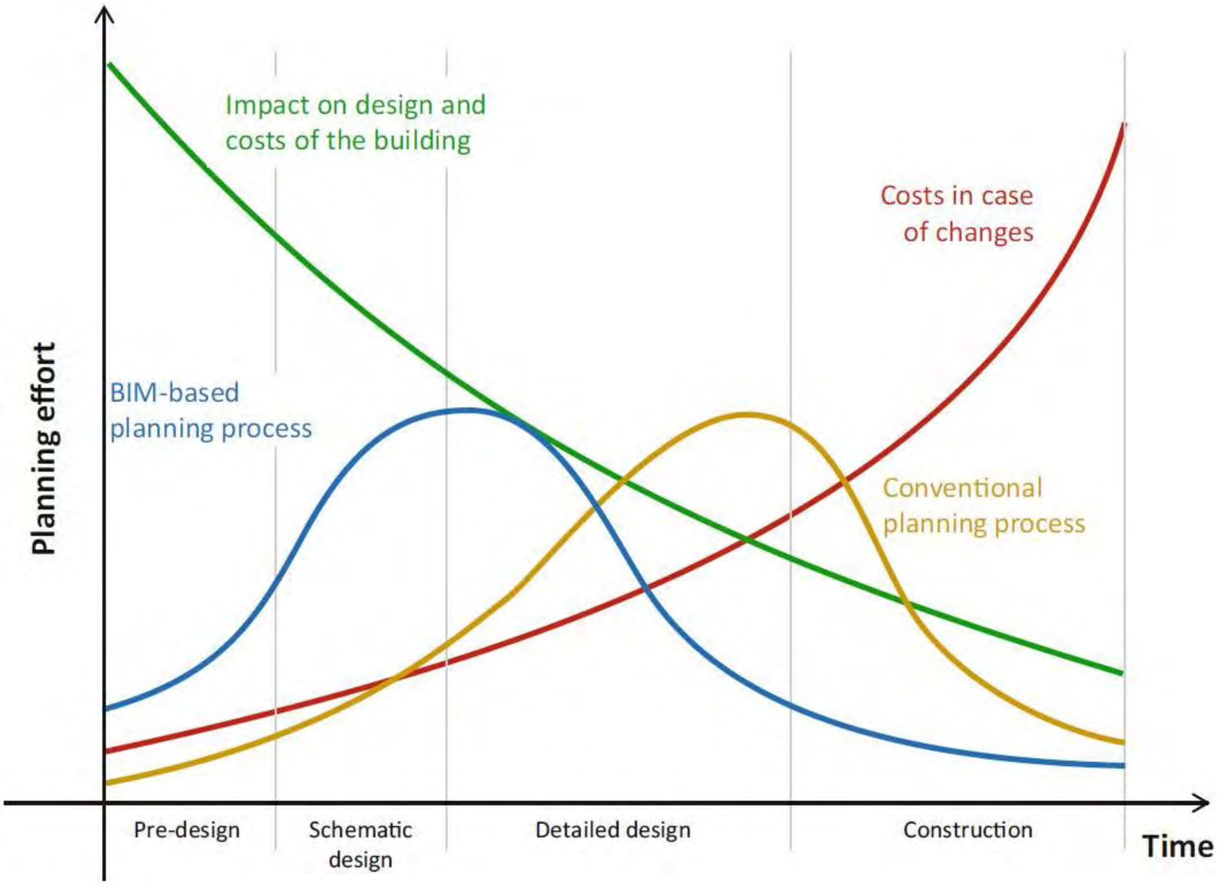


Figure 1-1 MacLeamy Curve – Planning effort over time (WP-1202, 2004; Borrmann et al., 2018)

However, as depicted by the green and red curves in Figure 1-1, with the progress of building design and construction, the impact on design and costs rapidly decreases, while the costs in case of changes dramatically increases. In this context, developing methods for applying LCA in early design stages is the research trend and focus, for which numerous obstacles need to be overcome. First and foremost, there are a lot of uncertainties due to the lack of information (Rezaei, Bulle & Lesage, 2019), e.g., unspecified materials and missing layers, which is required to decide upon holistic calculations to compare design variants by considering different use cases, such as LCA or Life Cycle Cost (LCC) calculations. In addition, another limitation of current LCA is that most calculations are carried out manually or semi-automatically, which is inefficient and time-consuming (Azizoglu & Seyis, 2020).

BIM can effectively integrate LCA and LCC in early design stages by shifting planning effort and design decisions to earlier phases (see Figure 1-1), making it possible to influence the design, performance and costs in early design phases, greatly improving the quantification and balance of the ecological and economic impact of buildings (Lu et al., 2021; Borrmann et al., 2018).

The open and international BIM standard, industry foundation classes (IFC), provides a standardized data model that codifies all information in a logical way, e.g., the identity, semantics, characteristics, attributes and relationships of objects, abstract concepts like performance and costing, and the processes of installation and operations (buildingSMART International, 2021b). In the meantime, it promotes vendor-neutral and usable capabilities across a wide range of hardware devices, software platforms and interfaces for various use cases (buildingSMART International, 2021a). With its characteristics and advantages, it plays a crucial role in the semi-automated and automated data exchange process of BIM-LCA integration, and thereby becomes a hot research point of which in recent years.

Although much progress has been made in the field of BIM-based LCA data exchange via the IFC format, it still does not work perfectly, which specifically manifests in that data loss and misinterpretation occur from time to time (Borrmann et al., 2018; Safari & AzariJafari, 2021). Therefore, current BIM-based LCA implementation relies heavily on the laborious manual work (Safari & AzariJafari, 2021).

With regard to this, this master's thesis aims to explore the way towards employing NLP to optimize the data exchange process of BIM-based LCA, which is expected to reduce data loss, enhance the BIM-LCA integration and improve the automation and digitization.

1.2 Outline

In Chapter 2, the theoretical and research background is presented, including the technologies LCA, BIM, IFC and NLP, as well as their state-of-the-art utilization in the AEC industry. In addition, several of the most commonly-used LCA databases and NLP tools are introduced. At the end of this chapter, the strategies and some relevant studies of BIM-LCA integration and BIM data mapping methods are summarized.

On the basis of the literature review in Chapter 2, the research questions are determined and phased first in Chapter 3. Then, an NLP-based automated method of mapping LCA data to BIM models is proposed, including the methodology, a prototypical workflow and the employed BIM/ LCA/ NLP/ computational tools.

In accordance with the workflow structured in Chapter 3, a case study is carried out and presented in Chapter 4, for the purpose of verifying the feasibility and effectiveness of the automated mapping method.

Chapter 5 discusses the results of the case study and the limitations of the automated mapping method.

The last chapter (Chapter 6) is an overall conclusion, which summarizes the results and contributions of the thesis, and makes an outlook on the future work.

2 State of the Art

2.1 LCA – Life Cycle Assessment

2.1.1 Definition and Methodological Framework of LCA According to ISO 14040

The international standards ISO 14040 defines life cycle as “consecutive and interlinked stages of a product system, from raw material acquisition or generation from natural resources to final disposal”, and Life Cycle Assessment (LCA) as “compilation and evaluation of the inputs, outputs and the potential environmental impacts of a product system throughout its life cycle” (ISO 14040, 2006).

Moreover, ISO 14040 standardizes the methodological framework of LCA, which contains four phases, namely the goal and scope definition, life cycle inventory analysis, life cycle impact assessment, and interpretation (ISO 14040, 2006). As shown in Figure 2-1, the LCA results can be used for various purposes in multiple fields, such as agriculture, energy systems, and significantly the AEC industry (Klöpffer, 2014).

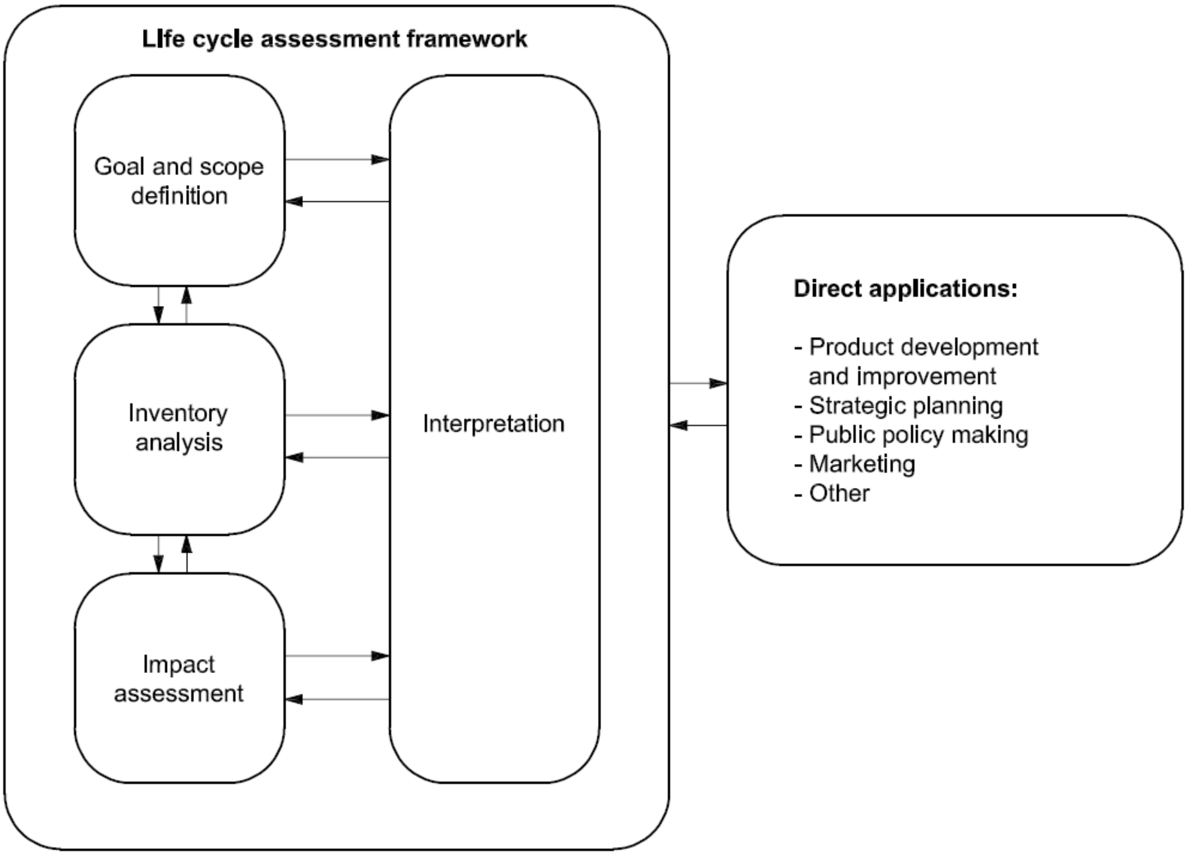


Figure 2-1 Methodological framework of LCA (ISO 14040, 2006)

2.1.2 Building Life Cycle Stages According to EN 15643-2

In accordance with the standard EN 15643-2, the life cycle of a building can be divided into four stages: product stage (A1 – A3), construction stage (A4 – A5), use stage (B1 – B7), and end of life stage (C1 – C4) (EN 15643-2, 2011). More specifically, the four stages together with the division of sub-stages are illustrated in Figure 2-2.

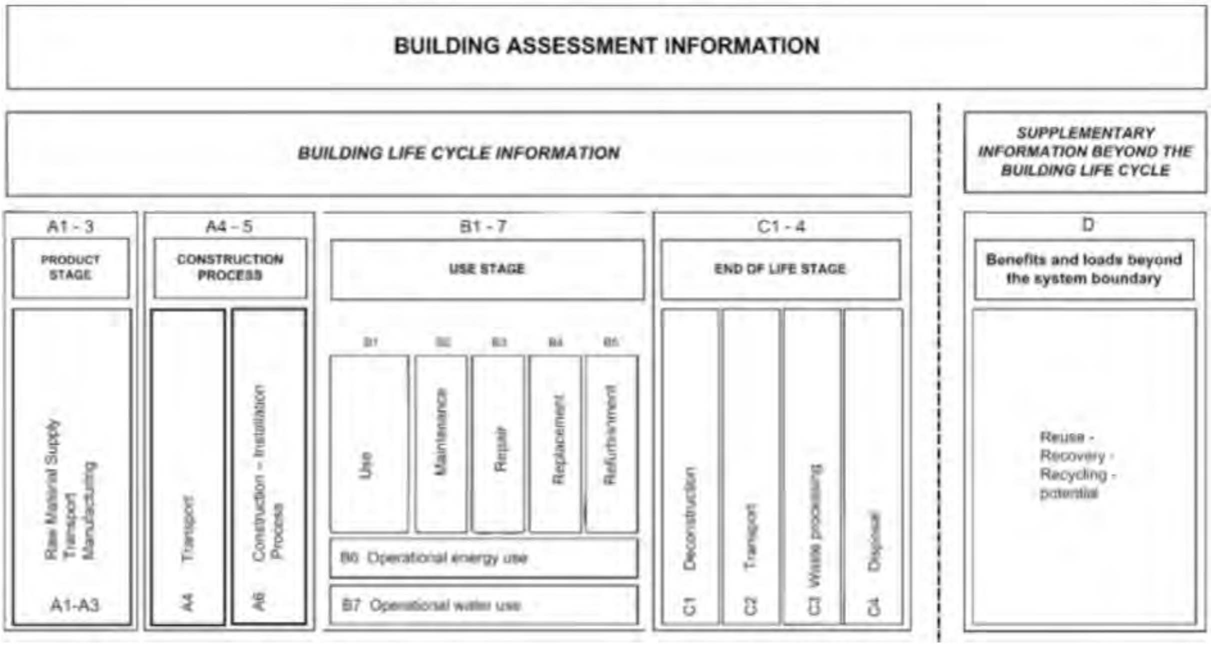


Figure 2-2 Building life cycle stages (EN 15643-2, 2011)

In addition to the four life cycle stages, the supplementary information – benefits and loads beyond the system boundary – needs to be considered when implementing a building LCA, i.e., the potential for reuse, recovery and recycling.

2.1.3 LCA Applied in the AEC Industry

In recent years, the application of LCA in the AEC industry has seen increasing interest, especially the utilization in early design stages, when it’s in time to make changes to the project, instead of being applied as a post-design evaluation in later design stages (Potrč Obrecht et al., 2020).

In light of the SLR conducted by Nwodo and Anumba, the general objective of building LCA in early design stages can be summarized as minimizing environmental impacts, carbon emissions, energy and cost (Nwodo & Anumba, 2019). To this end, a lot of scientists have dived into this area and made some achievements. For instance, some of them took uncertainty and sensitivity analyses emphasizing on construction materials to evaluate life cycle embedded energy or carbon emission of buildings

(Häfliger et al., 2017; Schneider-Marin et al., 2020). Some of them carried out multi-criteria analysis for buildings' sustainability assessment (Hu, 2019; Streimikiene et al., 2020). There are also scientists who explored the potential for implementing LCA in Green Building Rating System (Dalla Valle, 2021; Sartori et al., 2021), and created automated LCA methods and tools for the decision-making on building design options (Kumanayake & Luo, 2017).

Additionally, the German standard DIN 276 classifies and defines building components into eight general cost groups, which are further divided into several subgroups (DIN 276, 2018). As listed in Table 2-1, the cost group number and cost group name are associated with the cost group to which the building element belongs, which can be exploited to calculate the building LCC. For example, in the LCA web application eLCA the building environmental impacts can be determined and evaluated based on the cost groups.

Table 2-1 Part of the building cost groups according to DIN 276

Cost Group Number and Cost Group Name			
100	**Grundstück**		
200	**Vorbereitende Maßnahmen**		
300	**Bauwerk – Baukonstruktionen**		
	310	Baugrube	
	320	Gründung	
	330	Außenwände	
		331	Tragende Außenwände
		332	Nichttragende Außenwände
		333	Außenstützen
	340	Innenwände	
	350	Decken	
	360	Dächer	
400	**Bauwerk – Technische Anlagen**		
500	**Außenanlagen und Freiflächen**		
600	**Ausstattung und Kunstwerke**		
700	**Baunebenkosten**		
800	**Finanzierung**		

Furthermore, the WorldGBC released a report entitled “Bringing embodied carbon upfront” in 2019, for the sake of eliminating the life cycle emissions of buildings and construction sector by 2050 through coordinated actions that tackle embodied carbon (WorldGBC, 2019).

In conclusion, the application of LCA in the AEC industry is of great help for optimizing the decision-making of design options and construction materials by weighing up the pros and cons. Furthermore, with the technological updates especially the gradual maturity of BIM and CS technologies, the development of BIM-based LCA and BIM-LCA integration is becoming a research hotspot, which will be later presented in Chapter 2.4.

2.1.4 LCA/ LCI Databases

Ideally, the primary data on material and energy inputs should be obtained when conducting an LCA study. However, some data are not lightly available, for example, on raw material extraction and processing (Kalverkamp & Karbe, 2018). Therefore, numerous LCA or Life Cycle Inventory (LCI) databases have been established to support LCA studies, e.g., European databases like Ecoinvent, GaBi and ELCD, American databases like Athena and USLCI, and other databases like Base Carbone, BEDEC, CPM LCA, ProBas, ECORCE, KBOB and Ökobaudat (Martínez-Rocamora et al., 2016).

Correspondingly, most building LCA studies take use of predefined datasets for the materials or components to simplify the process and reduce the workload of data collection, selection and validation (Hollberg et al., 2021). The following paragraphs and Table 2-1 will outline and compare six of the most prevailing LCA/ LCI databases.

1. *Ecoinvent*[1]

The Ecoinvent database is a commercial database from Switzerland. As one of the first comprehensive LCI databases and one of the world’s leading data systems, it contains exceeding 14,000 datasets from both Switzerland and Europe in the fields of transport, energy, chemicals, agriculture, metals, building materials, waste treatment, etc. The database is mainly based on unit processes. In contrast to accumulated process data, these data records do not involve multiple unit process chains (Kalverkamp & Karbe,

[1] https://www.ecoinvent.org/

2018). With its consistency and transparency, Ecoinvent is perfectly suitable for construction purposes, since almost all categories of construction materials are incorporated.

2. GaBi[2]

The GaBi LCA database is also a market-oriented database that offers more than 15,000 plans and processes based mostly on the primary data collection across the globe, spanning most industries, e.g., agriculture, building and construction, chemicals and materials, electronics, energy, food, education, healthcare. Thereinto, in excess of 1,000 processes are related to construction materials and predominantly cradle-to-gate (Martínez-Rocamora et al., 2016). In addition to its wide range, it is annually updated by more than 300 life cycle experts from over 20 countries to keep all datasets up-to-date.

3. ELCD[3]

Originated from a project of the Joint Research Center of the European Commission, European Life Cycle Database (ELCD) comprises over 300 entries including some key materials, energy, transport, waste management and other areas, which is much fewer compared to Ecoinvent and GaBi. Its limited number of datasets for construction materials indicates that it needs to be complemented with other databases. Unfortunately, ELCD has been discontinued in the middle of 2018, after which, it is no longer available online, while is still downloadable as a ZIP package.

4. USLCI[4]

The U.S. Life Cycle Inventory (USLCI) Database was developed by the National Renewable Energy Laboratory of the U.S. Department of Energy, providing individual gate-to-gate, cradle-to-gate, and cradle-to-grave processes, and taking into account input and output flows of energy and materials exclusively for the U.S. This database focuses on metals, wood materials, and plastics, which constitute approximately 80 out of the 600 processes in the whole database (Martínez-Rocamora et al., 2016).

5. ProBas[5]

[2] https://gabi.sphera.com/international/databases/gabi-databases/
[3] https://eplca.jrc.ec.europa.eu/ELCD3/
[4] https://www.nrel.gov/lci/
[5] https://www.probas.umweltbundesamt.de/php/index.php

The Web Portal Prozessorientierte Basisdaten für Umweltmanagement-Instrumente (ProBas, Process-oriented Basic Data for Environmental Management Instruments) is a well-structured German database that integrates a large amount of publicly available data sources comprising energy, materials and products, transport, waste disposal and so on, in order to provide the broadest possible spectrum of life cycle data. Exceeding 8,000 unitary data records can be queried using extensive search and filter functions, containing around 700 unitary processes relevant to construction materials.

6. Ökobaudat[6]

With the support of German Federal Ministry of the Interior, Building and Community, Ökobaudat is a platform with a standardized LCA database as the core, involving datasets on building materials, construction, transport, energy and disposal prosses. Outstandingly, it offers more than 1,400 strictly checked and selected datasets for building products, including both generic and specific environmental declaration datasets from diverse companies or associations. All datasets comply with DIN EN 15804 and the Assessment System for Sustainable Building (BNB, Bewertungssystem Nachhaltiges Bauen). In addition to providing a web user interface, Ökobaudat also has a standardized interface for data exchange with some other applications and software tools, and it can be downloaded as a ZIP archive or in CSV format as well.

[6] https://www.oekobaudat.de/en.html

2.2 BIM – Building Information Modeling

2.2.1 Definition of BIM According to NBIMS-US V3

One of the most acknowledged definitions of BIM comes from National BIM Standard – United States Version 3 (NBIMS-US V3) published by National BIM Standards Project Committee. It defines BIM as "a business process for generating and leveraging building data to design, construct and operate the building during its life cycle (from earliest conception to demolition)", and Building Information Model* being interpreted as "a digital representation of physical and functional characteristics of a facility" (NIBS, 2015). In other words, BIM is known as not only the process of using building data to create and modify the BIM model, but also a powerful digital and computerized assistant in the process of facility construction, operation and maintenance.

2.2.2 BIM Dimensions

The BIM dimensions refer to the levels of information in a given BIM model (BibLus, 2021). Every time a specific type of information is added to the BIM model, a corresponding dimension is generated (BibLus, 2021).

Table 2-2 BIM dimensions (United BIM, 2020)

	3D BIM	4D BIM	5D BIM	6D BIM	7D BIM
Definition	(x, y, z) Geometric and graphical information	3D BIM + Duration, Timeline, Schedule	4D BIM + QTO, Cost estimation, Budgetary analysis	5D BIM + Sustainability, Energy efficiency	6D BIM + Facility management

Up to now, there are seven recognized BIM dimensions. Besides traditional two-dimensional (2D) drawings (plans, elevations and sections) and three spatial dimensions of building designs (3D BIM), BIM model has been extended to more dimensions incorporating information about time (4th dimension), cost (5th dimension), sustainability (6th dimension), and facility management (7th dimension) (United BIM,

* Building Information Model will be abbreviated as BIM model hereinafter.

2020). The specific information contained in BIM model with different dimensions and their respective definitions are listed in Table 2-2.

As the most basic type of BIM model, 3D BIM comprises the three geographical dimensions (x, y, z) of a structure, which is commonly used for 3D visualization and rendering of the design (Höflich & Maier Consult GmbH, 2021). Apart from this, model checking like code checking and clash detection is another usage, i.e., the verification of the model adherence to the project or to standards requirements, and the preventive analysis of the possible geometric conflicts (BibLus, 2021).

The fourth dimension of a BIM model stands for the project duration and timeline, showing how the project evolves over different phases, which helps to improve site planning and optimize the schedule (Axis Steel Detailing, LLC., 2021). Compared with traditional scheduling tools such as Gantt and Pert charts, 4D BIM is more flexible to dynamically simulate and adjust project phases rather than displaying project tasks and timeline with static bar chart or flowchart (BibLus, 2021). Moreover, it improves the data delivery efficiency and enhances the communication between stakeholders (BibLus, 2021).

The fifth dimension is associated with quantity takeoff (QTO) and cost estimation. 5D BIM is particularly useful in decision-making in the early phases of a project for it supports QTO and cost estimation under different design and construction scenarios (Axis Steel Detailing, LLC., 2021). Meanwhile, it helps with budgetary analysis by visualizing the predicted and actual costs of a project over time with real-time notifications of cost overrun (Axis Steel Detailing, LLC., 2021). However, since the QTO and cost estimation are updated simultaneously with project evolvement, a high risk of data loss exists with large probability (BibLus, 2021).

On the basis of 5D BIM, sustainability is introduced into 6D BIM to accomplish self-sustainability and energy-efficiency. It supports exhaustive analyses in terms of economic and environmental sustainability by creating overall and partial energy usage estimates during the initial design phase (BibLus, 2021; Axis Steel Detailing, LLC., 2021). Besides, 6D BIM offers the opportunity for detailed visualizing and planning component logistics and disposal, which in the long run can facilitate operational management (Höflich & Maier Consult GmbH, 2021; Axis Steel Detailing, LLC., 2021).

The additional focal point of 7D BIM is facility management. It provides an “as-built” model that not only can visualize the planning and scenario of maintenance and operation, but allows stakeholders to track important asset data like status, warranty information, technical specifications, and maintenance and operation manuals (Höflich & Maier Consult GmbH, 2021; Axis Steel Detailing, LLC., 2021). In the meantime, 7D BIM is also commonly applied to BIM strategy based LCA/ LCC (Höflich & Maier Consult GmbH, 2021).

Laying a solid foundation for the digitalization and automation of the AEC industry as well as creating great potential for achieving various purposes, BIM is now considerably employed by individuals, enterprises and government departments throughout the entire life cycle of buildings and diverse infrastructure (3Units Technology, 2021).

2.2.3 Little/ Big BIM and Closed/ Open BIM

To stepwisely shift conventional drawing-based workflows to model-based ones without unduly unsettling the established basic functioning, different technological levels of BIM implementation are distinguished, namely little BIM, big BIM, closed BIM, and open BIM (Borrmann et al., 2018). Thereof, little BIM is opposite to big BIM, which refer to the extent of BIM usage, and closed BIM is opposite to open BIM, which care about the variety of software vendors. Figure 2-3 illustrates the characteristics, differences, and relationships of the four BIM implementation options.

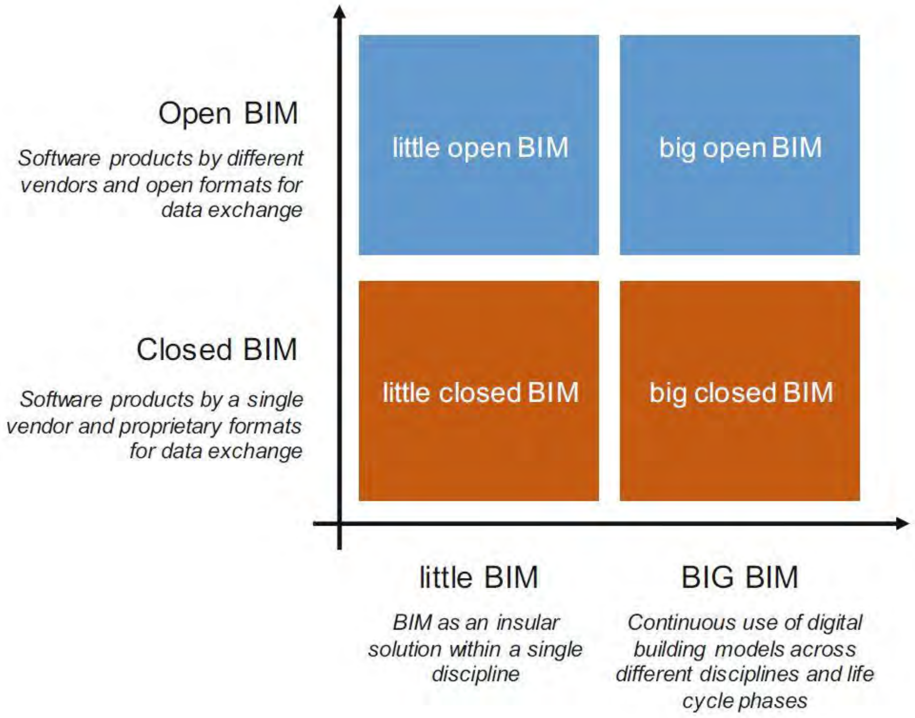

Open BIM
Software products by different vendors and open formats for data exchange
Closed BIM
Software products by a single vendor and proprietary formats for data exchange
little open BIM
big open BIM
little closed BIM
big closed BIM
little BIM
BIM as an insular solution within a single discipline
BIG BIM
Continuous use of digital building models across different disciplines and life cycle phases

Figure 2-3 Options of BIM implementation (Borrmann et al., 2018)

Little BIM describes the application of a specific BIM software by an individual stakeholder to realize a discipline-specific design task. The building model created in this software is not used across different software and is not handed over to other stakeholders, accordingly all external communications take place using derived drawings. This BIM implementation can offer efficiency gains, whereas the potential of comprehensively using digital building information remains untapped (Borrmann et al., 2018). In contrast, big BIM involves consistently model-based communication among all stakeholders and during the entire life cycle of a facility. Regarding the data exchange and the coordination of the workflows, digital technologies such as model servers, databases and project platforms are employed in a comprehensive manner (Borrmann et al., 2018).

When implementing closed BIM, all the employed software products come from just one vendor. Although some software companies provide a large range of software products for the facility design, construction and operation, there will always be a need to exchange data with other products that either serve a specific purpose or are used by other stakeholders in the overall process (Borrmann et al., 2018). In this context, open BIM utilizes open and vendor-neutral data formats to exchange data across products provided by different software vendors (Borrmann et al., 2018), supporting seamless collaboration for all project participants and thereby promoting the interoperability to benefit the project throughout its life cycle (Petrie, 2021). Meanwhile, it extends the breadth and depth of BIM deliverables by creating common alignment and language by adhering to international standards and commonly defined work processes. Furthermore, open BIM facilitates a common data environment that provides opportunities for users to develop new workflows and software applications, as well as improve technology automation (Petrie, 2021).

2.2.4 IFC – Industry Foundation Classes

To this day, several open and vendor-neutral standards have been created to achieve the interoperability and improve the data exchange between different BIM software applications, such as Industry Foundation Classes (IFC) developed by the international non-profit organization buildingSMART* and Construction Operations Building information exchange (COBie) devised by United States Army Corps of Engineers,

* https://www.buildingsmart.org/

allowing all project teams and stakeholders to access to any information at the same time.

Among all the standards, IFC is the most widely recognized and adopted non-proprietary (open BIM) standard for exchanging and sharing BIM information, providing a vendor-neutral file format and an object-oriented data model (buildingSMART International, 2021a). It is registered by ISO as the official international standard ISO 16739-1: 2018, forming the basis for many national guidelines that stipulate the implementation of BIM.

According to ISO 10303-11, the IFC schema is described by the standard data modeling language EXPRESS, which is designed exclusively for defining a data model and can be translated into the markup language XML (Extensible Markup Language) (ISO 10303-11, 2004). Whereas, it is impossible to describe concrete instances of the data model using EXPRESS, instead of which various approaches can be adopted, such as the STEP (STandard for the Exchange of Product model data) Physical File, XML instances or storing data in databases (Borrmann et al., 2018).

The IFC schema characterizes a data model that codifies both geometric and semantic data in a logical manner, to be specific, the identity semantics (name, machine-readable unique identifier, object type, etc.), the characteristics or attributes (material, thermal properties, etc.), relationships (locations, connections, ownership, etc.) of objects (walls, columns, slabs, etc.), abstract concepts (performance, costing, etc.), processes (installation, operations, etc.), and people (owners, designers, suppliers, etc.) (buildingSMART International, 2021a). On this account, the IFC data model is immensely extensive and complex.

In order to improve the maintainability and extensibility, the IFC data model is structured into four conceptual layers (Figure 2-4) with the principle that elements in the upper layer can reference elements in the layers below but not vice versa (Borrmann et al., 2018).

The lowest layer, Resource layer, includes all individual schemas related to resource definitions which do not derive from *IfcRoot*. Therefore, they have no Globally Unique Identifier (GUID) and shall not be used independently of a declaration at a higher layer (buildingSMART International, 2021b). In IFC specification, the GUID is generated for object instances that complies with the Universal Unique Identifier (UUID) standard and its implementation. It is compressed for exchange purpose following a published compression function, which results in a fixed 22-character length string (buildingSMART International, 2021g).

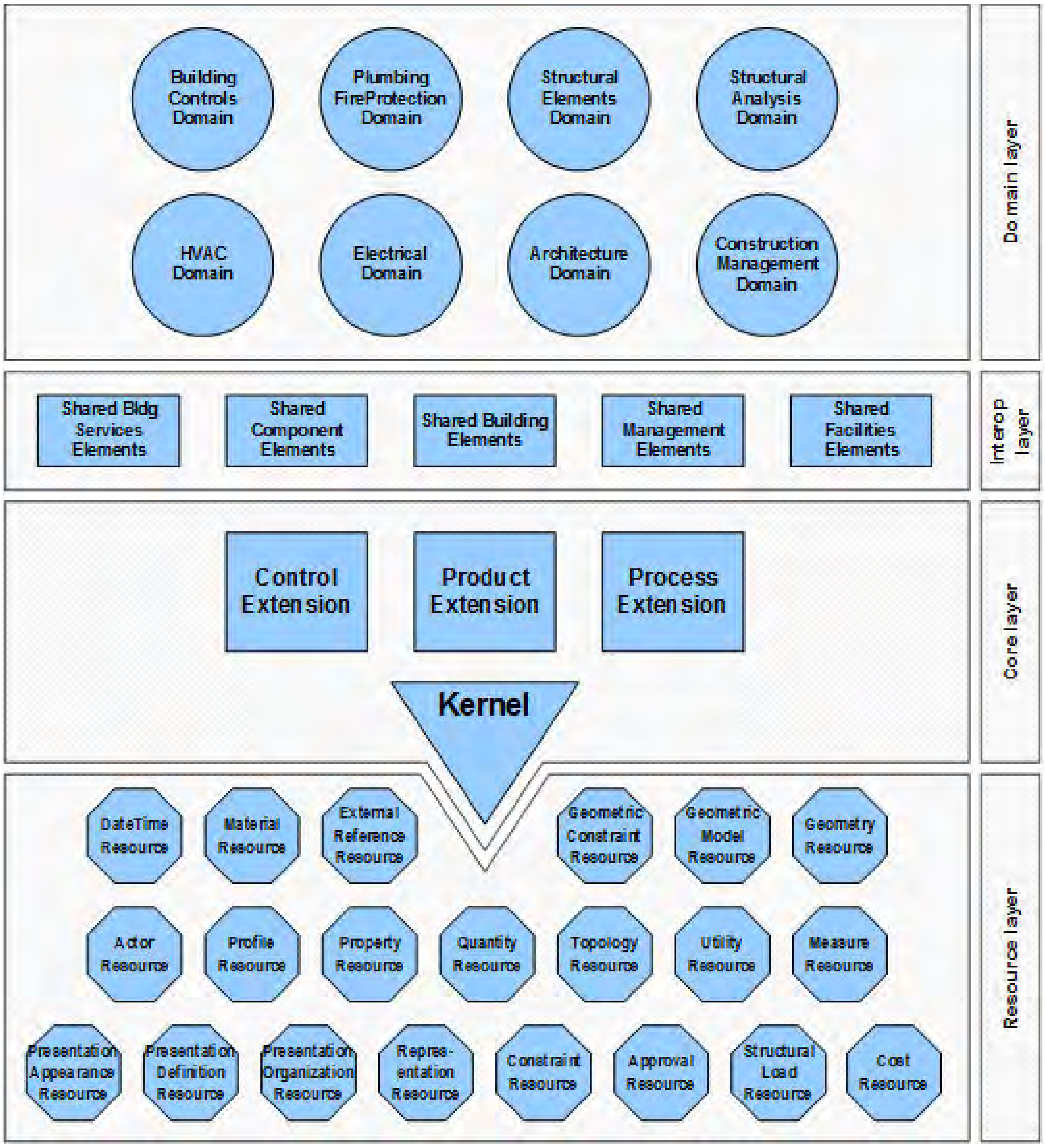


Figure 2-4 IFC data schema architecture with conceptual layers (buildingSMART International, 2021b)

The next layer is the Core layer, in which most general entity definitions are contained as well as all entities are defined, e.g., the basic structures, key relationships and general concepts. The Kernel schema represents the core of the IFC data model and comprises basic abstract classes, e.g., *IfcRoot*, *IfcObject*, *IfcProcess*, *IfcProduct*, *IfcRelationship* (Borrmann et al., 2018).

Above the Core layer lies the Interoperability layer, also regarded as the shared layer (Borrmann et al., 2018), in which the important building element classes (e.g., *IfcWall*, *IfcColumn*, *IfcBeam*) are defined for inter-domain information exchange and sharing across different disciplines.

The highly specialized classes defined in the highest layer, Domain layer, can only be applied to particular domains typically for intra-domain information exchange and sharing, and they form the leaf notes in the inheritance hierarchy tree.

IFC classes are generally described as what they inherit from, which can be divided into two categories: rooted and non-rooted classes (OSArch, 2021). Figure 2-5 illustrates part of the IFC inheritance hierarchy. Similar to any other object-oriented data model, the inheritance hierarchy plays a crucial role in the IFC, defining specialization and generalization relationships, and following a semantic approach – the meaning of objects is the basis for modeling inheritance relationships (Borrmann et al., 2018).

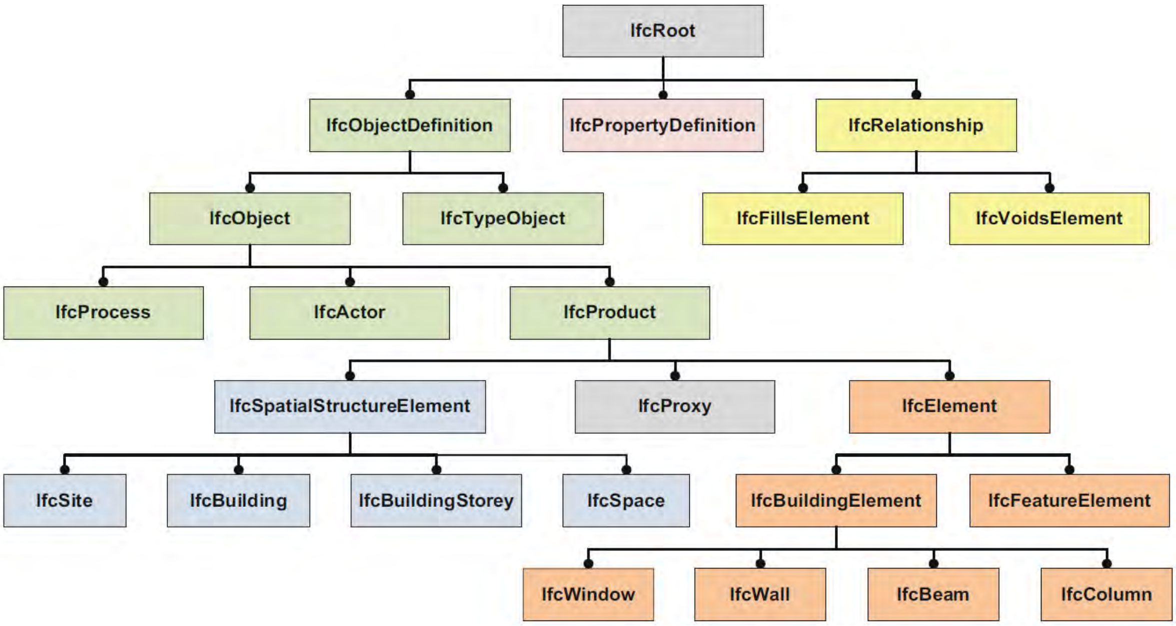


Figure 2-5 Part of the IFC inheritance hierarchy (Borrmann et al, 2018)

The class *IfcBuildingElement* is a generation of all elements that participate in a building system, which are primarily part of the construction of a building, i.e., its structural and space separating system, and are all physically existent and tangible things (buildingSMART International, 2021b). The elaborate hierarchical inheritance of this class is shown in Figure 2-6.

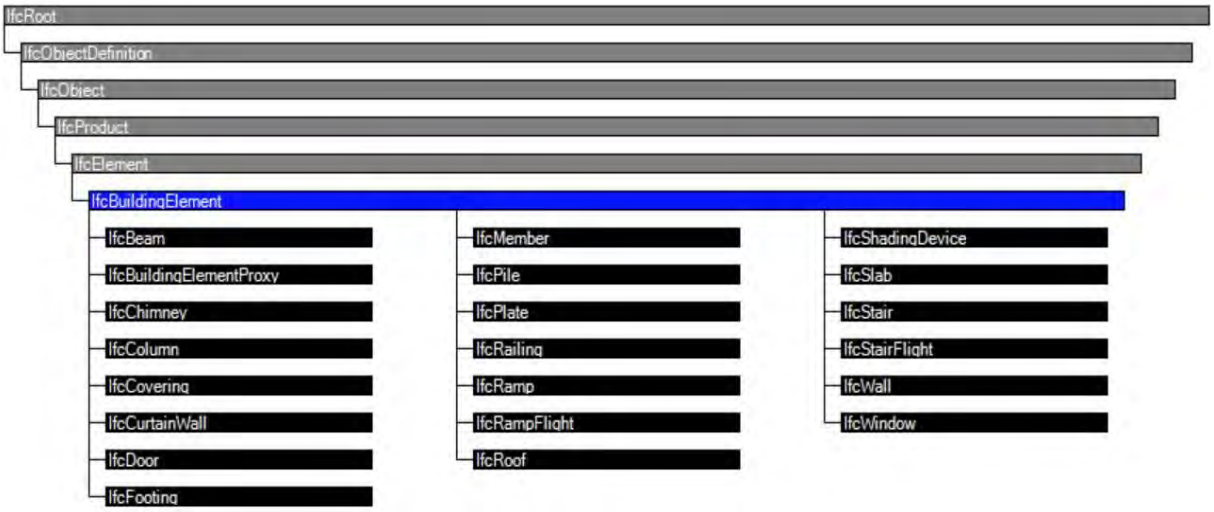


Figure 2-6 Entity inheritance of IFC classes – IfcBuildingElement (buildingSMART International, 2021b)

There are several IFC versions in service, two of the most commonly-used of which are IFC2x3 and IFC4. The IFC Certification is based on the Model View Definition (MVD, which will be introduced in Section 2.2.4). IFC2x3 owns only one certification, namely IFC2x3 Coordination View 2.0 (CV2.0), while the IFC4 Certification has been split into two more specific view definitions to better support the purpose of IFC data exchange (buildingSMART International, 2019). The first is Design Transfer View, which is aiming at supporting the transfer of model data to be used for further design, analysis, estimation and facility management tasks. It can be understood as a CV2.0 with some extended range. The second is Reference View, which aims to support the architecture, structural analysis and building services, the coordination of the planning disciplines, especially the clash detection and resolution of issues resulting from geometry (buildingSMART International, 2019).

Compare to IFC2x3, IFC4 is an advanced schema that has extended and overcomes certain limitations of IFC2x3 such as it attains the supports of mvdXML and BIM Collaboration Format (BCF, which will be introduced in 2.2.4) (buildingSMART International, 2021f). The former enables the MVDs machine-readable, as well as the checking tool suitable and the platform automatically configured for multiple MVDs driven by mvdXML. The latter can be used via API to integrate issue management during auditing into BIM applications, and can cooperate with mvdXML for auto-configuration of IFC-interfacesAPI for access to checking tools (buildingSMART International, 2021f). The detailed differences between the two IFC versions are listed in Table 2-3.

Table 2-3 Comparison of IFC2x3 and IFC4 (buildingSMART International, 2021e)

	IFC2x3	IFC4
Certification	IFC2x3 Coordination View 2.0	IFC4 Design Transfer View IFC4 Reference View
Differences	- Platform based on internet hosted Oracle database - Checking tool only suitable for one MVD (not machine-readable)	- Platform based on Azure Cloud and therefore scalable - Better UI - Support of mvdXML and BCF - Improved test cases

In conclusion, with the rich schema and the full range of data storage, IFC is currently used much more beyond exchanging information between software products. It is also employed as a means to archive project information, whether in the design, procurement, and construction phases, or as a collection of "as-built" information for long-term preservation or operation purposes (buildingSMART International, 2021a). On one hand, as a digital data format, it can avoid the need to manually re-enter the data or information that has been already created. On the other hand, it can reduce the accompanying risk of manual work errors.

IFC schema forms the basis for data exchange and interoperability in the modern AEC industry, which are exactly the challenges of the BIM-LCA integration. Therefore, numerous BIM-LCA studies have put emphasis on the IFC schema, which will be introduced in section 2.4.3.

2.2.5 IDM, MVD and BCF

Information Delivery Manual (IDM), Model View Definition (MVD), BIM Collaboration Format (BCF) were created to supplement the IFC schema to form a comprehensive and complete BIM system (Borrmann et al., 2018).

The first is a methodology to capture and specify processes and information flow during the lifecycle of a facility (buildingSMART International, 2021c), i.e., IDM regulates which information is delivered by whom, when and to which recipient (Borrmann et al., 2018). Today, it is recorded as an ISO standard (ISO 29481) and can be used to document existing or new processes and describe the associated information that have to be exchanged between parties (buildingSMART International, 2021c).

The second is a selection of entities of the entire IFC schema that defines which parts of the IFC data model need to be implemented for a specific data exchange scenario (Borrmann et al., 2018). MVDs can be as broad as nearly the entire schema (e.g., for archiving a project) or as specific as a couple object types and associated data (e.g., for pricing a curtain wall system) (buildingSMART International, 2021d). Moreover, it can add additional restrictions to the IFC schema, and even overrule some agreements (buildingSMART International, 2021d).

The last one BCF allows different BIM applications to communicate model-based issues with each other by leveraging IFC models that have been previously shared among project collaborators (buildingSMART International, 2021e). There are two different ways to utilize BCF – via a file-based exchange or via a web service. More specifically, BCF works by transferring XML formatted data, which is contextualized information about an issue or problem directly referencing a view, captured via PNG and IFC coordinates, and elements of a BIM, as referenced via their IFC GUIDs, from one application to another (buildingSMART International, 2021e).

2.3 NLP – Natural Language Processing

Natural language (NL) represents the language which is used by human and evolves over time with continuous usage and repetition (Kumar, 2017), the scope of which contains text, speech, cognition, and their interactions (Yalçın, 2020). Concerned with the interaction between humans and machines, Natural Language Processing (NLP) is an interdisciplinary subfield which has components from fields of linguistics, CS, AI, and so on (Yalçın, 2020). Nowadays, it has been heavily utilized in a variety of applications such as chatbot, speech recognition and generation, language translation, spam detection, question responses, and sentiment analysis (Rebala, Ravi & Churiwala, 2019; Singh, 2018).

2.3.1 NLP Steps and Tasks

NLP deals with how machines or computers process NL commands, whether written or verbal, and then engage with any action as requested (Kumar, 2017). From a machine learning view, an NLP analysis commonly comprises five major steps: (1) reading the corpus, (2) tokenization, (3) cleaning/ stopwords removal, (4) stemming/ lemmatization, (5) converting into numerical form/ word embedding (Singh, 2018).

To be more specific, a corpus is regarded as the entire collection of text documents (Singh, 2018). Tokenization is known as the method of dividing the given sentence or collection of words of a text document into separate or individual words, which removes the unnecessary characters such as punctuation (Singh, 2018). In spite of this, the tokens column still contains a certain amount of stopwords such as “this”, “the” and “to”, which largely increase the computation complexity without adding much value (Singh, 2018). Therefore, a cleaning step is recommended to remove these meaningless words from the tokens. After dropping the stopwords, the inflectional forms and derivationally related forms of words need to be converted to common base forms through the stemming or lemmatization step (Manning, Raghavan & Schütze, 2018).

The last major step word embedding is one of the key breakthroughs of deep learning for solving NLP problems (Kanani, 2019), which aims to transform words into vectors of numbers that preserve a form of syntactic and semantic relationships between words (Rebala, Ravi & Churiwala, 2019). Conceptually, it involves a mathematical embedding which transforms sparse vector representations of words into a dense, continuous vector space. The outcomes, word vectors (also called word embeddings), are basically a type of word representation that allows words with similar meaning to have similar representation (Kanani, 2019). One of the most popular word embedding tools is Word2Vec, which is a two-layer neural net that processes text by vectorizing words (Nicholson, 2020).

With the general steps hereinabove, NLP is capable of coping with tasks including identifying sentence boundaries in documents, extracting relationships from documents, and searching and retrieving of documents among others (Akerkar, 2018). Moreover, basic NLP tasks also involve tokenization and parsing, lemmatization and stemming, part-of-speech tagging, language detection and identification of semantic relationships (Akerkar, 2018).

2.3.2 NLP Tools

This section will outline and compare five of the most commonly employed NLP tools.

1. NLTK[1]

Natural Language Toolkit (NLTK) is a free, open-source and community-driven platform with comprehensive Application Programming Interface (API) documentation for building Python programs to work with human language data (NLTK Project, 2021). It provides interfaces to more than 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, and wrappers for industrial-strength NLP libraries (NLTK Project, 2021).

NLTK is powerful for simple text analysis. However, it is inefficient when it comes to working on a massive amount of data on account that it requires significant resources (Bushkovskyi, 2019). Additionally, NLTK was initially created for academic use and it mostly supports the English language so that it offers no pre-trained German language model.

2. Gensim[2]

Gensim is an open-source and platform-independent machine learning library, which can process arbitrarily large corpora using data-streamed algorithms without "dataset

[1] https://www.nltk.org/#
[2] https://radimrehurek.com/gensim/#

must fit in RAM” limitations (Řehůřek, 2021). The key feature of Gensim is word vectors, which converts the content of the documents into the sequences of vectors and clusters, and then classifies them. Therefore, it is probably applied to perform word-to-vector and sentence-to-vector tasks (Bushkovskyi, 2019). Gensim does not require costly annotations or hand tagging of documents because it uses unsupervised models (Kharkovyna, 2021). Hence, it is known as the fastest library for training of vector embeddings in Python. Furthermore, the Gensim community also publishes pre-trained models for specific domains like legal or health (Řehůřek, 2021).

3. CoreNLP[3] and Stanza[4]

CoreNLP is a one-stop-shop for NLP in Java, whose centerpiece is the pipeline. Pipelines take in raw text, run a series of NLP annotators on the text, and produce a final set of annotations, CoreDocuments and data objects (Stanford NLP Group, 2020a). The annotations include token and sentence boundaries, parts of speech, named entities, numeric and time values, dependency and constituency parses, coreference, sentiment, quote attributions, and relations (Stanford NLP Group, 2020a). The CoreDocuments contain all of the annotation information, accessible with a simple API (Stanford NLP Group, 2020a). With its high scalability, CoreNLP supports information scraping from open sources, sentiment analysis, conversational interfaces, and text generation (Bushkovskyi, 2019). CoreNLP is a good choice for processing large amounts of data and performing complex operations (Bushkovskyi, 2019). Currently, it supports eight languages, namely Arabic, Chinese, English, French, German, Hungarian, Italian, and Spanish (Stanford NLP Group, 2020a).

Although CoreNLP is based on Java, its creators developed an alternative for Python with the same functionality, called Stanza, which provides pre-trained NLP models for a total of 66 human languages (Stanford NLP Group, 2020b). It is a collection of accurate and efficient tools, which can be used in a pipeline, to convert a string of human language text into lists of sentences and words, to generate base forms of those words, their parts of speech and morphological features, to give a syntactic structure dependency parse, and to recognize named entities (Stanford NLP Group, 2020b).

[3] https://stanfordnlp.github.io/CoreNLP/
[4] https://stanfordnlp.github.io/stanza/

4. spaCy[5]

Being a free and open-source library for advanced NLP in Python, spaCy is designed particularly for production use and helps build applications that process and analyze large volumes of text (Explosion, 2021). It provides a one-stop-shop for tasks commonly involved in any NLP project, incorporating tokenization, lemmatization, part-of-speech tagging, entity and sentence recognition, dependency parsing, word-to-vector transformation, and text cleaning and normalization (Mc., 2017). As of now, spaCy has released 64 trained pipelines for 19 languages (Explosion, 2021).

Unlike NLTK or CoreNLP, spaCy is not a platform or an API since it does not provide software or web applications as service. In addition, notwithstanding it can be used to power conversational applications, it is not designed specifically for chatbots, but only provides the underlying text processing capabilities (Explosion, 2021).

5. BERT

Since 2017, Transformers have taken NLP by storm, providing enhanced parallelization and better modeling of long-range dependencies (Rogers, Kovaleva & Rumshisky, 2020). Developed by Google, BERT stands for Bidirectional Encoder Representations from Transformers (Devlin et al., 2019), which is the most well-known Transformer-based model.

Different from the above four tools, BERT is a language representation model instead of a platform or library. It is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers (Devlin et al., 2019). The conventional workflow for BERT consists of two stages: pre-training and fine-tuning (Rogers, Kovaleva & Rumshisky, 2020). Pretraining involves two self-supervised tasks: masked language modeling (MLM, prediction of randomly masked input tokens) and next sentence prediction (NSP, predicting if two input sentences are adjacent to each other). In fine-tuning for downstream applications, one or more fully connected layers are typically added on top of the final encoder layer (Rogers, Kovaleva & Rumshisky, 2020).

The pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of NLP tasks, e.g., question answering,

[5] https://spacy.io/

language inference and sentiment analysis, without substantial task-specific architecture modifications (Devlin et al., 2019).

2.3.3 Text Similarity Metrics in NLP

The text similarity denotes how two text documents close to each other in terms of their context and meaning (Kanani, 2020b). There are various text similarity metrics pervasively applied in NLP, e.g., Jaccard similarity, Euclidean distance, and cosine similarity, all of which possess their own specification, and the outcomes of which range from 0 to 1 with 0 stands for the lowest similarity while 1 indicates the largest (Kanani, 2020b).

1. Jaccard Similarity

Jaccard similarity is also known as Jaccard index and Intersection over Union (Formula 2.1), which matches documents by means of counting the maximum number of identical words between the documents (Prabhakaran, 2018; Kanani, 2020a).

$$Jaccard\ similarity = J(doc_1, doc_2) = \frac{doc_1 \cap doc_2}{doc_1 \cup doc_2} \tag{2.1}$$

However, this approach has an inherent flaw that with the increase of the document size, the number of common words tend to simultaneously increase even if the documents talk about different topics (Prabhakaran, 2018).

2. Euclidean Distance

Euclidean distance is one of the most well-known formulas for computing the distance between two points based on the Pythagorean theorem (Neto, 2021). To calculate it, at first the points need to be subtracted from the vectors, raised to squares, and then added up and taken the square root, as shown in Formula 2.2 (Neto, 2021).

$$\begin{aligned} Euclidean\ distance = d(p,q) = d(q,p) &= \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} \\ &= \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2} \end{aligned} \tag{2.2}$$

Being applied to NLP, the Euclidean distance is regarded as the most intuitive text similarity metric (Briggs, 2021). Whereas, it might be imprecise when it comes to documents with large size (Prabhakaran, 2018).

3. Cosine Similarity

Cosine similarity is a metric to measure the text similarity between two documents irrespective of their sizes. Mathematically, it calculates the cosine of the angle between two n-dimensional vectors projected in a multi-dimensional space by taking the dot product between the vectors and dividing it by the multiplication of the vector norms (Prabhakaran, 2018; Neto, 2021), the mathematical formula of which is:

$$cosine\ similarity = \cos(\theta) = \frac{A \cdot B}{||A||\ ||B||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\ \sqrt{\sum_{i=1}^{n} B_i^2}} \tag{2.3}$$

Where in NLP A and B represent the word embeddings of two documents, and cannot be empty.

Compare to the first two metrics, the last one cosine similarity is advantages in that its precision will not be significantly affected regardless the document size. In other words, even though the two being compared documents are reflected to be far apart by the Jaccard similarity or Euclidean distance due to the size of the document, the chance still exists that they are oriented closer together, or vice versa (Prabhakaran, 2018). On this account, most of current NLP tools utilize the cosine similarity.

2.3.4 NLP Applied in the AEC Industry

According to the SLR made by Di Giuda et al. (Di Giuda, Locatelli & Seghezzi, 2020), NLP has been applied in the AEC industry mainly in cooperation with BIM technologies to cope with issues of data processing, automation level improving, etc. Additionally, Wu et al. pointed out in their state-of-the-art SLR that BIM combined with NLP is increasingly exploited in smart construction such as information extraction and exchange, multi-model information integration, and achieving accuracy-efficiency trade-off (Wu et al., 2022).

Furthermore, Yalcinkaya and Singh revealed twelve research themes that indicate the patterns and trends in research of introducing Latent Semantic Analysis (an NLP technique) into BIM, e.g., information exchange and interoperability, safety management, urban/ building space design and analysis, design codes and code compliance, maintaining and managing facilities (Yalcinkaya & Singh, 2015; Pauwels, Zhang & Lee, 2017). Moreover, NLP is growingly appearing in the domain of construction engineering to improve construction safety, which intends to retrieve important information from the safety reports and make content analysis for better interpretability and less ambiguity (Pan & Zhang, 2021).

Jung and Lee comparatively analyzed a method to automatically classify case studies of BIM in construction projects by BIM use (Jung & Lee, 2019). To automate and expedite the analysis tasks of the study, they deployed NLP and commonly-used unsupervised learning for text classification, namely latent semantic analysis and latent Dirichlet allocation.

In building design codes and regulations, some progress has been achieved in the application of NLP, but it has not been realized in commercial use (Sacks, Girolami & Brilakis, 2020). Zhang and El-Gohary proposed a unified automated compliance checking system that integrates semantic NLP techniques and EXPRESS data-based techniques to automatically extract and transform both regulatory and design information (in BIM) for automated compliance reasoning (Zhang & El-Gohary, 2017). Similarly, Song et al. put forward an NLP and deep learning-based approach for supporting automated rule checking system, which describes a semantic analysis process of regulatory sentences and its utilization for the system (Song et al., 2018).

2.4 BIM-based LCA and BIM-LCA Integration

Relying on BIM technologies, the limitation of LCA/ LCEA/ LCC that is difficult to implement in early design stages can be gradually overcome. Thus, over the past decade, BIM incorporated with sustainability and LCA (i.e., 6D BIM and 7D BIM) has attracted more and more attention, simultaneously a lot of studies have been carried out. Accordingly, the number of relevant publications has seen a concomitant rise with a significant yearly increase since 2014 (Potrč Obrecht et al., 2020; Liu et al., 2021). Thereinto, studies of applying LCA in early design stages accounts for exceeding 50% of the BIM-LCA publications (Liu et al., 2021).

Safari and AzariJafari pointed out that there is an extensive area of opportunity for integrating BIM and LCA (Safari & AzariJafari, 2021), such as creating local LCA databases following the LOD models, and developing a comprehensive framework for assessing different sources of uncertainties. Seyis identified and classified 21 advantages and 7 disadvantages associated with BIM-based LCA (Seyis, 2020). The disadvantages – also opportunities from a different angle – can be grouped into two categories: standardization and data processing. In this context, the latest development of the data processing concentrated on BIM/ LCA data mapping methods will be presented in Section 2.4.3 and Chapter 2.5.

2.4.1 BIM-LCA Integration Types and Their Pros and Cons

Wastiels and Decuypere divided BIM-LCA integration into five types, as shown in Figure 2-7, which is acknowledged as the most comprehensive classification of BIM-LCA integration strategies to date (Potrč Obrecht et al., 2020).

The kernel of the first type is exporting the BOQ, which reflects the inventory of the building materials that a conventional LCA always starts with. The BOQ is at first exported from the BIM environment and then directly imported in a dedicated LCA software, where the LCA calculation, visualization and analysis are implemented. According to their investigation, this workflow is the most prevailingly performed one in current BIM-LCA integration practice. However, this strategy relies heavily on manual works such as linking the building components along with their quantities to the predefined LCA profiles in the LCA software, which is laborious and might not support the iterative design (Wastiels & Decuypere, 2019).

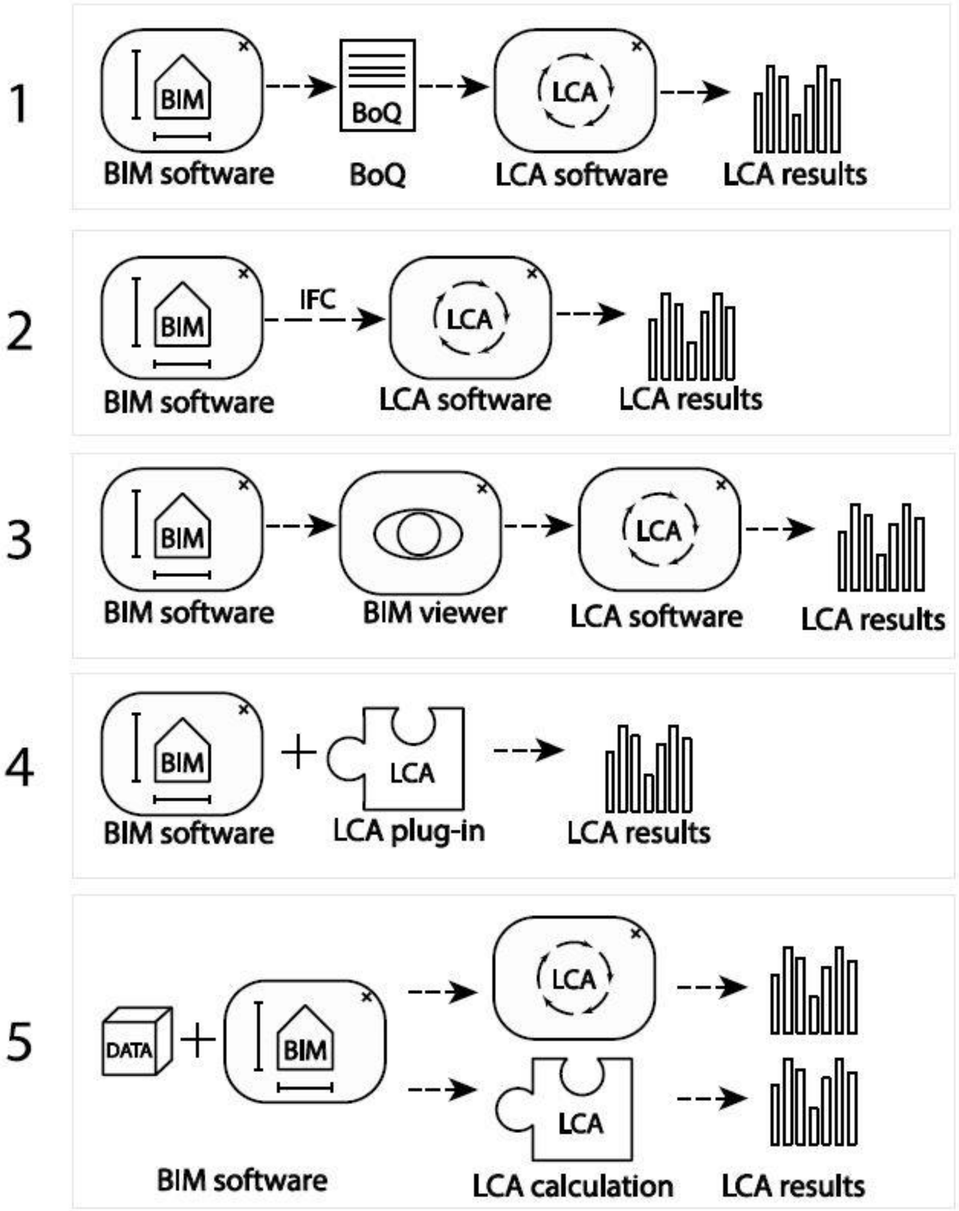


Figure 2-7 BIM-LCA integration types (Wastiels & Decuypere, 2019; Potrč Obrecht et al., 2020)

Compared with the first integration type, the second approach automates the data export and import by introducing the open exchange format IFC to replace the BOQ spreadsheet, based on which the material quantities, e.g., surfaces and volumes, can be determined. Furthermore, when the GUID contained in IFC file is imported and stored in the LCA software, the iterative design could be supported, since the quantities and descriptions can be updated through a new version of the IFC file without losing the existing links to the LCA profiles (Wastiels & Decuypere, 2019).

In the third strategy, the LCA profiles are attributed in an intermediate step in a BIM Viewer between the export of IFC file and the LCA software. One of the advantages of this method is isolating the LCA profiles in a 3D environment while keeping the in-depth

LCA analysis in a dedicated LCA environment. Besides, the link between the IFC data and the LCA profiles can be maintained for further reference during an iterative optimization process (Wastiels & Decuypere, 2019).

The fourth type employs exclusively developed plug-ins that enable all LCA analyses within the BIM tools (Potrč Obrecht et al., 2020). Comparing to the previous strategies, an advantage of this approach is that the LCA results can be directly visualized in the geometric model created in the native BIM environment. Whereas, the LCA analyses can only be carried out on the basis of the LCA databases that are built-in in the plug-in (Wastiels & Decuypere, 2019). At present, a number of LCA instruments are in service working with plug-ins for BIM software such as Tally[1] and One Click LCA[2] for Autodesk Revit.

In the fifth approach, the LCA information is included in the BIM objects that are used in the BIM model instead of being attributed to the appropriate building components in a later stage or in separate tools (Wastiels & Decuypere, 2019). A great advantage is that the information regarding the environmental impacts can be analyzed along with the project development (Potrč Obrecht et al., 2020). Nevertheless, this strategy has not been efficiently realized yet due to lacking available BIM objects with LCA data and consensus on the way to structure LCA data and profiles (Wastiels & Decuypere, 2019).

Among the five BIM-LCA integration strategies, the first can be seen as a manual approach on account of the manual obtained BOQ spreadsheet and the manual linking work. The second and the third can be categorized as semi-automated BIM-based LCA methods in that they still need manual links between the IFC file and the LCA profiles. The four and five can be regarded as automated approaches because they have integrated LCA information in the BIM tools through plug-ins or as built-in data, with which the laborious manual mapping work can be avoided.

To better transversely compare the three kinds of integration methods (manual, semi-automated, and automated), the strengths and weaknesses of which are listed in the following table (Table 2-5) (Wastiels & Decuypere, 2019).

[1] https://apps.autodesk.com/RVT/en/Detail/Index?id=3841858388457011756&appLang=en&os=Win64
[2] https://apps.autodesk.com/RVT/en/Detail/Index?id=3065869958781255107&appLang=en&os=Win64

Table 2-4 Pros and cons of manual, semi-manual and automated BIM-based LCA methods

	Manual Method	Semi-automated Method	Automated Method
Pros	- The most conventional and commonly-used approach currently - The most accurate	- Manually obtaining the BOQ is avoided by introducing IFC format or BIM viewer tools between BIM software and LCA software - Being able to support iterative design	- The fastest approach with the highest automation level - Being able to support iterative design - Being able to instantly visualizing the LCA results in the geometric model
Cons	- Laborious and time-consuming manual work (linking building components/ IFC data to predefined LCA profiles) - might not support iterative design	- Manual linking/ mapping work still needed - not accurate enough in terms of QTO	- The least frequently-used approach currently - The least accurate due to the limited volume of LCA database or the lack of LCA data

Notwithstanding most BIM-LCA integration research so far is still focused on manual and semi-automatic solutions (Safari & AzariJafari, 2021), and there are deficiencies in the current implementation of automated methods, the full automation of BIM-based LCA is the future direction.

2.4.2 BIM-based LCA for Decision-making

For the past few years, BIM-based LCA and environmental impact analysis to support decision-making of design options or material combinations has become a hot research theme, several key studies of which will be introduced in this section.

Figueiredo et al. advanced a decision-making framework for construction professionals and researchers involving the integration of Life Cycle Sustainability Assessment (LCSA), Multi-Criteria Decision Analysis (MCDA), and BIM to choose suitable materials for buildings (Figueiredo et al., 2021). Whereas, it covers the construction, operation and end-of-life phases of the building, ignoring the early design phase.

Forth et al. developed a framework for optimizing building designs taking BIM as a main source of truth for a multicriteria analysis, which compares multicriterial variants to support decision-making during the early design stages, including the selection process and feedback communications of design changes (Forth et al., 2021). For a holistic variant comparison, they considered several criteria with the main focus on LCA (Forth et al., 2021).

Schneider-Marin et al. introduced a BIM-based method to analyze the contribution of the main functional parts of buildings to discover potentials of embedded energy demand and GHG emission reduction (Schneider-Marin et al., 2020). They conducted a sensitivity analysis to exhibit the variance in results due to the early inherent uncertainties, indicating where to strategically reduce uncertainties to improve the precision, and considerably avoiding misleadingly outcomes. Although their study brought sensitivity analysis in the early design process, and showed how a simplified and fast BIM-based LCA calculation provides valuable guidance, a comprehensive framework for evaluating different sources of uncertainties has not been developed.

Tushar et al. (Tushar et al., 2021) were the first to employ BIM-based LCA method combined with energy rating tool to quantify, compare and improve the building design options in terms of passive design strategies, such as orientation, shading, sealing, glazing and insulation, to reduce carbon footprint and energy consumptions in residential dwellings. In their research, a sensitivity analysis was performed to optimize the operational energy efficiency, and finally an evidence-based analytical framework was proposed for providing a BIM-based optimization platform to validate and justify the impact categories of environmental-friendly and energy-efficient design of buildings.

To support decision-making in the early design phase, Röck et al. exploited an BIM-based approach to assess a wide range of construction options and their embodied environmental impact (Röck et al., 2018), enabling to identify design-specific hotspots and promote the transparency of communication by visualizing them on the building model. In their research, an automated link between the aggregated LCA database (MS Excel) and the BIM model (Autodesk Revit) was established via a custom script developed through the visual scripting software (Autodesk Dynamo).

2.4.3 BIM-based LCA/ LCC Data Exchange and Processing via IFC Format

Despite much progress in recent years, BIM-based LCA data exchange via the IFC format still does not work perfectly, which concretely manifests in the time-to-time occurrence of data loss and misinterpretation (Borrmann et al., 2018; Safari & AzariJafari, 2021). To work out this problem, there is a growing trend towards the automatic compilation of data exchange, including IFC data automatic recognition, extraction and mapping to LCA databases.

Lawrence et al. proposed a generic method to create and maintain the LCC calculation using flexible mappings between the building model and the cost estimation (Lawrence et al., 2014). With the flexibility of modern query languages, this method is in favor of allowing estimator to encode a broad variety of relationships between the design and estimation.

In a study conducted by Kim et al. (Kim et al., 2016), the information stored in the IFC file was first extracted and grouped into different element sets (ObjectPlacement, LayerSet, material layer, etc.), and proceeded to match the property data in the user defined library, then mapped to building energy analysis model. Although the mapping process was complemented with manually extending material information, their study has inspired the development of automated mapping method in some sense.

Considering that the lack of semantic information within BIM models can lead to ineffective decision-making processes, Santos et al. explored the potential of BIM as a data repository for LCA and LCC information and its capacity for supporting the automatic/ semi-automatic environmental and economic assessment (Santos et al., 2019). As the results of their research, a BIM-LCA/ LCC analysis framework was developed, as well as the compatibility of IFC4 for the proposal of an IDM and an MVD was verified.

Horn et al. came up with an approach for integrating LCA in all phases of digital planning based on a single IFC data format (Horn et al., 2020), having taken in to account IDM, MVD, and varying levels of development and resulting data availability during integral planning phases, as well as resulting LCA application context. They concluded that an open BIM approach for LCA integration in model-based design is feasible, but requires a few adjustments in IFC, LCA, and planning practice.

Theißen et al. proposed a BIM-LCA framework that forms the basis for BIM semi-automation of the whole building LCA by enabling the integration of LCA data using the IFC (Theißen et al., 2020). Their project includes the development of an IDM in the line with a German assessment system for sustainable buildings (BNB), based on which, a MVD is also developed to define the software subset of the IFC data model to meet the exchange requirements for the whole building LCA. Furthermore, they presented the solution approaches for adapting LCA databases and IFC4. The DIN EN 15804 compliant and open access Ökobaudat is used as the LCA database.

Khosakitchalert et al. put forward a method comprising of five processes (Khosakitchalert, Yabuki & Fukuda, 2020), including extracting material information from the original model, converting, recreating, joining and eliminating the components (walls, floors, etc.). This so-called automatic compound element modification (ACEM) method improves the accuracy of the incompletely or incorrectly extracted quantities of compound elements from BIM models by taking information from IFC-based clash detection to eliminate excess quantities and add missing quantities. It effectively saves the time of editing the BIM model and contributes a lot to the QTO of LCA/ LCC calculations. Analogously, Bernardino-Galeana et al. presented a structured proposal for the accomplishment of LCC with IFC based on the ISO 15686 standard (Bernardino-Galeana et al., 2021), expected to promote the LCC analysis in early design stages to achieve more economically efficient buildings.

To sum up, more and more scientists and researchers have attached importance to the data recognition, extraction and mapping problems in the fields of BIM and LCA, but a widely acknowledged framework has not yet been established. There is still a long way to go to achieve fully automated data processing.

2.5 BIM (IFC) Data Extraction and Mapping Methods

The emergence of BIM relies on the development of modern computer science and technologies, and there is still great potential of optimizing BIM with CS technologies. For this purpose, a lot of studies investigate the probable applications of CS technologies (NLP, neural networks, etc.) in the BIM data processing. In some cases, CS technologies are employed to tackle the IFC data extraction and mapping problem, which are regarded as complex tasks, involving dealing with different data structures and semantics that have to be aligned.

Dankers et al. demonstrated an approach of linking the machine to human readable data in their paper, allowing non-CAD users or architectural experts to access the data (Dankers, Van Geel & Segers, 2014). A web-platform for integrating model-based and non-model-based data is the core part of the approach, where a mapping process was constructed from IFC properties to a national building element classification system as a way of structuring the objects and information. However, they translated the semantic information by means of a basic reasoning system rather than NLP, which is not accurate enough.

Wu et al. put forward a NL-based intelligent retrieval engine for the BIM object database and Revit modeling, which outperforms traditional keyword-based retrieval instruments such as Autodesk Seek and BIMobject (Wu et al., 2019). Their research was comprised of four main steps – constructing a domain ontology for semantic understanding and establishing a BIM object database framework for testing the engine, extracting information from the natural sentences of users through NLP, forming a final query in light of “keyword” and “restriction sequence”, and presenting and ranking the results through the mapping from the final query to the BIM object database.

Ismail et al. dealt with the problem of building related data extraction and processing (Ismail, Strug & Ślusarczyk, 2018). They retrieved information from the IFC files, then used the tool IfcWebServer to transform the information into the graph model, and stored the model in a graph database which allows for specific graph queries. However, the scope of transformation and queries does not consider all the geometry information or the process of creating geometry objects, which could be addressed by the next step of their research – developing an interface between the graph database and the IFC geometry engine.

Costa and Sicilia analyzed the potential of using semantic web query languages to promote automatic transformation of BIM data (Costa & Sicilia, 2020), having identified fourteen data mapping patterns (1 to 1, 1 to N, N to 1, N to N on class/ attribute level, etc.) and three data transformation cases with consideration of the semantic and structural differences.

Barzegar et al. designed an IFC-based database schema in consideration of spatial analysis requirements, and provided a methodology to convert BIM data into this schema, which comprises of seven steps, namely designing the architectural model and adding legal data, georeferencing, IFC data validation and cleaning, mapping process, database data validation and cleaning, spatial analysis, and visualization (Barzegar et al., 2021). Their research demonstrated the feasibility of the proposed spatial database, and greatly facilitated spatial analyses required by different stakeholders.

Koo et al. explored the feasibility of applying 3D geometric deep neural networks (DNN) in IFC data extraction, which confirms DNN as a viable solution to distinguishing BIM element subtypes (Koo., Jung & Yu, 2021). To be more specific, they compared the applicability of multi-view convolutional neural networks (MVCNN) and PointNet in extracting unique features of IfcDoor and IfcWall element subtypes. The test results indicated MVCNN as having the better prediction performance, while PointNet's accuracy was hampered by resolution loss due to selective use of point cloud data.

Even though LCA is not involved in the technological achievements or studies above, what they have figured out is indeed of great reference value and has great potential for filling up the gaps in implementing and improving BIM-based LCA from numerous sides, particularly in the realization of entirely automated data exchange.

3 An Automated Method of Mapping LCA Data to BIM Models

3.1 Research Questions

Chapter 2 has presented the state-of-the-art development and employment of LCA, BIM, and NLP in the AEC industry, and analyzed the current BIM-LCA integration approaches as well as the relevant research and practical applications.

According to the advantages and disadvantages analysis of the manual, semi-automated, and automated BIM-LCA integration strategies, the automated approach is discovered to be much faster than the other two approaches and is regarded as the promising direction. It is able to eliminate the laborious and time-consuming manual work, e.g., linking the BOQ to the existing predefined LCA profiles, or mapping the LCA-related IFC data to the LCA database. However, this method does have deficiencies, such as the lack of datasets due to the volume of the BIM software built-in LCA database, and the low accuracy of the BIM-to-LCA data mapping due to the time-to-time data loss or misinterpretation. Therefore, it is the least frequently adopted method currently.

Compared with closed BIM, open BIM uses open and vendor-neutral data formats to satisfy the demand for exchanging and sharing data across various software applications and among all project participants during the life cycle of a facility (Borrmann et al., 2018; Petrie, 2021). In the meantime, it promotes a common data environment that provides opportunities to enhance the level of automation (Petrie, 2021). Thereby, for the purpose of improving the automated BIM-based LCA method, the BIM-LCA data exchange via the IFC format is a great starting point, as it is the most prevailing open BIM standard, and most BIM software on the market supports the export of IFC files. If the LCA-related information stored in the IFC files can be automatically and accurately mapped to the LCA/ LCI databases, data loss and manual work in the BIM-LCA data exchange process are supposed to be reduced and avoided to a large extent.

Furthermore, the technology NLP is known as an advanced and powerful instrument to coping with semantic tasks, which is increasingly applied to the AEC industry to process semantic data in the BIM-LCA integration (Di Giuda, Locatelli & Seghezzi, 2020).

In this context, the research of this thesis aims to explore the potential of employing NLP technique to improve the automated BIM-LCA integration strategy in terms of the data mapping process. To accomplish this objective, the following two research questions are proposed to be addressed:

- Can NLP help with the automated method of mapping LCA data to BIM models?
- Is the automated mapping result efficient and accurate enough to support the downstream LCA analyses?

3.2 Methodology

In order to resolve the two research questions, an automated method of mapping LCA data to BIM models was developed, with introducing the technology NLP, the methodology of which is depicted in Figure 3-1.

First of all, the BIM model is created in the BIM software, from which the IFC file can be exported. There are a large number of BIM software that supports adding and editing building component or material information, as well as exporting IFC files with different versions of IFC certification, e.g., Autodesk Revit[1], Graphisoft ArchiCAD[2], and Allplan[3].

Afterwards, the LCA related data is extracted from the exported IFC data model. As presented in Section 2.2.4, the IFC model is immensely extensive in data volume and data types, and very complex in data structure. Nevertheless, not all information contained in the IFC file is requisite and valuable for an LCA study, since it involves not only project site, building element, etc., but also ownership, owner history, etc. On this account, this step is necessary for distinguishing the IFC data which is relevant to the LCA domain from other non-LCA related data.

Then, the objects to be mapped are determined and selected from the extracted LCA data, which describe the most characteristic attributes of the building elements, e.g., material/ component names, material categories, and material properties. At the same time, the corresponding data is selected from the LCA database according to the identical mapping objects.

In LCA research, there are generally two ways to query datasets in LCA databases, i.e., generally LCA databases have two service modes. One is standalone, and the other is built into the LCA tools such as LCA software, web applications, and plug-ins. Therefore, in the flowchart of the methodology (Figure 3-1), the patterns of the LCA database and the LCA tool partially overlap.

[1] https://www.autodesk.eu/products/revit/overview?term=1-YEAR&tab=subscription
[2] https://graphisoft.com/solutions/archicad
[3] https://www.allplan.com/de/

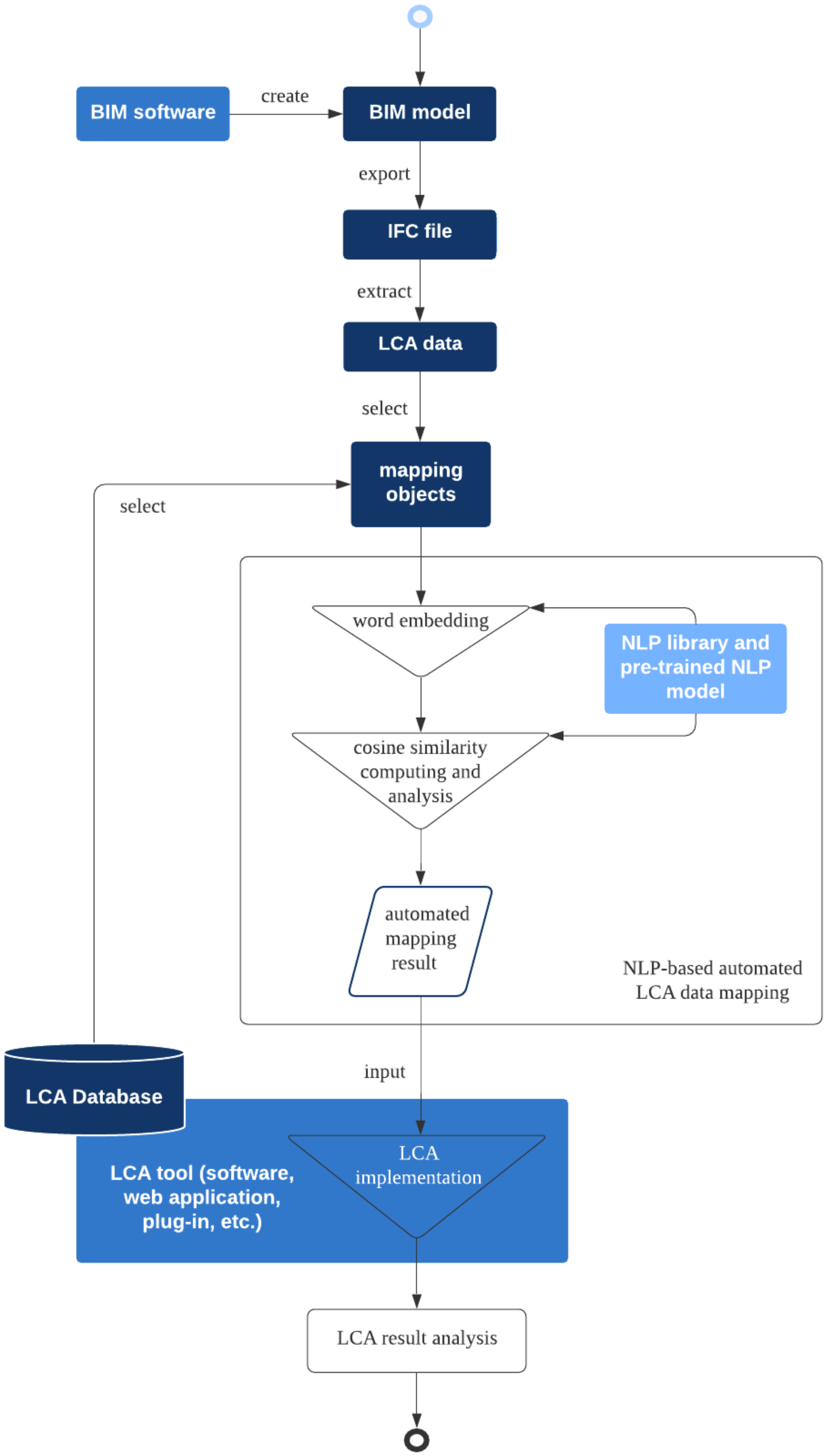


Figure 3-1 Flowchart of the methodology

When the mapping objects are prepared, NLP libraries and pre-trained NLP models are employed to tackle the issue of mapping the selected IFC data to the LCA database. More specifically, the items in the mapping objects are first encoded respectively for the IFC data and the LCA database through the pre-trained NLP model to obtain the word vectors, based on which the semantic similarities between the items can be computed.

In light of the introduction in Section 2.3.2, there are plenty of NLP libraries and pre-trained NLP models, e.g., NLTK, spaCy, BERT, which can be imported and loaded in C++, python, or java, to deal with multiple NLP tasks like tokenization and parsing. Taking into account that attribute descriptions of the building elements are commonly short and contain few stopwords, thus the NLP-based automated mapping method focuses on the numerical forms converted from the mapping objects rather than comparing their tokens.

In addition, cosine similarity is the best-performing text similarity metric for the reason that its precision will not be significantly affected regardless the document size. Hence, it is utilized to compute the semantic similarities in this method, the formula of which was listed in Section 2.3.3.

Aiming to inspect the quality of the NLP-based automated mapping method, the mapping result is input into the LCA tool. On this basis, an LCA analysis is implemented to reflect the building LCA impacts from overall perspective or in different building LCA stages through the calculated LCA indicators, e.g., global warming potential (GWP) and acidification potential (AP).

To conclude, for the first research question, this automated BIM-LCA data mapping method employs NLP technology to overcome the obstacle of machines in understanding human language, which occurs when recognizing the semantic information and computing the semantic similarity. For the second research question, three different NLP libraries and pre-trained NLP models are adopted, namely Gensim, spaCy, and BERT, the performance of which are observed and compared in context of the same workflow and case study. The processes and outcomes will be elaborated in the following chapters.

3.3 Workflow and Tools

For the sake of structuring and visualizing the automated method step by step in accordance with the methodology, a workflow is exploited as a prototype implementation and illustrated in Figure 3-2, where several various BIM/ LCA/ computational tools are used to cope with different tasks.

At first, a BIM model is created in the software Autodesk Revit 2022, from which an IFC file of the BIM model can be directly exported. The version of this exported IFC data model is IFC4 and the certification is IFC4 Design Transfer View, since IFC4 is the most advanced IFC schema that supports BIM collaboration better than the previous versions like IFC2x3, and IFC4 Design Transfer View is designed exclusively for the model data transfer usage to support further design, analysis and management purposes (buildingSMART International, 2019).

The next step is extracting the LCA-related data from the IFC file. The German Federal Institute for Research on Building, Urban Affairs and Spatial Development (BBSR) released an LCA web application eLCA v0.9.7 beta[1] to evaluate the building environmental impacts during its life cycle (eLCA, 2021). This application provides a tool, with which the LCA analysis required data embedded in the IFC file can be easily extracted as a CSV file, comprising the information of *GUID*, *Component Name*, *IFC Class*, *Cost Group Number/ Name*, *Material Name*, *Mass*, *Area*, etc. Table 3-1 displays an example of the data it extracts and stores, all the information needs to be automatically mapped and imported in eLCA for the evaluation of the building environmental impacts.

Table 3-1 An example of the extracted IFC data obtained through the eLCA tool

GUID	IFC Class	Cost Group	Component Name	Material Name	Mass	Area
1eEeqlWvH0$ uJ0CRYFb9GD	IfcWall	331 Tragende Außenwände	Basiswand: STB+WD 200+200	'HPL-Platte', 'Ortbeton - C30/37'		

[1] https://www.bauteileditor.de/

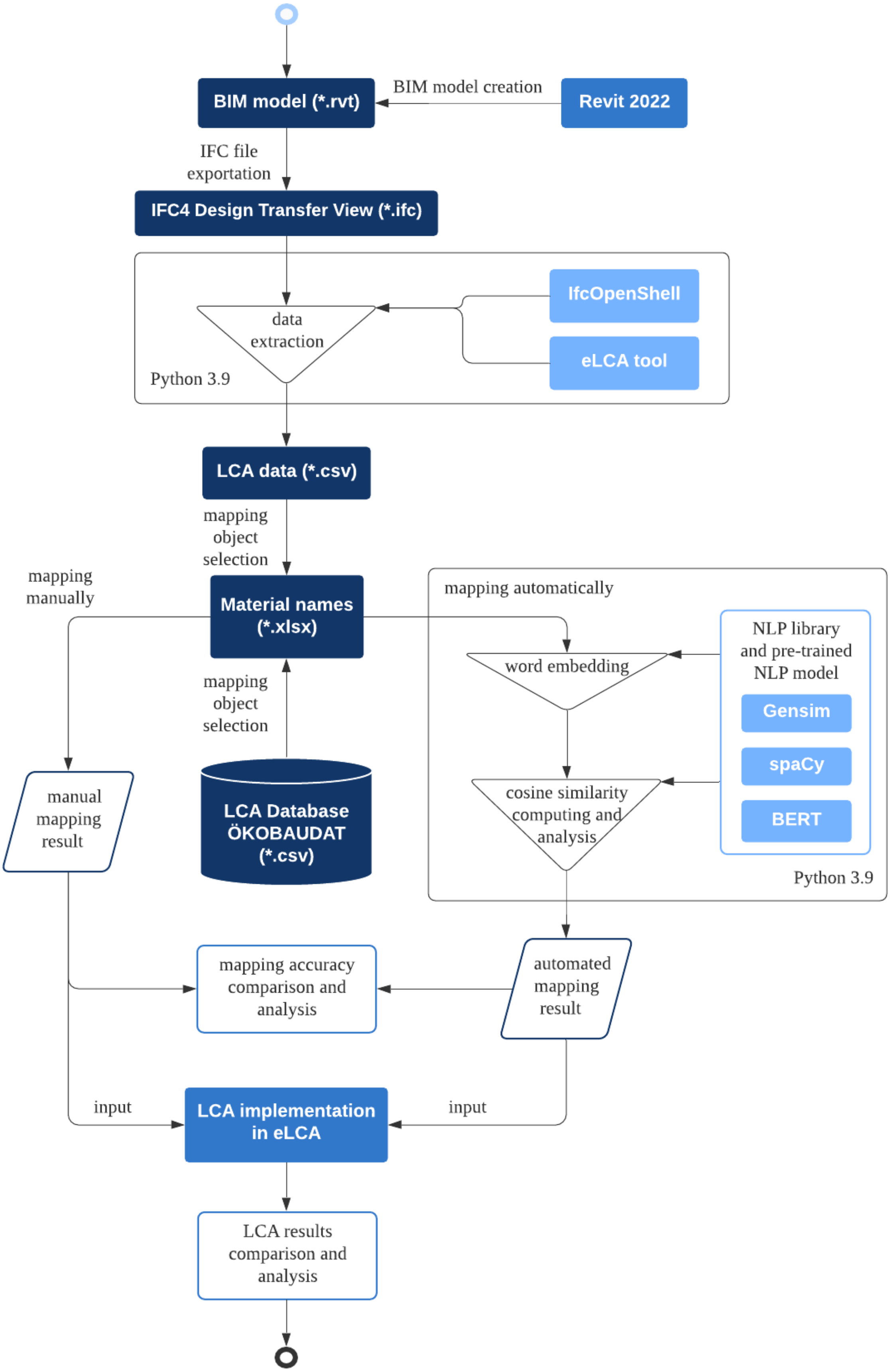


Figure 3-2 Workflow of the automated method of mapping LCA data to BIM models

It is noticed that if a building component is composed of multiple building materials, all the materials will be stored in one cell. In case that *Material Name* is selected as the mapping object, to split the materials that share the same cell can make the data mapping process more convenient. Besides, due to the multi-layered building components, the extraction of the mass and area works not well, which specifically manifests in the data loss.

This tool can be downloaded from its official website, and is expressed as a ZIP file that has packed several python scripts and operation manuals, within which, the IFC2LCA script has imported and taken use of the open-source library IfcOpenShell[2]. It derives from a project initiated by Krijnen in 2011 and is constantly developed, being geared towards providing an easy interface to extract and manipulate geometries in IFC files (Krijnen, 2021).

Making the appropriate choice of the LCA database is critical to ensuring the reliability of the mapping results. Table 3-2 compares the LCA/ LCI databases introduced in Section 2.1.4. Weighing up the size of the LCA/ LCI databases and the volumes of the data concerning the AEC industry they contained, together with considering the variation in climate, geography and material properties, Ökobaudat with the most updated version (2021_II) is determined as the target database for the workflow.

Table 3-2 Comparison of the LCA/ LCI databases

	Ecoinvent	GaBi	ELCD	USLCI	ProBas	Ökobaudat
Free of Charge	No	No	Yes	Yes	Yes	Yes
Data Volume*	> 14,000	> 15,000	> 300	> 600	> 8,000	> 1,400

Above all, the datasets in Ökobaudat are based on the background databases GaBi and Ecoinvent, both of which are European databases, being subject to strict quality requirements. Secondly, Ökobaudat owns the most abundant datasets for building products among all the databases. Thirdly, its data exchange interface enables via which other applications and software tools read datasets from Ökobaudat, and with certain permissions import them directly into Ökobaudat (BBSR, 2021). For instance,

[2] http://ifcopenshell.org/

* The “Data Volume” refers to the total number of datasets contained in the database except for Ökobaudat, which represents the number of datasets exclusively for building products.

the building components and LCA calculations in eLCA are dependent on the Ökobaudat datasets (eLCA, 2021). Last but not the least, it is publicly available and free of charge.

Alongside the web user interface, Ökobaudat can be downloaded as a ZIP archive or in CSV format as well. Table 3-3 shows an example of the data it stores, where *UUID* represents the universally unique identifier; *Category* indicates the hierarchical classes into which the material is classified; and *Modul* indicates the corresponding building LCA stage. Moreover, there are a lot of LCA-related material properties such as surface weight, bulk density, GWP, ozone depletion potential (ODP), photochemical ozone creation potential (POCP) and AP, and also some other information like expiry year and declaration owner, which are not listed below.

Table 3-3 An example of the data in the LCA database Ökobaudat

UUID	Name (DE)	Name (EN)	Category	Modul
64850805-e3ec-44f4-b7b6-dd10515b5e56	Dachziegel		'Mineralische Baustoffe' / 'Steine und Elemente' / 'Dachziegel'	A1

Not all information contained in the extracted IFC data is necessary to be mapped to the database, since they are bound in line with the building material, which means, if an attribute is successfully mapped, all other attributes will be automatically attached and mapped. Thereby, a data selection step is inserted before the mapping process, which can reduce the workload to a large extent. As a result, the *Material Name* is determined to be the mapping object in this workflow, for the reason that it is the most representative attribute of the data in Ökobaudat, and from the perspective of the extracted IFC data, it is the most complex attribute because the items in it need to be split.

Accordingly, for both the LCA-related IFC data and the LCA database Ökobaudat, the material names are separately picked out from the datasets and prepared to be mapped through both manual attempt and NLP-based automated methods. The manual mapping approach is set as the control group while keeping all the other steps and tools unchanged, i.e., it is assumed to be 100% accurate. The accuracy of the mapping results will be calculated following Formula 3.1.

$$accuracy = \frac{number\ of\ the\ correctly\ mapped\ items}{total\ number\ of\ the\ items\ to\ be\ mapped} \times 100\% \qquad (3.1)$$

It should be noted that some of the materials in Ökobaudat do not have English names for this LCA database is developed by BBSR and the majority of the data declaration owners come from Germany. Therefore, in order to ensure the mapping quality, the pre-trained NLP model used for the automated mapping method is preferably German language model.

Table 3-3 compares the five NLP tools introduced in Section 2.3.2. As is shown, Stanza, spaCy and BERT officially provide pre-trained German language models, while NLTK and Gensim don not.

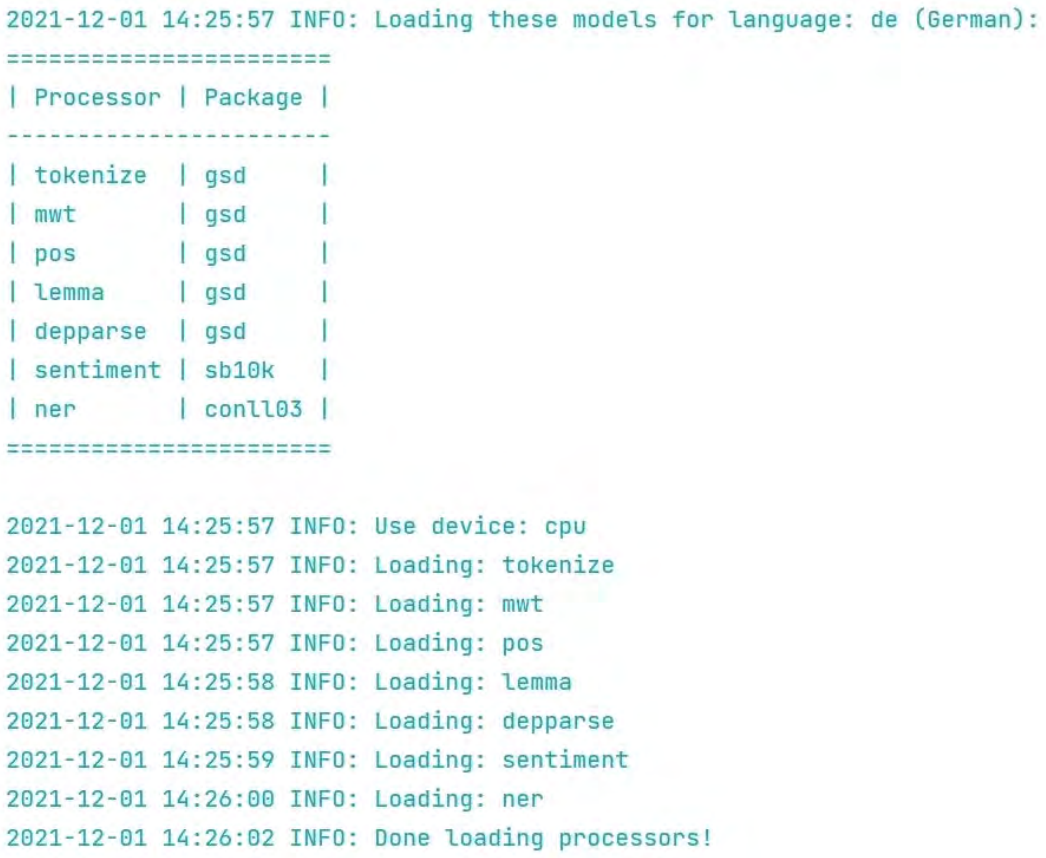
Table 3-4 NLP tools

	NLTK	CoreNLP/ Stanza	Gensim	spaCy	BERT
Official Pre-trained German Language Model	No	Yes	No	Yes	Yes
Word Embedding Processor	Yes	No	Yes	Yes	Yes

However, there is no word embedding processor in the Stanza language model, i.e., the cosine similarity between sentences cannot be computed based on the document obtained through the NLP model since it involves no word vector. Figure 3-3 exhibits the processors contained in the Stanza German language model.

In addition, NLTK is helpful for simple text analysis instead of a large amount of data (Bushkovskyi, 2019), and it mostly supports English language. Hence, it is not a good choice for the automatic mapping in this research, either.

Furthermore, despite Gensim offers no official pre-trained German language model, Müller trained one with Gensim Word2Vec library on the German Wikipedia and German news articles (Müller, 2015), which is able to measure the semantic similarity between two German words straightly.



```
2021-12-01 14:25:57 INFO: Loading these models for language: de (German):
=======================
| Processor | Package |
-----------------------
| tokenize  | gsd     |
| mwt       | gsd     |
| pos       | gsd     |
| lemma     | gsd     |
| depparse  | gsd     |
| sentiment | sb10k   |
| ner       | conll03 |
=======================

2021-12-01 14:25:57 INFO: Use device: cpu
2021-12-01 14:25:57 INFO: Loading: tokenize
2021-12-01 14:25:57 INFO: Loading: mwt
2021-12-01 14:25:57 INFO: Loading: pos
2021-12-01 14:25:58 INFO: Loading: lemma
2021-12-01 14:25:58 INFO: Loading: depparse
2021-12-01 14:25:59 INFO: Loading: sentiment
2021-12-01 14:26:00 INFO: Loading: ner
2021-12-01 14:26:02 INFO: Done loading processors!
```

Figure 3-3 Processors contained in the Stanza German language model

Comprehensively considering the strengths and weaknesses of the five NLP tools, the three tools – Gensim, spaCy, and BERT – together with their corresponding pre-trained German language models, namely *german.model*[3], *de_core_news_lg*[4], and *bert_base_german_cased*[5], are respectively employed for the automatic mapping of material names between the extracted IFC data and the LCA database Ökobaudat, concretely in the word embedding and cosine similarity computing tasks.

The data mapping through the three different NLP tools are accomplished by similar python codes with the same mapping algorithm, which is illustrated in Figure 3-4. The entire automated mapping process is implemented in PyCharm[6] in the environment of Python 3.9[7].

[3] https://devmount.github.io/GermanWordEmbeddings/
[4] https://spacy.io/models/de
[5] https://huggingface.co/bert-base-german-cased
[6] https://www.jetbrains.com/pycharm/
[7] https://www.python.org/

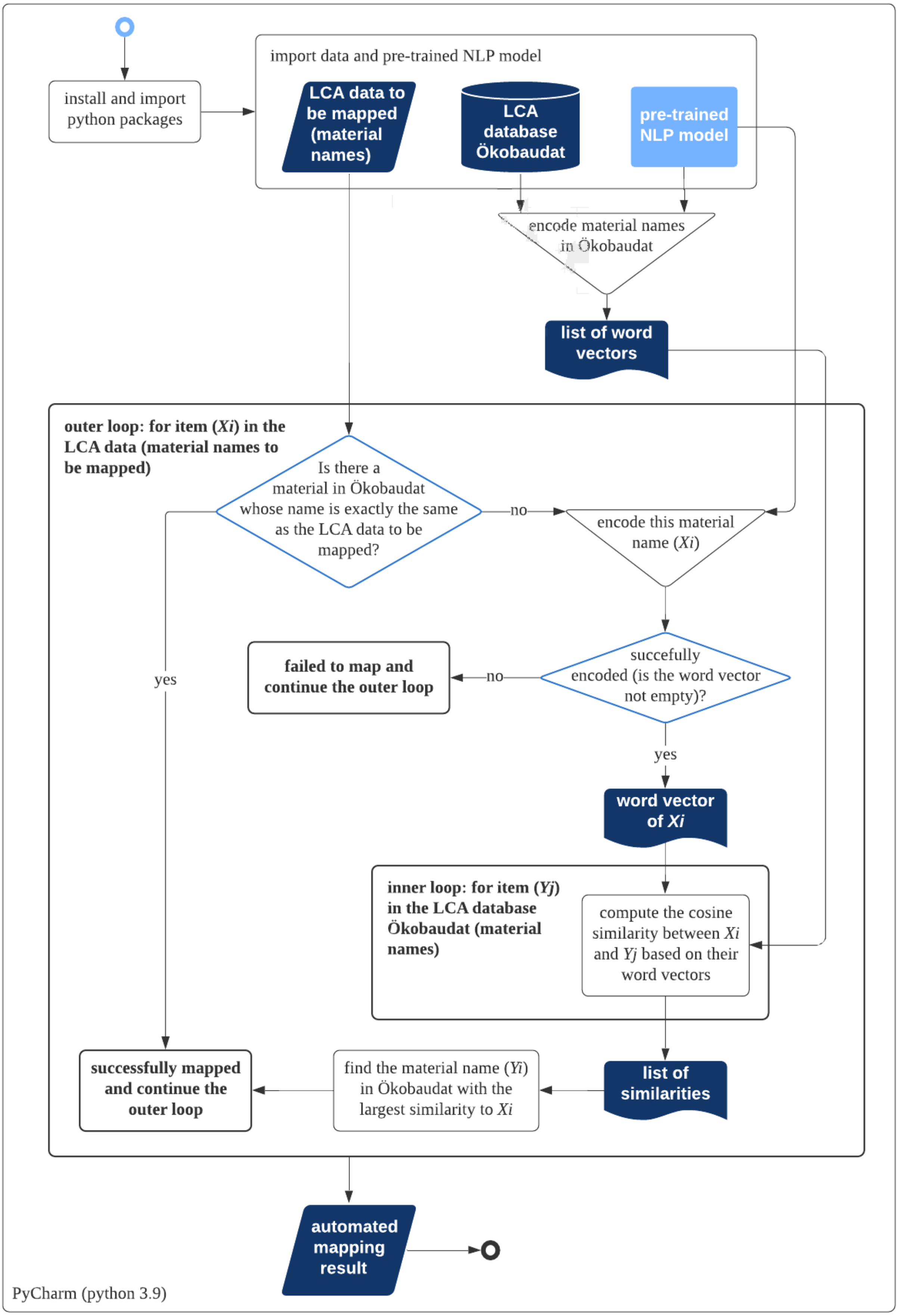

install and import python packages
import data and pre-trained NLP model
LCA data to be mapped (material names)
LCA database Ökobaudat
pre-trained NLP model
encode material names in Ökobaudat
list of word vectors
outer loop: for item (Xi) in the LCA data (material names to be mapped)
Is there a material in Ökobaudat whose name is exactly the same as the LCA data to be mapped?
no
encode this material name (Xi)
yes
failed to map and continue the outer loop
no
succefully encoded (is the word vector not empty)?
yes
word vector of Xi
inner loop: for item (Yj) in the LCA database Ökobaudat (material names)
compute the cosine similarity between Xi and Yj based on their word vectors
successfully mapped and continue the outer loop
find the material name (Yi) in Ökobaudat with the largest similarity to Xi
list of similarities
automated mapping result
PyCharm (python 3.9)

Figure 3-4 Algorithm of the automated mapping

After installing and importing the necessary python libraries and packages in the beginning, the extracted IFC data and Ökobaudat, as well as the pre-trained NLP model are imported and loaded, followed by the text encoding of the material names in Ökobaudat.

Through the spaCy model or the BERT model, the material names can be directly processed and transformed into word vectors with a simple line of python code. Whereas, the Gensim model enables only basic single words to be directly embedded, while lots of terminologies in the AEC industry are compound words and some material names in this study consist of multiple words. To make up for this shortcoming, a function is written to divide the multi-word material names into individual words and convert them into word vectors.

The main part of the algorithm is comprised of a nested loop, where all materials in Ökobaudat will be traversed once (the inner loop) for each material in the exported IFC data (the outer loop).

More specifically, in each process of the outer loop (for each material named Xi in the IFC data), the first thing is to check whether there is a material in Ökobaudat the name of which is exactly the same as Xi. This step is based on the consideration of the status quo that most building LCA studies use predefined datasets in LCA databases to reduce the workload of data collection, selection and validation (Hollberg et al., 2021). When the answer of this conditional statement is yes, the material Xi will be automatically and successfully mapped according to the identical name, i.e., mapping through simple computer programming instructions without introducing NLP technology. If no, Xi will be encoded through the NLP model.

Nevertheless, due to the limitation of the training corpus or the NL recognition capability of the NLP model, the word embedding is not always smoothly, which manifests in the empty word vector, i.e., all elements of the n-dimensional array are zero. This phenomenon also occurs when encoding the material names in Ökobaudat. Consequently, here comes the second conditional statement before the inner loop, which aims to filter out the material names failed to be encoded from the successfully encoded ones.

For each Xi with non-empty word vector, the word vectors of the material names (Yj) in Ökobaudat are traversed, during which the cosine similarity between Xi and each Yj are computed based on their word vectors and then sorted in sequence. After traversing the entire LCA database, the largest similarity can be found, and accordingly the material Yi in Ökobaudat that is most similar to Xi can be discovered.

In the last step of the workflow, the mapping results of the NLP-based automated approach are analyzed in comparison with the manual method, and the performance of the three NLP tools are compared to find out the most accurate and reliable one. Meanwhile, separately based on the manual mapping result and the NLP-based automated mapping results, LCA analyses are implemented in the web application eLCA. The differences and errors of the eLCA outcomes are regarded as an additional parameter to validate the methodology.

4 Case Study

In order to verify the feasibility of the automated method proposed in Chapter 3, as well as to quantitatively analyze and compare the contributions of the three NLP technologies to the mapping task, a case study was conducted.

4.1 BIM Model Creation and IFC File Exportation

Figure 4-1 displays a BIM model of a joint building, which was created in Autodesk Revit 2022. The building comprises of a two-story house, a three-story house and a connecting corridor, containing roofs, walls, slabs, foundation, doors and windows, which are the most commonly-used building elements. Notwithstanding the automated method can be applied to more complex designs, it is not necessary, since this simple model suffices for the implementation of the methodology in accordance with the workflow.

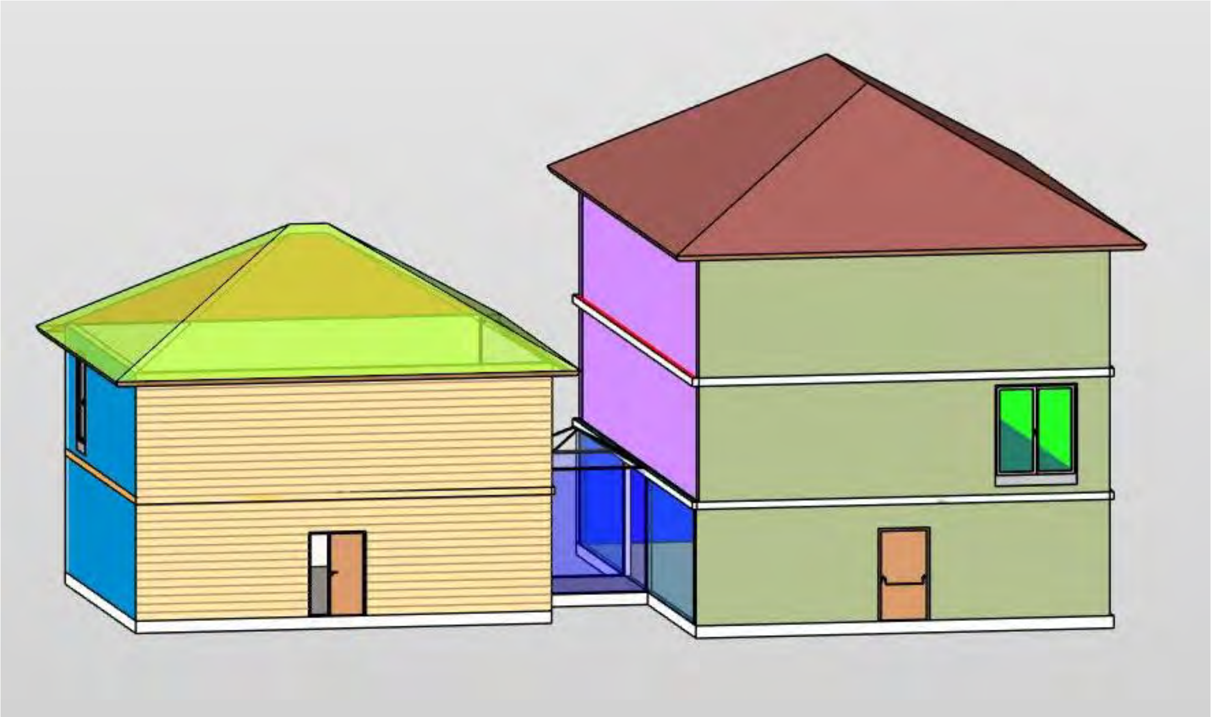
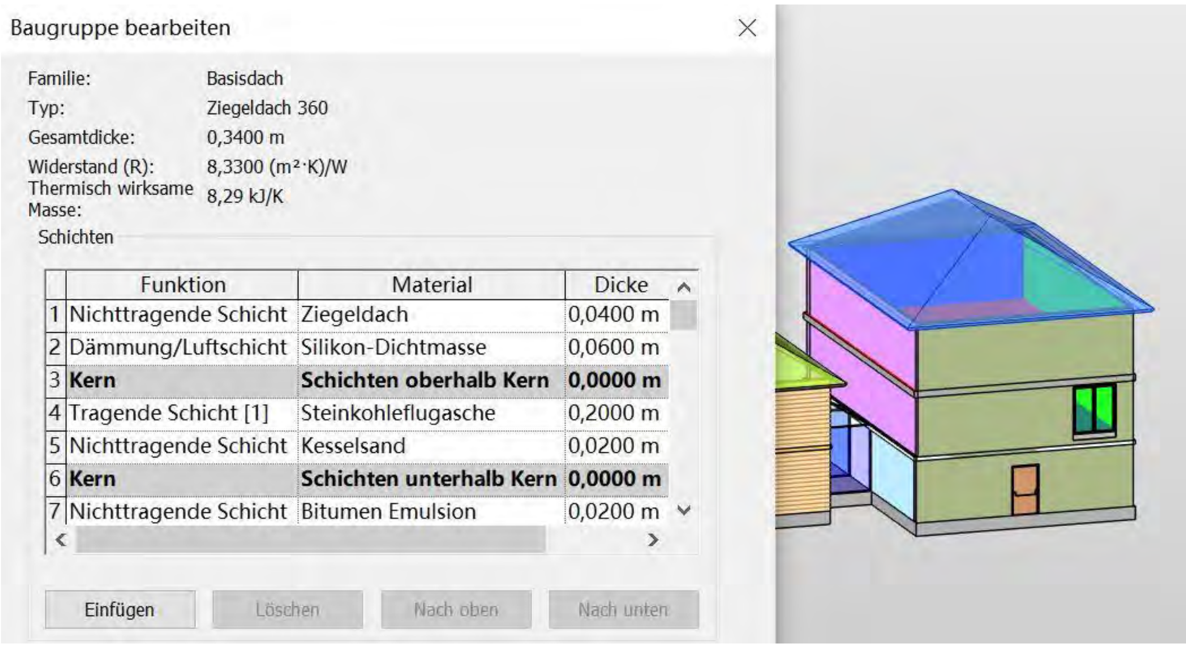


Figure 4-1 The building for the case study modeled in Revit 2022

To reduce errors and ensure the reliability of the mapping results, the number and categories of the construction materials ought to be as much and abundant as possible. For this purpose, each building component is composed of multiple layers, which are separately assigned to different materials. For example, as shown in Figure 4-2, the roof of the three-story house consists of five material layers, namely “Ziegeldach”, “Silikon-Dichtmasse”, “Steinkohleflugasche”, “Kesselsand”, and “Bitumen Emulsion”, which are called “tile roof”, “silicone sealant”, “coal fly ash”, “boiler sand”, and “bitumen emulsion” in English.


Baugruppe bearbeiten

Familie: Basisdach
Typ: Ziegeldach 360
Gesamtdicke: 0,3400 m
Widerstand (R): 8,3300 (m²·K)/W
Thermisch wirksame Masse: 8,29 kJ/K

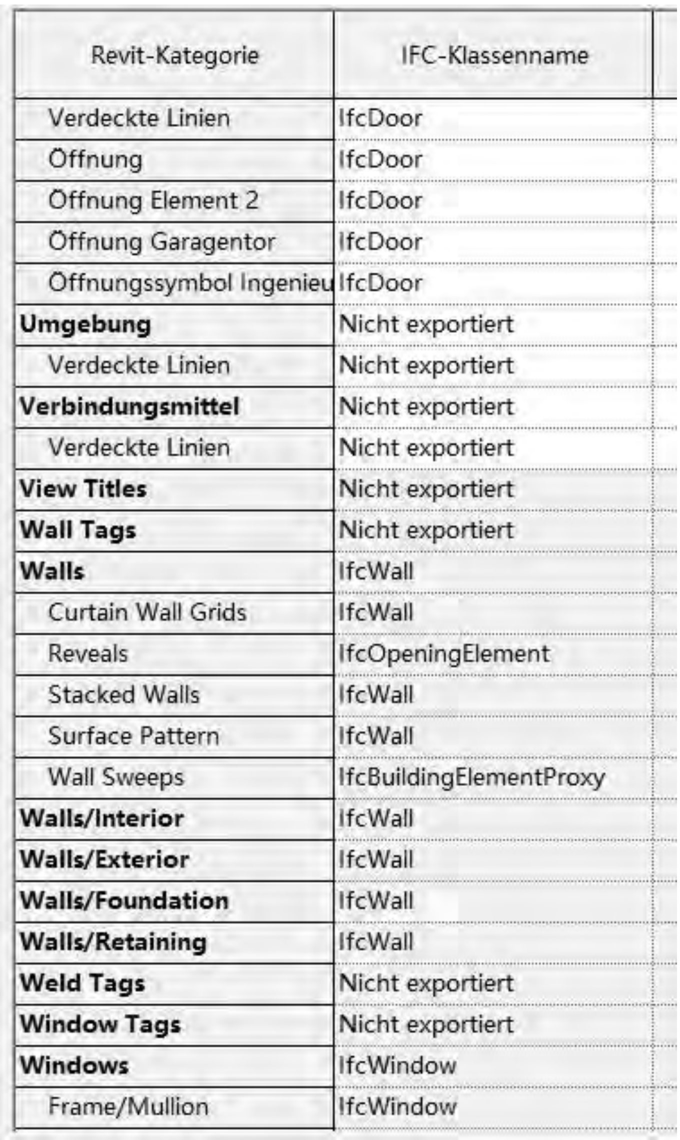
Schichten

	Funktion	Material	Dicke
1	Nichttragende Schicht	Ziegeldach	0,0400 m
2	Dämmung/Luftschicht	Silikon-Dichtmasse	0,0600 m
3	**Kern**	**Schichten oberhalb Kern**	**0,0000 m**
4	Tragende Schicht [1]	Steinkohleflugasche	0,2000 m
5	Nichttragende Schicht	Kesselsand	0,0200 m
6	**Kern**	**Schichten unterhalb Kern**	**0,0000 m**
7	Nichttragende Schicht	Bitumen Emulsion	0,0200 m

Figure 4-2 An example of the material composition of a building element (one of the roofs)

It needs to be pointed out that some materials are not suitable for constructing the building components to which they belong, e.g., “coal fly ash” and “bitumen emulsion” used for the roof. Nevertheless, this is not conflicting in that the research is centered on the mapping performance rather than the architectural reality.

After creating the BIM model, the next step in the workflow is to export the IFC file, before which, it is needed to set the IFC Options. An IFC Options that has been set up is aimed at determining what information is needed in the downstream processes, and it can be saved as a TXT file for loading in the future work which has the same requirement for the exported information. Figure 4-3 shows part of the IFC Options in this case study, where the names of IFC classes, e.g., *IfcWall*, *IfcRoof*, were added to the corresponding building elements that need to be exported, which contain the information of the materials. The non-LCA related data such as component tags, and the information not involved in this BIM model like MEP design were set not to be exported.



Revit-Kategorie	IFC-Klassenname
Verdeckte Linien	IfcDoor
Öffnung	IfcDoor
Öffnung Element 2	IfcDoor
Öffnung Garagentor	IfcDoor
Öffnungssymbol Ingenieu	IfcDoor
Umgebung	Nicht exportiert
Verdeckte Linien	Nicht exportiert
Verbindungsmittel	Nicht exportiert
Verdeckte Linien	Nicht exportiert
View Titles	Nicht exportiert
Wall Tags	Nicht exportiert
Walls	IfcWall
Curtain Wall Grids	IfcWall
Reveals	IfcOpeningElement
Stacked Walls	IfcWall
Surface Pattern	IfcWall
Wall Sweeps	IfcBuildingElementProxy
Walls/Interior	IfcWall
Walls/Exterior	IfcWall
Walls/Foundation	IfcWall
Walls/Retaining	IfcWall
Weld Tags	Nicht exportiert
Window Tags	Nicht exportiert
Windows	IfcWindow
Frame/Mullion	IfcWindow

Figure 4-3 Part of the IFC Options

Finally, for the purpose of data exchange, the IFC4 Design Transfer View was selected as the version of the exported IFC file, with exporting the general IFC property sets and the basic quantities.

4.2 LCA-related IFC Data Extraction and Processing

The data embedded in the exported IFC file was extracted by means of performing the eLCA IFC2LCA python script. On account that the obtained CSV file is inconvenient for human reading and the manual mapping, a python script (csv2xlsx.py) was written to convert CSV file to human-readable MS Excel file. The data in the converted Excel file is regarded as the raw data, part of which is displayed in Table 4-1.

Table 4-1 Part of the raw IFC data

GUID	IFC Class	Cost Group	Component Name	Material Name
1eEeqlWvH0$uJ0CRYFb9GD	IfcWall	331 Tragende Außenwände	Basiswand: STB+WD 200+200	'HPL-Platte', 'Ortbeton - C30/37'
3Froekll18vhjGoGPvxx4u	IfcSlab	351 Deckenkonstruktionen	Geschossdecke: STB 300	'Ortbeton - C30/37', 'Walzplattiertes Grobblech', 'Doppelbodensystem Typ LIGNA'
3Froekll18vhjGoGPvxxBT	IfcSlab	351 Deckenkonstruktionen	Geschossdecke: FB 180 Teppich	'Fussboden - Teppich', 'Ortbetonestrich', 'Keramische Fliesen', 'Nadelschnittholz - getrocknet', 'Schuettung aus Polystyrolschaumstoff-Partikeln'
0OEkutEW5EMeg5GLjflcAQ	IfcRoof	300 Bauwerk Baukonstruktionen	Basisdach: Ziegeldach 360	'Ziegeldach', 'Silikon-Dichtmasse', 'Steinkohleflugasche', 'Kesselsand', 'Bitumen Emulsion'

The first attribute *GUID* expresses the globally unique identifier utilized by the IFC specification for exchange purpose of the object instances (buildingSMART International, 2021g). The *IFC Class* describes the specific IFC entity class (*IfcWall*, *IfcSlab*, *IfcRoof*, etc.) of the building element. In this case study, all the IFC classes inherit from the class IfcBuildingElement (see Figure 2-6), which is a subclass of the *IfcRoot*. Thus, all the building elements possess their own GUID.

The third column *Cost Group* involves the cost group numbers and names of the building components, which are classified and defined in the DIN 276, and can be input in LCA tools (e.g., eLCA) to support the LCA analysis. In this case study, all the cost groups belong to subgroups of the cost group 300 Bauwerk – Baukonstruktionen. The *Component Name* can also be queried in LCA tools to check if there is already a pre-defined one.

The last attribute *Material Name* expresses the materials, which constitute the building components (see Figure 4-2). It was determined as the mapping object in this workflow and case study for it is the most representative and complex attribute of the data in both the IFC file and the LCA database Ökobaudat.

Obviously, the data in the same row belongs to one independent building element, consequently some cells in the *Material Name* column comprise more than one material, which need to be split. In addition, there are several repetitive materials (also reflected in the table above), which will be condensed by the name to reduce the workload. To deal with the two issues, two python scripts (split_data.py and condense_data.py) were written separately and ran sequentially.

In the end, a total of 40 non-repetitive building materials were obtained from the IFC file, the entire and elaborated information of which is listed in Appendix A. Table 4-2 exhibits part of the extracted, split, and condensed LCA-related IFC data.

Table 4-2 Part of the extracted, split, condensed and selected IFC data

GUID	IFC Class	Cost Group	Component Name	Material Name
1eEeqlWvH0$uJ0CRYFb9GD	IfcWall	331 Tragende Außenwände	Basiswand: STB+WD 200+200	HPL-Platte
1eEeqlWvH0$uJ0CRYFb9GD	IfcWall	331 Tragende Außenwände	Basiswand: STB+WD 200+200	Ortbeton - C30/37
2CJ9v0fLT2QB1aN$GEssv5	IfcWall	341 Tragende Innenwände	Basiswand: KS 240	Kalksandstein
3FroekIl18vhjGoGPvxxBT	IfcSlab	351 Deckenkonstruktionen	Geschossdecke: FB 180 Teppich	Keramische Fliesen
0OEkutEW5EMeg5GLjflcAQ	IfcRoof	300 Bauwerk Baukonstruktionen	Basisdach: Ziegeldach 360	Bitumen Emulsion
0OEkutEW5EMeg5GLjfId_E	IfcPlate	346 Elementierte Innenwandkonstruktionen	Fassadenelemente: Dachverglasung	Glas - Sicherheitsverglasung
1rqGkujSH97xGP0fn8zaQL	IfcDoor	344 Innenwandöffnungen	TU DF 1 - Rahmenstock mit Glasseitenteil: DL - 800 x 2000 - MB 1400	Lack
2rtQ5uukz4NBMM$uyTMi1R	IfcWall	342 Nichttragende Innenwände	Basiswand: GK VS 150	Lehmplatte

4.3 LCA Database Ökobaudat Condensation

The LCA database Ökobaudat can be downloaded directly from its official website as a CSV file. In this case study, the most updated version OBD_2021_II was adopted. For the same reason as the extracted IFC data – for better human reading and more convenient manual mapping, it was also converted to an Excel file by running csv2xlsx.py, part of which is listed in Table 4-3.

Table 4-3 Part of the LCA database Ökobaudat

UUID	Name (DE)	Name (EN)	Modul
64850805-e3ec-44f4-b7b6-dd10515b5e56	Dachziegel		A1
64850805-e3ec-44f4-b7b6-dd10515b5e56	Dachziegel		A2
64850805-e3ec-44f4-b7b6-dd10515b5e56	Dachziegel		A3
64850805-e3ec-44f4-b7b6-dd10515b5e56	Dachziegel		A4
64850805-e3ec-44f4-b7b6-dd10515b5e56	Dachziegel		A5
64850805-e3ec-44f4-b7b6-dd10515b5e56	Dachziegel		B1
64850805-e3ec-44f4-b7b6-dd10515b5e56	Dachziegel		B4
64850805-e3ec-44f4-b7b6-dd10515b5e56	Dachziegel		B5
64850805-e3ec-44f4-b7b6-dd10515b5e56	Dachziegel		C1
64850805-e3ec-44f4-b7b6-dd10515b5e56	Dachziegel		C2
64850805-e3ec-44f4-b7b6-dd10515b5e56	Dachziegel		C3
64850805-e3ec-44f4-b7b6-dd10515b5e56	Dachziegel		C4
64850805-e3ec-44f4-b7b6-dd10515b5e56	Dachziegel		D

There are a lot of duplicate data since the database takes into account life cycle stages of buildings (see Figure 2-2), which share the same other attributes like UUID and material names (German name, English name, and French name). On account that this thesis focuses on the mapping of material names between the LCA-related IFC data and the LCA database rather than the LCA calculations at different stages, the repeated data was cut down to only one for each material by running the script condense_data.py. As a result, the number of unique material names in Ökobaudat is 1059. Through condensing the material names, the run time particularly the traversal time of subsequent mapping work can be saved to a large extent.

To be noticed that some of the materials do not have English names, which is one of the concerns that the material names were edited to be German when creating the BIM model. Hence, the NLP tools adopted later will support German language models as well.

4.4 Manual Mapping

The materials were manually mapped according to their names, which spent about 30 minutes. Table 4-4 exhibits part of the manual mapping result.

Table 4-4 Part of the manual mapping result

Materials in BIM Model	Materials in LCA Database Ökobaudat
Ortbeton - C30/37	Transportbeton C30/37
Ziegel mit Daemmstoff	Mauerziegel (Daemmstoff gefuellt)
Gipskartonplatte (Feuerschutz) 0,0125 m	Gipskartonplatte (Feuerschutz) (Dicke 0,0125 m)
Gipskartonplatte (impraegniert) 0,0125 m	Gipskartonplatte (impraegniert) (Dicke 0,0125 m)
Fussboden - Teppich	Fussbodenbelag mehrschichtiges Nadelvlies (Teppichboden, 1400 g/m^2)
Ortbetonestrich	Zementestrich
Schuettung aus Polystyrolschaumstoff-Partikeln	Schuettung aus Polystyrolschaumstoff-Partikeln (ohne Bindemittel)
Ziegeldach	Dachziegel
Bitumen Emulsion	Bitumen Emulsion (40% Bitumen, 60% Wasser)
Mehrschichtige Massivholzplatte	3- und 5-Schicht Massivholzplatte (Durchschnitt DE)
Glas - Sicherheitsverglasung	Flachglas - Guardian - Verbundsicherheitsglas (VSG)
Lack	Lacksysteme Holzfassade deckend (Decklacksystem)
Glas - Isolierverglasung	Isolierglas 2-Scheiben
Daemmung - XPS	XPS-Daemmstoff
Lehmplatte	Lehmbauplatte

For the sake of simulating the reality that currently most LCA studies use available datasets in LCA databases to reduce the workload of data collection, selection and validation (Hollberg et al., 2021), a part of the material names in the processed IFC data was set to be quite similar to the material names in Ökobaudat, by mean of choosing certain predefined materials in Revit or adapting the material names of the LCA data in Ökobaudat. As a consequence, half of the 40 materials can be found in Ökobaudat with the identical names.

4.5 Automated Mapping Using NLP Libraries and Pre-trained NLP Models

According to the workflow and the algorithm of the NLP-based automated mapping method presented in Chapter 3.3, the material names of the data in Ökobaudat were encoded through the pre-trained NLP model as a list of word vectors before the nested loop. Even if the number of unique material names in Ökobaudat is 1059, not all of them could be smoothly recognized and encoded. The ones failed to be encoded were expressed as empty word vectors, the cosine similarity computed based on which will be zero.

In the beginning of the outer loop, 20 materials in the processed IFC data were successfully mapped through the first conditional statement, i.e., according to the exactly same names. Then the remaining 20 materials were encoded for mapping by introducing the NLP technology, and the failed ones were described as empty word vectors as well. For each of the successfully encoded materials in the processed IFC data, at most 1059 similarities will be computed and compared to detect the most similar material in the database.

If the materials are automatically mapped only in line with the exactly same names, the mapping accuracy shall be equal to 50.0%. Regarding the manual mapping result as the correct solution, i.e., setting the accuracy of the manual mapping to 100.0%, and considering the basic accuracy of the automated mapping as 50.0%, the introduction of NLP tools aims to explore the extent to which they can improve the accuracy of automated mapping.

Three different NLP tools, namely Gensim, spaCy and BERT, and the corresponding pre-trained German language models were employed, to encode the material names in the processed IFC data and the LCA database Ökobaudat, and compute their semantic similarities based on the obtained word vectors.

4.5.1 Word Embedding and Semantic Similarity Computation through Gensim

Since Gensim does not provide official pre-trained German language model, this case study took use of the NLP model trained by Müller in his bachelor’s thesis (Müller, 2015), called “german.model”. In addition, this model can encode only basic single words, whereas most of the material names in this case study are compound words or consist of multiple words. To tackle this problem, a function is written to divide the multi-word material names into individual words and encode them into word vectors through this pre-trained NLP model. Table 4-5 lists part of the mapping result.

Table 4-5 Part of the mapping result obtained through Gensim

Materials in BIM Model	Materials in LCA Database Ökobaudat
Ortbeton - C30/37	Zementmoertel
Ziegel mit Daemmstoff	Dachziegel
Gipskartonplatte (Feuerschutz) 0,0125 m	[EV]*
Gipskartonplatte (impraegniert) 0,0125 m	[EV]
Fussboden - Teppich	Decke - HT Labor+Hospitaltechnik GmbH - Decke fuer Hygienebereiche
Ortbetonestrich	[EV]
Nadelschnittholz - getrocknet	[EV]
Schuettung aus Polystyrolschaumstoff-Partikeln	Schuettung aus Polystyrolschaumstoff-Partikeln (ohne Bindemittel)
Ziegeldach	Lehmputz
Bitumen Emulsion	Bitumen Emulsion (40% Bitumen, 60% Wasser)
Mehrschichtige Massivholzplatte	[EV]
Glas - Sicherheitsverglasung	Glas - Promat GmbH - Promaglas G30 type 1
Lack	Folie aus Polytetrafluorethylen (PTFE)
Glas - Isolierverglasung	Glas - Promat GmbH - Promaglas G30 type 1
Daemmung - XPS	Fussbodenheizung PEX (100 mm Abstand)

The yellow marks indicate the wrong outcomes referring to the manual mapping result. The accuracy of the mapping result obtained with the pre-trained Gensim Word2Vec NLP model is 57.5%. Considering that half (20) of the materials can be found in Ökobaudat with identical names, the items that are correctly mapped through the Gensim tool and the pre-trained German Word2Vec model are merely 3 out of 20.

The reason for the low accuracy lies in the limited training corpus of the model. First of all, this model is derived from a study in 2015, which has been six years since then, so that the corpora are not up-to-date for now. Besides, the training corpus is composed

* [EV] stands for empty vector, which denotes the name of the material in the BIM model cannot be recognized and encoded.

of German Wikipedia and news, which is not an exclusive corpus of the AEC industry and building materials. Hence, even though this model includes word embedding pipeline, it does not have vectors for every word in the material names in the processed IFC data and Ökobaudat. Only 12 out of 20 items in the IFC material names, and 508 out of 1059 items in the Ökobaudat material names have been successfully queried and embedded to word vectors, which means this model has no vector for the remaining approximately half of the entire data.

4.5.2 Word Embedding and Semantic Similarity Computation through spaCy

Table 5-6 shows part of the automated mapping result obtained through spaCy.

Table 4-6 Part of the mapping result obtained through spaCy

Materials in BIM Model	Materials in LCA Database Ökobaudat
Ortbeton - C30/37	Transportbeton C30/37
Ziegel mit Daemmstoff	Betonpflasterstein mit Edelsplittvorsatzbeton
Gipskartonplatte (Feuerschutz) 0,0125 m	Gipskartonplatte (Feuerschutz) (Dicke 0,0125 m)
Gipskartonplatte (impraegniert) 0,0125 m	Gipskartonplatte (Feuerschutz) (Dicke 0,0125 m)
Fussboden - Teppich	Flachglas - Guardian - Flachglas unbeschichtet
Ortbetonestrich	[EV]
Nadelschnittholz - getrocknet	Flachglas - Guardian - Flachglas unbeschichtet
Schuettung aus Polystyrolschaumstoff-Partikeln	Fernwaerme aus Abfaellen
Ziegeldach	Dachziegel
Bitumen Emulsion	Acrylat Dichtmasse
Mehrschichtige Massivholzplatte	binderholz Massivholzplatte
Glas - Sicherheitsverglasung	Flachglas - Guardian - Flachglas unbeschichtet
Lack	Innenfarbe Dispersionsfarbe scheuerfest
Glas - Isolierverglasung	Flachglas - Guardian - Flachglas beschichtet
Daemmung - XPS	Flachglas - Guardian - Flachglas unbeschichtet

For German language, spaCy offers four different pre-trained NLP models. In this case study, the large one, namely “de_core_news_lg” was adopted, which was trained based on the sources such as TIGER Corpus[1], Tiger2Dep[2] and WikiNER[3]. It is equipped with processors and pipelines of tok2vec, tagger, parser, lemmatizer, etc.

The accuracy of the mapping result is 65.0%. Compared with the basic mapping accuracy (50.0%), spaCy has improved the accuracy of automated mapping method by 15.0% by successfully mapping 6 out of 20 materials that own different names.

SpaCy possesses a strength among the three NLP tools that with which the semantic similarity can be computed with ease, as the model contains a pipeline that can directly obtain the similarity. Nevertheless, it has the same but much lighter weakness as the previously adopted Gensim model that there are some words for which the spaCy model has no vector. Three items in the processed IFC data and 107 items in Ökobaudat cannot be converted into word vectors through the pre-trained spaCy German language model.

4.5.3 Word Embedding and Semantic Similarity Computation through BERT

The pre-trained NLP model for the word embedding and semantic similarity computing tasks through BERT is the BERT German language model named "bert_base_german_cased", the most updated version of which was released in April, 2020. This model was trained using Google's Tensorflow code on the latest German Wikipedia dump, the OpenLegalData dump and news articles. Table 4-7 exhibits part of the mapping result.

Through BERT, 13 out of 20 materials with different names have been correctly mapped. Accordingly, the mapping accuracy of automated method has increased from the basic (50.0%) to 82.5%, which is a significant progress.

Not only the BERT model has performed excellently in the semantic similarity computation, but also it serves well in the word embedding compared with the previous two NLP tools, which manifests in that all the material names both in the processed IFC data and in Ökobaudat have been successfully recognized and transformed into word vectors.

[1] https://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/tiger/
[2] https://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/tiger2dep/
[3] https://figshare.com/articles/dataset/Learning_multilingual_named_entity_recognition_from_Wikipedia/5462500

Table 4-7 Part of the mapping result obtained through BERT

Materials in BIM Model	Materials in LCA Database Ökobaudat
Ortbeton - C30/37	Transportbeton C30/37
Ziegel mit Daemmstoff	XPS-Daemmstoff
Gipskartonplatte (Feuerschutz) 0,0125 m	Gipskartonplatte (Feuerschutz) (Dicke 0,0125 m)
Gipskartonplatte (impraegniert) 0,0125 m	Gipskartonplatte (impraegniert) (Dicke 0,0125 m)
Fussboden - Teppich	Fussbodenbelag mehrschichtiges Nadelvlies (Teppichboden, 1400 g/m^2)
Ortbetonestrich	Faserzementplatte (Fassade)
Nadelschnittholz - getrocknet	Nadelschnittholz - getrocknet (Durchschnitt DE)
Schuettung aus Polystyrolschaumstoff-Partikeln	Schuettung aus Polystyrolschaumstoff-Partikeln (zementgebunden)
Ziegeldach	Vormauerziegel
Bitumen Emulsion	Extrudierter Polystyrol Daemmstoff (XPS)
Mehrschichtige Massivholzplatte	binderholz Massivholzplatte
Glas - Sicherheitsverglasung	Sonnenschutzlammellen Metall
Lack	Lacksysteme Holzfassade deckend (Decklacksystem)
Glas - Isolierverglasung	Isolierglas 2-Scheiben
Daemmung - XPS	XPS-Daemmstoff

4.6 Comparison of Mapping Results

Table 4-8 compares the mapping results obtained through the three different NLP tools in aspects of the mapping accuracy and the program run time, taking the manual mapping result as the control group.

Table 4-8 Comparison of mapping results

	Manual	Gensim	spaCy	BERT
Automated	No	Yes	Yes	Yes
Accuracy	100.0%	57.5%	65.0%	82.5%
Time*	30 min	6 s	8 s	60 s

As it is shown in the table above, when the accuracy ascends, the time to obtain the mapping result increases simultaneously. The three NLP-based data mapping methods have increased the basic accuracy (50.0%) of the automatic mapping by 7.5%, 15.0%, and 32.5%, respectively.

The Gensim approach is the fastest NLP-based automated mapping method as it is advertised on its official website. Its weak capability in the word embedding is another cause for the fast speed, which results in the lowest accuracy. Therefore, the Gensim approach has little contribution to the automatic mapping.

The spaCy approach has a higher accuracy than the Gensim approach, and the run time of which increases not too much. However, when it comes to being compared with the BERT approach, its performance seems mediocre.

The accuracy of the mapping result obtained through the BERT model is the highest among the three NLP-based automated mapping methods. Whereas, the program run time of this approach is several times that of the other two, but it is still much shorter than the time of manual mapping. It can be predicted that when the BIM model becomes more complex and the volume of the to be mapped IFC data becomes enormous, mapping in this way will become a little time-consuming. Unfortunately, this scenario has a great probability to happen, on account that in practical applications,

* For Gensim, spaCy and BERT methods, the "Time" stands for the run time of the python script, which is dependent on the computer configuration.

the building designs are generally much more complex and elaborate than the BIM model in the case study.

The prominent performance of BERT can be attributed to its principle. One of the major limitations of NLP is that the standard language models are unidirectional, which restricts the power of the pre-trained representations (Devlin et al., 2019). BERT has overcome this restriction by bidirectionally training the language model, i.e., it has pre-trained deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers (Devlin et al., 2019).

In this context, for the same text, the size of the word vector obtained through BERT is much larger than other NLP tools/ pre-trained models. For instance, in this case study, the size of the word vectors of the material names encoded by Gensim or spaCy is 300, which encoded by BERT model and Transformer is 768. Hence, with the unique pre-training and encoding principle, and the consequently much larger word vector size, the semantic similarity (cosine similarity) computed based on the BERT word embeddings is much more convincing than others.

Table 4-9 Part of the similarities of mapping results computed through different NLP tools

Materials in BIM Model	Word Embedding and Similarity Computing through Gensim		Word Embedding and Similarity Computing through spaCy		Word Embedding and Similarity Computing through BERT	
	Manual Mapping result	Gensim Mapping result	Manual Mapping result	spaCy Mapping result	Manual Mapping result	BERT Mapping result
Ortbeton - C30/37	0.563370	0.741386	0.942882	0.942882	0.961197	0.961197
Ziegel mit Daemmstoff	0.682718	0.802028	-0.017330	0.877658	0.886623	0.889431
Fussboden - Teppich	0.681897	0.800892	0.476967	0.816554	0.892529	0.892529
Ortbetonestrich	[EV]	[EV]	[EV]	[EV]	0.822909	0.875407
Ziegeldach	0.528889	0.627138	0.760532	0.760532	0.841518	0.846852
Bitumen Emulsion	1.000000	1.000000	0.244121	0.776563	0.870319	0.875211
Lack	[EV]	0.656246	0.151196	0.592134	0.796803	0.796803
Daemmung - XPS	[EV]	0.766853	[EV]	0.741554	0.939645	0.939645

Table 4-9 lists part of the cosine similarities of the mapping results, including the similarities of the manual mapping result encoded and computed separately through the three NLP tools. The tables of the complete mapping results and similarities are attached in Appendix B.

The yellow marks highlight the unequal similarities, which leads to the different mapping results between the manual mapping and the mapping through NLP tools. Apart from the difference in the computed similarities, the empty vectors (EV) are another cause for the unsatisfactory performance of Gensim and spaCy. For the materials that need to be mapped, 8 out of 20 and 3 out of 20 materials respectively cannot be converted into word vectors by the Gensim and spaCy pre-trained NLP models.

To be noticed that there is a negative similarity marked in blue, which is an outlier, as the cosine similarity ought to be in the range of [0, 1]. In order to investigate the cause of this anomaly, the two materials *Ziegel mit Daemmstoff* and *Mauerziegel (Daemmstoff gefuellt)* as well as their variants were encoded, and then the cosine similarities between them were computed, both of which were conducted through the spaCy approach. The results are displayed in Table 4-10.

Table 4-10 Similarities of the two materials and their variants

	Ziegel mit Daemmstoff	Ziegel Daemmstoff
Mauerziegel (Daemmstoff gefuellt)	-0.017330	0.075202
Mauerziegel Daemmstoff gefuellt)	0.057569	0.210049
Mauerziegel (Daemmstoff gefuellt	0.066416	0.188892
Mauerziegel Daemmstoff gefuellt	0.424563	0.686186
Mauerziegel Daemmstoff	0.501110	0.884823
Ziegel Daemmstoff	0.573850	1.000000

As reflected in the table above, the spaCy pre-trained NLP model is not good at removing punctuations and stopwords in the text, which have significant impact on the result of word embedding. Thus, the similarities computed based on the word vectors vary dramatically, and the negative value is caused by the floating rounding error.

In summary, the first research question has been resolved, for that the three NLP tools have made varying degrees of contribution to the automated method of mapping LCA data to BIM models. Thereinto, BERT is the most reliable and recommendable one for assisting the automated BIM-to-LCA data mapping process, since it has the highest mapping accuracy among the three NLP-based automated mapping methods. Nevertheless, it may be a bit time-consuming with the large data volume in the modern society.

For the second research question, both the manual mapping result and the NLP-based mapping results will be input into the LCA web application eLCA respectively, based on which LCA analyses will be implemented to explore whether the automated mapping results are convincing and accurate enough compared with the manual method.

4.7 LCA Implementation, Results, and Comparison

Only the mapping results obtained manually and through the BERT-based automated method were input into eLCA to implement LCA analyses. One reason is that the mapping result obtained through the BERT approach is the most similar to the manual mapping result, i.e., the most accurate among the three NLP-based approaches. An additional reason is that there are many empty vectors in the other two NLP-based mapping results, so that the eLCA analysis cannot be performed due to the blank mapped items.

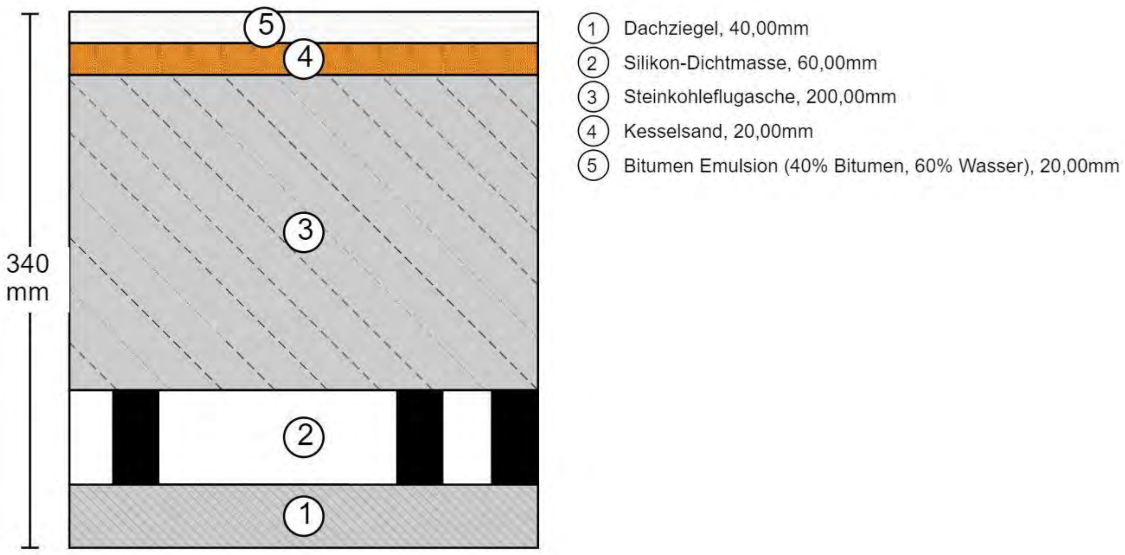

340 mm
5
4
3
2
1
1 Dachziegel, 40,00mm
2 Silikon-Dichtmasse, 60,00mm
3 Steinkohleflugasche, 200,00mm
4 Kesselsand, 20,00mm
5 Bitumen Emulsion (40% Bitumen, 60% Wasser), 20,00mm

Figure 4-4 An example of the material layers of a building element modeled in eLCA

In eLCA, the building elements were modeled following the workflow: *Materials → Building Component Layers → Building Component Parts → Building Component/ Element*. As an example, Figure 4-4 shows the construction of the building element displayed in Figure 4-2, which is composed of five material layers.

In most cases, the building element like wall and floor can be modeled correctly in eLCA, while it is ineffective for elements such as doors and windows, which usually have more complex structures rather than simply being constructed as layers. In addition, the material quantities of doors and windows commonly have little effect on the LCA result because of their relatively small volumes. Therefore, in this case study, the doors and windows were not modeled in eLCA.

Table 4-11 Values of LCA indicators obtained through eLCA

LCA Indicator	Unit	$\mathrm{total/m^2NGF^a}$ in the life cycle	
		Manual	BERT
GWP	kg CO2 equiv.	1.96045979E+02	1.96656422E+02
ODP	kg R11 equiv.	1.05455562E-08	1.05405134E-08
POCP	kg ethene equiv.	3.55110329E-02	3.61436753E-02
AP	kg SO2 eqv.	2.98992136E-01	2.98994851E-01
EP	kg PO4 equiv.	4.29325041E-02	4.28098388E-02
Total PE	MJ	3.52382815E+03	3.52073134E+03
PENRT	MJ	2.65196168E+03	2.64582828E+03
PERT	MJ	8.71866466E+02	8.74903055E+02
ADP elem.	kg Sb equiv.	1.82914469E-03	2.21211896E-03
ADP fossil	MJ	2.47545923E+03	2.46951850E+03

Table 4-11 exhibits part of the LCA indicator values during the life cycle, integrating the building life cycle stages of A1 – A3, C3, C4, and B (see Figure 2-2). The tables of complete indicators in separate building life cycle stages are attached in Appendix C. The indicators are only used to validate the effectiveness of the BERT-based automated mapping method, regardless of whether the results are reasonable and practical.

The errors of the indicators obtained based on the BERT mapping result were calculated in accordance with Formula 4.1, where *VI* represents Value of Indicator.

$$error = \frac{VI_{BERT} - VI_{manual}}{VI_{manual}} \tag{4.1}$$

Figure 4-5, 4-6 and 4-7 illustrate the errors in different life cycle stages, the values of which are attached in Appendix C as well. As shown in the figures, most LCA results based on the BERT automatic mapping have tiny errors with reference to the results obtained based on the manual mapping. However, there are several errors that are extremely large, i.e., exceeding 20% or even close to 200%, which is unacceptable. Particularly for the life cycle stage C4 – the disposal phase of the end-of-life stage, most of the indicators reveal large errors.

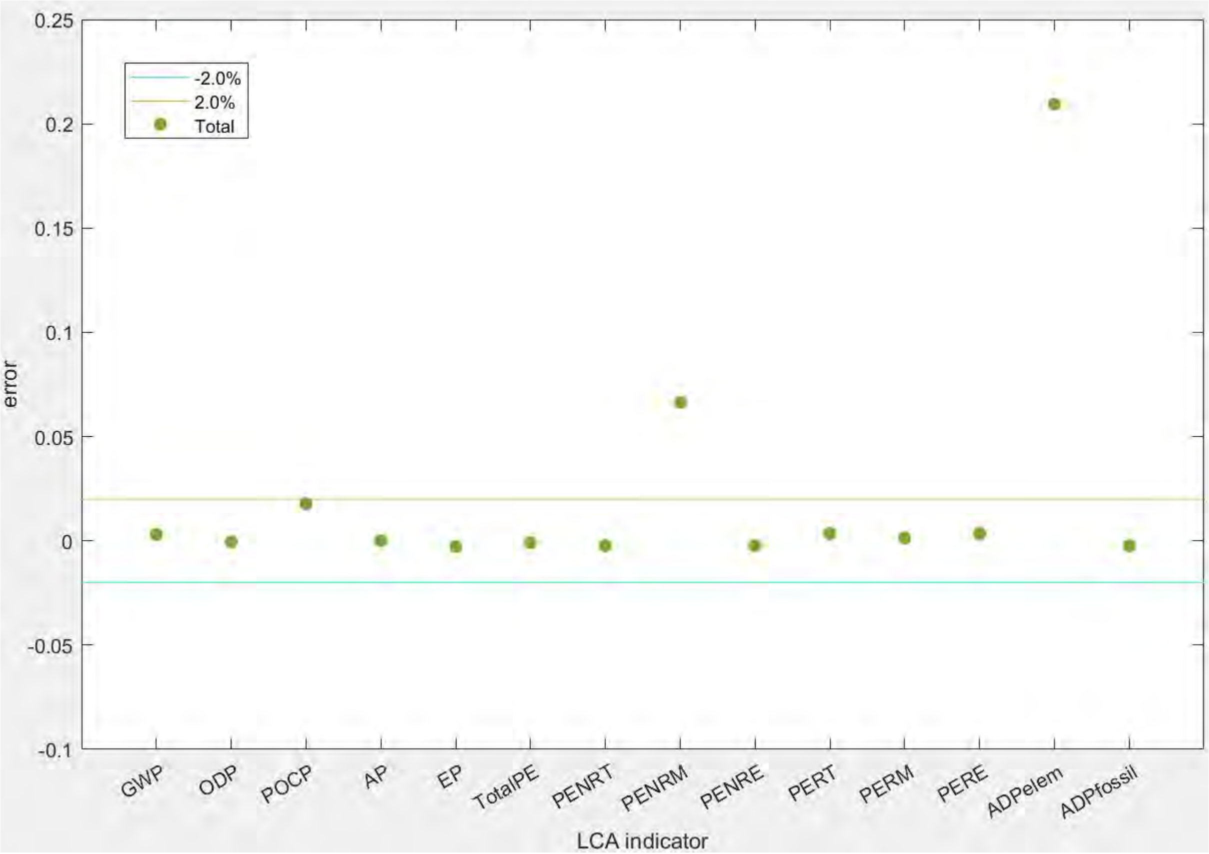


Figure 4-5 Errors of the LCA result based on BERT mapping – Total

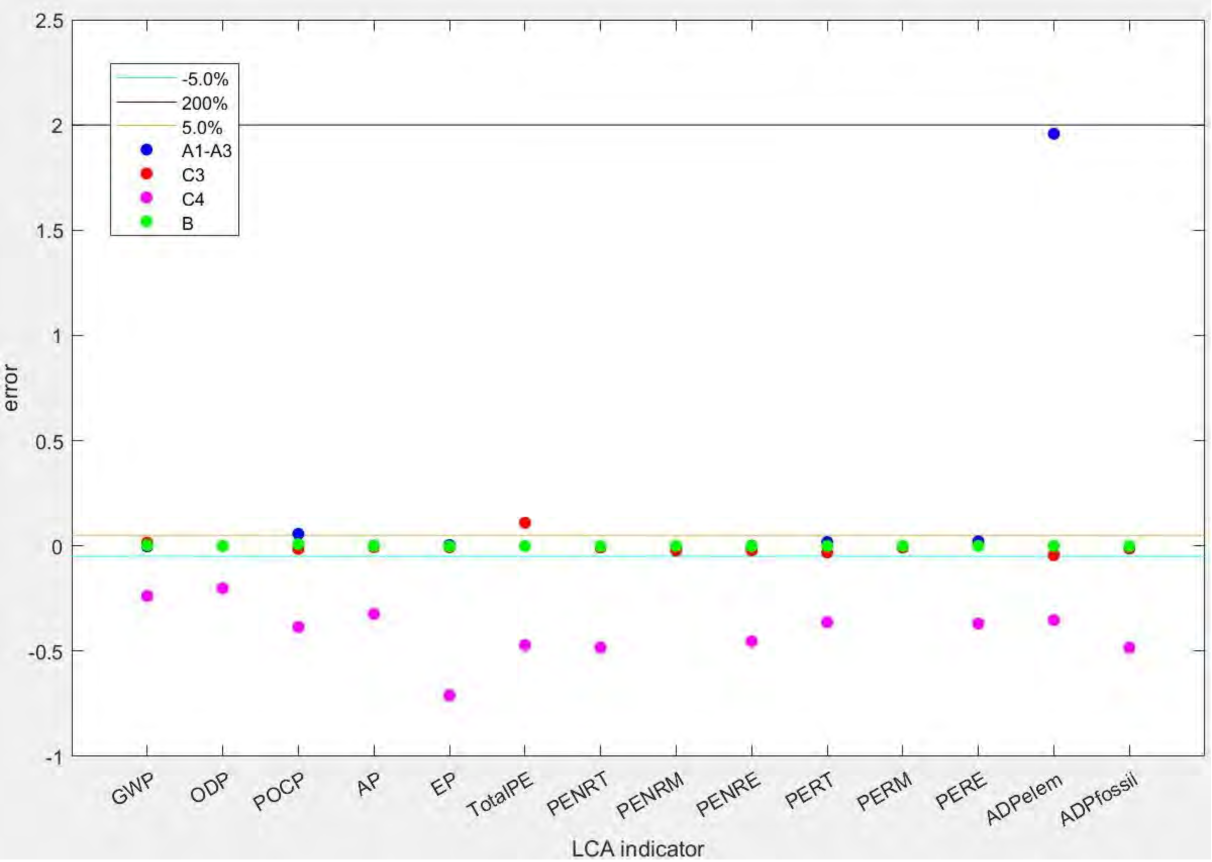


Figure 4-6 Errors of the LCA result based on BERT mapping – A1-A3, C3, C4, B

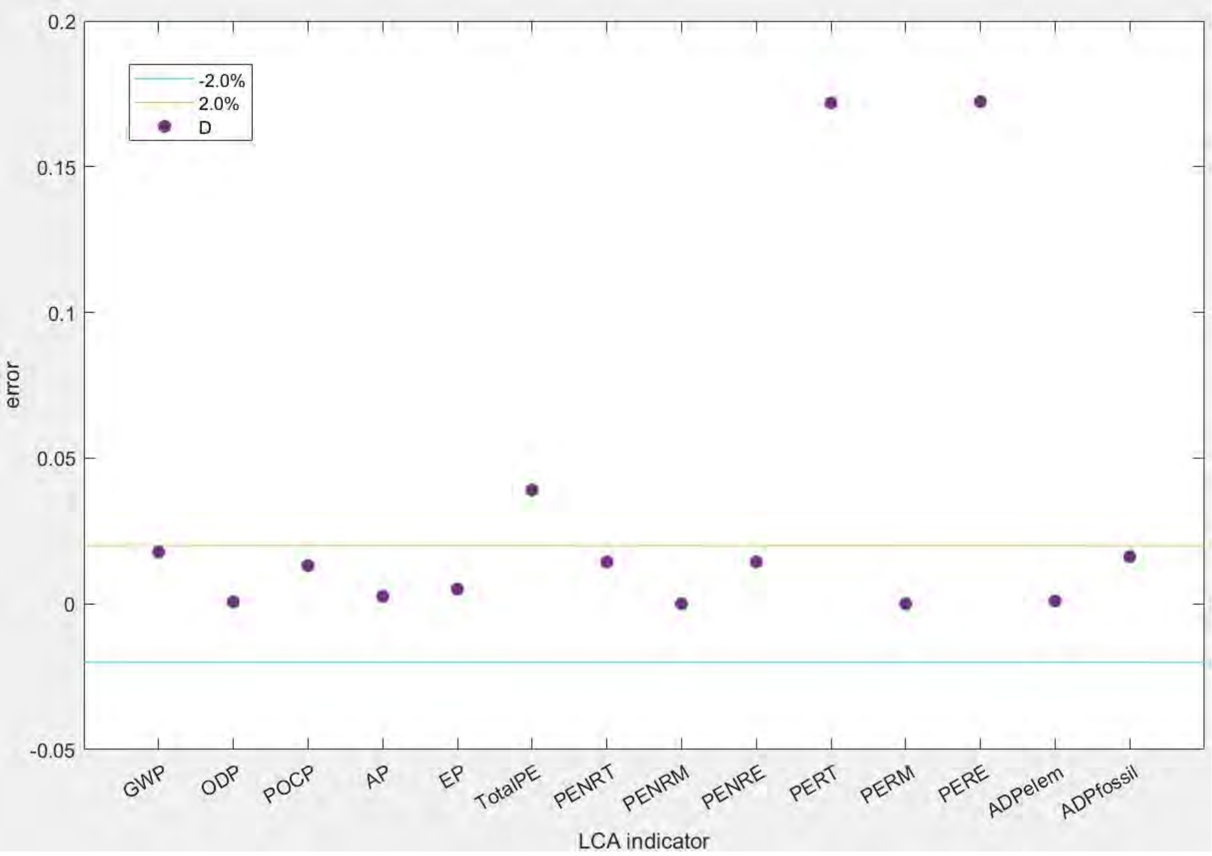


Figure 4-7 Errors of the LCA result based on BERT mapping - D

In conclusion, although BERT performs best among the three NLP tools in the automated IFC-to-LCA data mapping process, with the accuracy of 82.5%, it can only be used to conduct a rough estimation. When a reliable and accurate LCA study is required, additional manual work such as inspection or adjustment is still necessary.

5 Discussion

Even though the proposed NLP-based automated method of mapping LCA data to BIM models is proved to be promising in some sense, there are several limitations.

1. The three pre-trained NLP models employed in the case study are general models instead of models specifically trained for use in the AEC industry.

In addition to the first limitation, in the real world there is also no pre-trained NLP model prepared for practical applications in the AEC industry. For this reason, it is a great opportunity to train an NLP model with AEC industry related corpora, which is expected to fill the gap in the comprehensive automation and digitization of the building sector.

2. The datasets contained in the LCA database Ökobaudat are not complete and high-quality enough.

Ökobaudat is known as the best LCA database for being used in the AEC industry, since it owns the most abundant datasets for building products and it is constantly updated. Nevertheless, the datasets it contains are not complete and high-quality enough, which manifests in that for some materials predefined in the BIM software, the similar ones cannot be found in Ökobaudat. Besides, some data do not have English or French names, which is unfavorable for international projects. Moreover, not all data covers the entire building life cycle stages, which can lead to inaccurate or even missing result when conducting LCA analyses.

3. The BERT-based automatic mapping result is not accurate enough to independently support an accurate and reliable LCA study.

As presented in Chapter 4.7, compared with the LCA result based on manual mapping, the accuracy of BERT-based automatic mapping (82.5%) is not high enough to support downstream LCA calculations independently, which is reflected in the large errors of some LCA indicators. Thus, notwithstanding the BERT model performs well in the NLP-based automated method of mapping LCA data to BIM models, additional manual checking and adjustment are still required at present.

4. The workflow and the case study have only mapped material names and ignored other characteristics such as material properties or building components.

The level at which LCA data is available in the LCA database (e.g., EPD data for cement) is not always coincide with the level for which BIM objects are provided (e.g., IfcWall, IfcWindow) (Wastiels & Decuypere, 2019). Furthermore, LCA data is often valid for specific situations only (e.g., for a specific size of a certain building component), which might not be valid anymore after resizing the building elements in the BIM model (Wastiels & Decuypere, 2019).

The workflow and case study in this thesis have only cared about the automatic mapping according to material names, which is not sufficient for subsequent LCA calculations and analyses, since they require more elaborate information about the corresponding material properties or even the BOQ of the building elements. Therefore, mapping through other objects can be the next step, e.g., material properties/ categories, component names/ categories, and cost groups.

6 Conclusion and Outlook

Aiming to explore the potential of employing NLP technology to improve the automated method of BIM-LCA integration in the aspect of the IFC-to-LCA data mapping process, this master's thesis has resolved two research questions and made some contributions. The first is having found the way towards utilizing NLP tools to improve the automated method of mapping LCA data to BIM models. The second is having figured out whether the NLP-based automatic mapping is efficient and accurate enough to support the subsequent LCA calculations and analyses.

According to a SLR, this thesis summarized the up-to-date application of LCA, BIM, and NLP in the AEC industry, and compared three kinds of BIM-LCA integration strategies: manual, semi-automated, and automated approaches. On the basis of the state of the art, this thesis proposed an automated method of mapping LCA data to BIM models which introduces NLP technology to overcome the semantic obstacle. Simultaneously, A methodology and a prototypical workflow of this method were developed.

In order to verify the feasibility of the proposed method, a case study was implemented, which mapped the from the BIM model extracted LCA-related IFC data to the LCA database Ökobaudat respectively through manual attempt and three different NLP tools – Gensim, spaCy, and BERT. In the end, after quantitatively analyzing and comparing the contributions of the three NLP tools to the accuracy of the data mapping task, performing the automated method with the help of the pre-trained BERT model is concluded to be the most efficient and recommended with the mapping accuracy of 82.5%. However, it can be a little time-consuming with the increasing BIM model complexity and data volume, and additional manual work such as mapping result checking and adjustment is required to better support downstream LCA analyses.

This thesis does have some limitations, which were elaborated in Chapter 5. From another perspective, they can also be seen as the directions of future research.

First of all, there is no pre-trained NLP model dedicated to the AEC industry. In this vein, it has a great potential to train an NLP model which is exclusive for addressing the semantic problems to improve the automation and digitization of the BIM-based LCA study and application.

Secondly, this thesis can be extended to explore the possibility of mapping other LCA-related IFC information to the LCA database alongside the material names, such as the material properties/ categories, the building component names/ categories, and the information of cost groups. Forth et al. proposed a methodology that facilitates a multicriterial variant comparison and feedback communication of design changes, in which a standard component database that integrates all LCA-relevant information is developed, allowing certain IFC data to be automatically mapped to it (Forth et al., 2021). The automated mapping method developed by them has taken into account the mapping not only on material level but on building component level, and considered multicriteria beyond LCA (Forth et al., 2021), which has significant referent value for the future work of this thesis.

In the end, the outcome of this thesis can be further developed and optimized as a standalone BIM-to-LCA data mapping tool, or integrated into an existing BIM-based LCA plug-in or application.

References

3Units Technology. Building Information Modeling. Retrieved August 07, 2021, from https://3units.ch/en/engineering-department/progettazione-bim/

Akerkar, R. (2018). Natural language processing. *Artificial Intelligence for Business*, 53-62.

Autodesk. (2002). White Paper: Building Information Modeling.

Axis Steel Detailing, LLC. 3D, 4D, 5D, 6D, 7D – Dimensions of BIM. Retrieved September 02, 2021, from https://axissteel.com/the-dimensions-of-bim-explained/

Azizoglu, B., & Seyis, S. (2020). Analyzing the benefits and challenges of Building Information Modelling and Life Cycle Assessment integration. *Advances in Building Information Modeling*, 161–169.

Barzegar, M., Rajabifard, A., Kalantari, M., & Atazadeh, B. (2021). An IFC-based database schema for mapping BIM data into a 3D spatially enabled land administration database. *International Journal of Digital Earth, 14*(6), 736–765.

BBSR. Data exchange in the ÖKOBAUDAT environment. Retrieved August 07, 2021, from https://www.oekobaudat.de/en.html

Bernardino-Galeana, I., Llatas, C., Montes, M. V., Soust-Verdaguer, B., Canivell, J., & Meda, P. (2021). Life Cycle Cost (LCC) and sustainability. Proposal of an IFC structure to implement LCC during the design stage of buildings. *Critical Thinking in the Sustainable Rehabilitation and Risk Management of the Built Environment*, 404–426.

BibLus. (2021). The 7 dimensions of BIM - 3D, 4D, 5D, 6D, 7D BIM explained. Retrieved September 02, 2021, from https://biblus.accasoftware.com/en/bim-dimensions/

Borrmann, A., König, M., Koch, C., & Beetz, J. (2018). *Building Information Modeling: Technology foundations and industry practice*. Springer.

Briggs, J. (2021). Similarity Metrics in NLP. Retrieved November 08, 2021, from https://towardsdatascience.com/similarity-metrics-in-nlp-acc0777e234c buildingSMART International. (2021a). Industry Foundation Classes (IFC) – An Introduction. Retrieved July 17, 2021, from https://technical.buildingsmart.org/standards/ifc/

buildingSMART International. (2021b). IFC4 Documentation. Retrieved July 17, 2021, from https://standards.buildingsmart.org/IFC/RELEASE/IFC4/ADD2_TC1/HTML/

buildingSMART International. (2021c). Information Delivery Manual (IDM). Retrieved November 11, 2021, from https://technical.buildingsmart.org/standards/information-delivery-manual/

buildingSMART International. (2021d). Model View Definition (MVD) – An Introduction. Retrieved November 11, 2021, from https://technical.buildingsmart.org/standards/ifc/mvd/

buildingSMART International. (2021e). BIM Collaboration Format (BCF) – An Introduction. Retrieved November 10, 2021, from https://technical.buildingsmart.org/standards/bcf/

buildingSMART International. (2021f). IFC4 vs. IFC2x3. Retrieved November 10, 2021, from https://www.b-cert.org/Documentation/e6d094e3-7245-45e5-3154-08d500137b53

buildingSMART International. (2021g). IFC GUID. Retrieved November 21, 2021, from https://technical.buildingsmart.org/resources/ifcimplementationguidance/ifc-guid/

buildingSMART International. (2019). What is openBIM?. Retrieved November 10, 2021, from https://www.buildingsmart.org/ifc4-software-certification-delivers-first-milestone/

Bushkovskyi, O. (2019). Natural language processing tools and libraries in 2021. Retrieved November 27, 2021, from https://theappsolutions.com/blog/development/nlp-tools/#contents_6

Costa, G., & Sicilia, A. (2020). Alternatives for facilitating automatic transformation of BIM data using semantic query languages. *Automation in Construction, 120*, 103384.

Dalla Valle, A. (2021). Green buildings rating systems as driver for specific life cycle-oriented data within decision process. *Change Management Towards Life Cycle AE(C) Practice*, 79-86.

Dankers, M., Van Geel, F., & Segers, N. M. (2014). A web-platform for linking IFC to external information during the entire lifecycle of a building. *Procedia Environmental Sciences, 22*, 138-147.

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805v2 [cs.CL]

DIN 276. (2018). *Kosten im Bauwesen*.

Di Giuda, G. M., Locatelli, M., & Seghezzi, E. (2020). Natural language processing and BIM in AECO sector: A state of the art. *Proceedings of International Structural Engineering and Construction, 7*(2).

eLCA. (2021). ELCA v0.9.7 beta - Information eLCA Application. Retrieved September 11, 2021, from https://www.bauteileditor.de/information/

EN 15643-2. (2011). *Sustainability of construction works – Assessment of buildings – Part 2: Framework for the assessment of environmental performance*.

European Platform on Life Cycle Assessment. Retrieved July 20, 2021, from https://eplca.jrc.ec.europa.eu/ELCD3/

Ecoinvent version 3.7.1. Retrieved July 20, 2021, from https://www.ecoinvent.org/

Explosion. (2021). Spacy 101: Everything you need to know – spaCy usage documentation. Retrieved November 27, 2021, from https://spacy.io/usage/spacy-101

Figueiredo, K., Pierott, R., Hammad, A. W. A., & Haddad, A. (2021). Sustainable material choice for construction projects: A Life Cycle Sustainability Assessment framework based on BIM and Fuzzy-AHP. *Building and Environment, 196*, 107805.

Forth, K., Abualdenien, J., Borrmann, A., Fellermann, S., & Schunicht, C. (2021). Design optimization approach comparing multicriterial variants using BIM in early design stages. *Proceedings of the 38th International Symposium on Automation and Robotics in Construction (ISARC)*.

GaBi LCA Databases. Retrieved July 20, 2021, from https://gabi.sphera.com/international/databases/gabi-databases/

Hall, J. (2021). Top 10 benefits of BIM in construction. Retrieved August 03, 2021, from https://bim360resources.autodesk.com/connect-construct/top-10-benefits-of-bim-in-construction

Hauschild, M. Z., Rosenbaum, R. K., & Olsen, S. I. (2018). *Life Cycle Assessment: Theory and Practice*. Springer International Publishing.

Häfliger, I., John, V., Passer, A., Lasvaux, S., Hoxha, E., Saade, M. R., & Habert, G. (2017). Buildings environmental impacts' sensitivity related to LCA modelling choices of construction materials. *Journal of Cleaner Production, 156*, 805-816.

Hollberg, A., Kiss, B., Röck, M., Soust-Verdaguer, B., Wiberg, A. H., Lasvaux, S., Galimshina, A., & Habert, G. (2021). Review of visualising LCA results in the design process of buildings. *Building and Environment, 190*, 107530.

Horn, R., Ebertshäuser, S., Di Bari, R., Jorgji, O., Traunspurger, R., & Both, P. von. (2020). The BIM2LCA approach: An Industry Foundation Classes (IFC)-based interface to integrate Life Cycle Assessment in integral planning. *Sustainability, 12(16)*, 6558.

Höflich & Maier Consult GmbH. BIM - die Prozessbasierte Digitalisierung im Bereich des Bauwesens. Retrieved September 01, 2021, from https://h-m-consult.com/index.html

Hu, M. (2019). Building impact assessment - A combined Life Cycle Assessment and multi-criteria decision analysis framework. Resources, *Conservation and Recycling, 150*, 104410.

ISO 10303-11. (2004). *Industrial automation systems and integration – Product data representation and exchange – Part 11: Description methods: The EXPRESS language reference manual.*

ISO 14040. (2006). *Environmental management – Life cycle assessment – Principles and framework*.

ISO 14044. (2006). *Environmental management – Life cycle assessment – Requirements and guidelines*.

Ismail, A., Strug, B., & Ślusarczyk, G. (2018). Building Knowledge Extraction from BIM/IFC Data for Analysis in Graph Databases. *Artificial Intelligence and Soft Computing*, 652–664.

Iwaro, J., Mwasha, A., Williams, R. G., & Zico, R. (2014). An integrated criteria weighting framework for the sustainable performance assessment and design of building envelope. *Renewable and Sustainable Energy Reviews, 29*, 417-434.

Jung, N., & Lee, G. (2019). Automated Classification of Building Information Modeling (BIM) case studies by BIM use based on Natural Language Processing (NLP) and unsupervised learning. *Advanced Engineering Informatics, 41*, 100917.

Kalverkamp, M., & Karbe, N. (2018). Comparability of Life Cycle Assessments: Modelling and analyzing LCA using different databases. *Cascade Use in Technologies 2018*, 51-63.

Kanani, B. (2019). Introduction to Word Embeddings. Retrieved November 08, 2021, from https://studymachinelearning.com/introduction-to-word-embeddings/

Kanani, B. (2020a). Jaccard Similarity – Text Similarity Metric in NLP. Retrieved November 08, 2021, from https://studymachinelearning.com/jaccard-similarity-text-similarity-metric-in-nlp/

Kanani, B. (2020b). Cosine Similarity – Text Similarity Metric. Retrieved November 11, 2021, from https://studymachinelearning.com/cosine-similarity-text-similarity-metric/

Kharkovyna, O. (2021). Top 10 natural language processing (NLP) tools for beginners. Retrieved November 27, 2021, from https://medium.datadriveninvestor.com/top-10-natural-language-processing-nlp-tools-for-beginners-7e15e3042bc2

Khosakitchalert, C., Yabuki, N., & Fukuda, T. (2020). Automated modification of compound elements for accurate BIM-based quantity takeoff. *Automation in Construction, 113*, 103142.

Kim, H., Shen, Z., Kim, I., Kim, K., Stumpf, A., & Yu, J. (2016). BIM IFC information mapping to building energy analysis (BEA) model with manually extended material information. *Automation in Construction, 68*, 183-193.

Klöpffer, W. (2014). *Background and future prospects in Life Cycle Assessment*. Springer.

Koo, B., Jung, R., & Yu, Y. (2021). Automatic classification of wall and door BIM element subtypes using 3D geometric deep neural networks. *Advanced Engineering Informatics, 47*, 101200.

Krijnen, T. (2021). Built environment research and development – Projects. Retrieved December 01, 2021, from http://thomaskrijnen.com/#projects

Kumar, R. (2017). Natural language processing. *Machine Learning and Cognition in Enterprises*, 65-73.

Lawrence, M., Pottinger, R., Staub-French, S., & Nepal, M. P. (2014). Creating flexible mappings between building information models and cost information. *Automation in Construction, 45*, 107-118.

Liu, Z., Lu, Y., Shen, M., & Peh, L. C. (2021). Transition from Building Information Modeling (BIM) to Integrated Digital Delivery (IDD) in sustainable building management: A knowledge discovery approach based review. *Journal of Cleaner Production, 291*, 125223.

Lu, K., Jiang, X., Yu, J., Tam, V. W. Y., & Skitmore, M. (2021). Integration of Life Cycle Assessment and Life Cycle Cost using Building Information Modeling: A critical review. *Journal of Cleaner Production, 285*, 125438.

Manning, C. D., Raghavan, P., & Schütze, H. (2018). *Introduction to information retrieval*. Cambridge: Cambridge University Press.

Martínez-Rocamora, A., Solís-Guzmán, J., & Marrero, M. (2016). LCA databases focused on construction materials: A review. *Renewable and Sustainable Energy Reviews, 58*, 565-573.

Mc., C. (2017). A short introduction to NLP in python with spacy. Retrieved November 28, 2021, from https://towardsdatascience.com/a-short-introduction-to-nlp-in-python-with-spacy-d0aa819af3ad

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

Müller, A. (2015). Analyse von Wort-Vektoren deutscher Textkorpora. Retrieved November 11, 2021, from https://devmount.github.io/GermanWordEmbeddings/

Neto, J. (2021). Best NLP Algorithms to get Document Similarity. Retrieved November 08, 2021, from https://medium.com/analytics-vidhya/best-nlp-algorithms-to-get-document-similarity-a5559244b23b

NIBS. (2015). *National BIM Standard – United States Version 3*.

Nicholson, C. (2020). A Beginner's Guide to Word2Vec and Neural Word Embeddings. Retrieved November 08, 2021, from https://wiki.pathmind.com/word2vec

NLTK Project. (2021). NLTK - Documentation. Retrieved November 17, 2021, from https://www.nltk.org/#

Nwodo, M. N., & Anumba, C. J. (2019). A review of Life Cycle Assessment of buildings using a systematic approach. *Building and Environment, 162*, 106290.

OSArch. (2021). IFC classes. Retrieved November 21, 2021, from https://wiki.osarch.org/index.php?title=IFC_classes

Ökobaudat Sustainable Construction Information Portal. Retrieved July 25, 2021, from https://www.oekobaudat.de/en.html

Pan, Y., & Zhang, L. (2021). Roles of artificial intelligence in construction engineering and management: A critical review and future trends. *Automation in Construction, 122*, 103517.

Pauwels, P., Zhang, S., & Lee, Y. (2017). Semantic Web Technologies in AEC industry: A literature overview. *Automation in Construction, 73*, 145-165.

Petrie, R. (2021). IFC4 Software Certification Delivers First Milestone. Retrieved December 21, 2021, from https://www.buildingsmart.org/about/openbim/openbim-definition/

Potrč Obrecht, T., Röck, M., Hoxha, E., & Passer, A. (2020). BIM and LCA integration: A systematic literature review. *Sustainability*, *12*(14), 5534.

Prabhakaran, S. (2018). Cosine Similarity – Understanding the math and how it works (with python codes). Retrieved November 08, 2021, from https://www.machinelearningplus.com/nlp/cosine-similarity/

ProBas. Prozessorientierte Basisdaten für Umweltmanagementsysteme. Retrieved July 21, 2021, from https://www.probas.umweltbundesamt.de/php/index.php

Rebala, G., Ravi, A., & Churiwala, S. (2019). Natural language processing. *An Introduction to Machine Learning*, 117-125.

Řehůřek, R. (2021). Gensim: Topic modelling for humans. Retrieved November 17, 2021, from https://radimrehurek.com/gensim/#

Rezaei, F., Bulle, C., & Lesage, P. (2019). Integrating Building Information Modeling and Life Cycle Assessment in the early and detailed building design stages. *Building and Environment*, *153*, 158–167.

Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A Primer in BERTology: What We Know About How BERT Works. *Transactions of the Association for Computational Linguistics, 8*, 842-866.

Röck, M., Hollberg, A., Habert, G., & Passer, A. (2018). LCA and BIM: Visualization of environmental potentials in building construction at early design stages. *Building and Environment, 140*, 153-161.

Sacks, R., Girolami, M., & Brilakis, I. (2020). Building Information Modelling, Artificial Intelligence and Construction Tech. *Developments in the Built Environment, 4*, 100011.

Safari, K., & AzariJafari, H. (2021). Challenges and opportunities for integrating BIM and LCA: Methodological choices and framework development. *Sustainable Cities and Society, 67*, 102728.

Santos, R., Costa, A. A., Silvestre, J. D., & Pyl, L. (2019). Integration of LCA and LCC analysis within a BIM-based environment. *Automation in Construction, 103*, 127-149.

Sartori, T., Drogemuller, R., Omrani, S., & Lamari, F. (2021). A schematic framework for Life Cycle Assessment (LCA) and Green Building Rating System (GBRS). *Journal of Building Engineering, 38*, 102180.

Schneider-Marin, P., Harter, H., Tkachuk, K., & Lang, W. (2020). Uncertainty analysis of embedded energy and greenhouse gas emissions using BIM in early design stages. *Sustainability, 12*(7), 2633.

Seyis, S. (2020). Mixed method review for integrating building information modeling and life-cycle assessments. *Building and Environment, 173*, 106703.

Singh, P. (2018). Natural language processing. *Machine Learning with PySpark*, 191-218.

Song, J., Kim, J., & Lee, J. (2018). NLP and Deep Learning-based analysis of building regulations to support Automated Rule Checking System. *Proceedings of the 35th International Symposium on Automation and Robotics in Construction (ISARC)*.

Stanford NLP Group. (2020a). CoreNLP - Overview. Retrieved November 17, 2021, from https://stanfordnlp.github.io/CoreNLP/

Stanford NLP Group. (2020b). Stanza – A Python NLP Package for Many Human Languages. Retrieved November 17, 2021, from https://stanfordnlp.github.io/stanza/

Streimikiene, D., Skulskis, V., Balezentis, T., & Agnusdei, G. P. (2020). Uncertain multi-criteria sustainability assessment of green building insulation materials. *Energy and Buildings, 219*, 110021.

Theißen, S., Höper, J., Wimmer, R., Zibell, M., Meins-Becker, A., Rössig, S., . . . Lambertz, M. (2020). BIM integrated automation of whole building life cycle assessment using German LCA Data Base ökobaudat and Industry Foundation classes. *IOP Conference Series: Earth and Environmental Science, 588*(3), 032025.

Tushar, Q., Bhuiyan, M. A., Zhang, G., & Maqsood, T. (2021). An integrated approach of BIM-enabled LCA and energy simulation: The optimized solution towards sustainable development. *Journal of Cleaner Production, 289*, 125622.

United BIM. (2020). What are BIM dimensions – 3D, 4d, 5D, 6d, and 7D BIM explained: Definition & benefits. Retrieved August 02, 2021, from https://www.united-bim.com/what-are-bim-dimensions-3d-4d-5d-6d-7d-bim-explained-definition-benefits/

U.S. Life Cycle Inventory Database. Retrieved August 03, 2021, from https://www.nrel.gov/lci/

Valle, A. D. (2021). *Change management towards life cycle AE(C) practice*. Springer Nature.

Wastiels, L., & Decuypere, R. (2019). Identification and comparison of LCA-BIM integration strategies. *IOP Conference Series: Earth and Environmental Science, 323*, 012101.

WorldGBC. (2019). Bringing embodied carbon upfront.

WP-1202. (2004). *Collaboration, integrated information and the project lifecycle in building design, construction and operation*. Cincinnati, OH: Construction Users Roundtable.

Wu, C., Li, X., Guo, Y., Wang, J., Ren, Z., Wang, M., & Yang, Z. (2022). Natural language processing for smart construction: Current status and future directions. *Automation in Construction, 134*, 104059.

Wu, S., Shen, Q., Deng, Y., & Cheng, J. (2019). Natural-language-based intelligent retrieval engine for BIM object database. *Computers in Industry, 108*, 73–88.

Yalcinkaya, M., & Singh, V. (2015). Patterns and trends in building information modeling (BIM) research: A latent semantic analysis. *Automation in Construction, 59*, 68-80.

Yalçın, O. G. (2020). Natural language processing. *Applied Neural Networks with TensorFlow 2*, 187-213.

Zhang, J., & El-Gohary, N. M. (2017). Integrating semantic NLP and logic reasoning into a unified system for fully-automated code checking. *Automation in Construction, 73*, 45-57.

Appendix A: IFC Data

Table A Extracted, split, condensed and selected IFC data

GUID	IFC Class	Cost Group	Component Name	Material Name
1eEeqlWvH0$uJ0CRYFb9GD	IfcWall	331 Tragende Außenwände	Basiswand: STB+WD 200+200	HPL-Platte
1eEeqlWvH0$uJ0CRYFb9GD	IfcWall	331 Tragende Außenwände	Basiswand: STB+WD 200+200	Ortbeton - C30/37
1eEeqlWvH0$uJ0CRYFb9iT	IfcWall	331 Tragende Außenwände	Basiswand: Ziegel 300	Ziegel mit Daemmstoff
2CJ9v0fLT2QB1aN$GEssv5	IfcWall	341 Tragende Innenwände	Basiswand: KS 240	Kalksandstein
2CJ9v0fLT2QB1aN$GEssxk	IfcWall	341 Tragende Innenwände	Basiswand: GK 200	Gipskartonplatte (Feuerschutz) 0,0125 m
2CJ9v0fLT2QB1aN$GEssxk	IfcWall	341 Tragende Innenwände	Basiswand: GK 200	Gipskartonplatte (impraegniert) 0,0125 m
2CJ9v0fLT2QB1aN$GEssxk	IfcWall	341 Tragende Innenwände	Basiswand: GK 200	Steinkohleflugasche
3Froekll18vhjGoGPvxx4u	IfcSlab	351 Deckenkonstruktionen	Geschossdecke: STB 300	Walzplattiertes Grobblech
3Froekll18vhjGoGPvxx4u	IfcSlab	351 Deckenkonstruktionen	Geschossdecke: STB 300	Doppelbodensystem Typ LIGNA
3Froekll18vhjGoGPvxxBT	IfcSlab	351 Deckenkonstruktionen	Geschossdecke: FB 180 Teppich	Fussboden - Teppich
3Froekll18vhjGoGPvxxBT	IfcSlab	351 Deckenkonstruktionen	Geschossdecke: FB 180 Teppich	Ortbetonestrich
3Froekll18vhjGoGPvxxBT	IfcSlab	351 Deckenkonstruktionen	Geschossdecke: FB 180 Teppich	Keramische Fliesen
3Froekll18vhjGoGPvxxBT	IfcSlab	351 Deckenkonstruktionen	Geschossdecke: FB 180 Teppich	Nadelschnittholz - getrocknet

3FroekIl18vhjGoGPvxxBT	IfcSlab	351 Deckenkonstruktionen	Geschossdecke: FB 180 Teppich	Schuettung aus Polystyrolschaumstoff-Partikeln
0u48trA2DAOvk$adJTFrkA	IfcSlab	351 Deckenkonstruktionen	Geschossdecke: FB 180 Parkett	Mehrschichtparkett
0u48trA2DAOvk$adJTFrkA	IfcSlab	351 Deckenkonstruktionen	Geschossdecke: FB 180 Parkett	Transportbeton C20/25
0u48trA2DAOvk$adJTFrkA	IfcSlab	351 Deckenkonstruktionen	Geschossdecke: FB 180 Parkett	Kautschuk-Dichtmasse
0u48trA2DAOvk$adJTFrkA	IfcSlab	351 Deckenkonstruktionen	Geschossdecke: FB 180 Parkett	Holzwolle-Leichtbauplatte
3QyxzDci96i9xqU0xkN8jQ	IfcWall	331 Tragende Außenwände	Basiswand: LB 20	Transparente Platten PC
0OEkutEW5EMeg5GLjflcAQ	IfcRoof	300 Bauwerk Baukonstruktionen	Basisdach: Ziegeldach 340	Ziegeldach
0OEkutEW5EMeg5GLjflcAQ	IfcRoof	300 Bauwerk Baukonstruktionen	Basisdach: Ziegeldach 340	Silikon-Dichtmasse
0OEkutEW5EMeg5GLjflcAQ	IfcRoof	300 Bauwerk Baukonstruktionen	Basisdach: Ziegeldach 340	Kesselsand
0OEkutEW5EMeg5GLjflcAQ	IfcRoof	300 Bauwerk Baukonstruktionen	Basisdach: Ziegeldach 340	Bitumen Emulsion
0OEkutEW5EMeg5GLjflc5p	IfcRoof	300 Bauwerk Baukonstruktionen	Basisdach: KLH 200	Mehrschichtige Massivholzplatte
0OEkutEW5EMeg5GLjfld_E	IfcPlate	346 Elementierte Innenwandkonstruktionen	Systemelement: Dachverglasung	Glas - Sicherheitsverglasung
1rqGkujSH97xGP0fn8zaQL	IfcDoor	344 Innenwandöffnungen	TU DF 1 - Rahmenstock mit Glaseinteil: DL - 800 x 2000 - MB 1400	Lack
1rqGkujSH97xGP0fn8zaxR	IfcWindow	344 Innenwandöffnungen	TU DF 1 - Rahmenstock mit Glaseinteil: DL - 800 x 2000 - MB 1400	Glas - Isolierverglasung
2JlUyU8hP689vZx69$7kcL	IfcSlab	351 Deckenkonstruktionen	Geschossdecke: WD 300	Kunststoffprofil SBR
2JlUyU8hP689vZx69$7kcL	IfcSlab	351 Deckenkonstruktionen	Geschossdecke: WD 300	Polycarbonatplatte
2JlUyU8hP689vZx69$7kqm	IfcWall	331 Tragende Außenwände	Basiswand: STB 100	Ortbeton - C20/25
2JlUyU8hP689vZx69$7kqm	IfcWall	331 Tragende Außenwände	Basiswand: STB 100	Huettensteine

2JlUyU8hP689vZx69$7kqm	IfcWall	331 Tragende Außenwände	Basiswand: STB 100	Dachsteine
2EEBr4nBD5mQsjIRgiiFe7	IfcWall	331 Tragende Außenwände	Basiswand: STB 200	Feuerverzinktes Stahlband
2EEBr4nBD5mQsjIRgiiFe7	IfcWall	331 Tragende Außenwände	Basiswand: STB 200	Daemmung - XPS
2rtQ5uukz4NBMM$uyTMi1R	IfcWall	342 Nichttragende Innenwände	Basiswand: GK VS 150	Asphalttragschicht
2rtQ5uukz4NBMM$uyTMi1R	IfcWall	342 Nichttragende Innenwände	Basiswand: GK VS 150	Lehmplatte
2rtQ5uukz4NBMM$uyTMiRd	IfcWall	342 Nichttragende Innenwände	Basiswand: GK 100	Argeton
2rtQ5uukz4NBMM$uyTMiWB	IfcWall	342 Nichttragende Innenwände	Basiswand: GK 100_	BIMs Schotter
2rtQ5uukz4NBMM$uyTMiWB	IfcWall	342 Nichttragende Innenwände	Basiswand: GK 100_	Zementbauplatte
2rtQ5uukz4NBMM$uyTMiWB	IfcWall	342 Nichttragende Innenwände	Basiswand: GK 100_	Melaminharz-Schaum

Appendix B: Mapping Results and Similarities

Table B-1 Mapping results obtained through manual method and different NLP tools

Materials in BIM Model	Mapped Materials in LCA Database Ökobaudat			
Extracted Material Name	**Manual Mapping**	**Gensim**	**spaCy**	**BERT**
HPL-Platte	HPL-Platte	HPL-Platte	HPL-Platte	HPL-Platte
Ortbeton - C30/37	Transportbeton C30/37	Zementmoertel	Transportbeton C30/37	Transportbeton C30/37
Ziegel mit Daemmstoff	Mauerziegel (Daemmstoff gefuellt)	Dachziegel	Betonpflasterstein mit Edelsplittvorsatzbeton	XPS-Daemmstoff
Kalksandstein	Kalksandstein	Kalksandstein	Kalksandstein	Kalksandstein
Gipskartonplatte (Feuerschutz) 0,0125 m	Gipskartonplatte (Feuerschutz) (Dicke 0,0125 m)	[EV]	Gipskartonplatte (Feuerschutz) (Dicke 0,0125 m)	Gipskartonplatte (Feuerschutz) (Dicke 0,0125 m)
Gipskartonplatte (impraegniert) 0,0125 m	Gipskartonplatte (impraegniert) (Dicke 0,0125 m)	[EV]	Gipskartonplatte (Feuerschutz) (Dicke 0,0125 m)	Gipskartonplatte (impraegniert) (Dicke 0,0125 m)
Steinkohleflugasche	Steinkohleflugasche	Steinkohleflugasche	Steinkohleflugasche	Steinkohleflugasche
Walzplattiertes Grobblech	Walzplattiertes Grobblech	Walzplattiertes Grobblech	Walzplattiertes Grobblech	Walzplattiertes Grobblech
Doppelbodensystem Typ LIGNA	Doppelbodensystem Typ LIGNA	Doppelbodensystem Typ LIGNA	Doppelbodensystem Typ LIGNA	Doppelbodensystem Typ LIGNA
Fussboden - Teppich	Fussbodenbelag mehrschichtiges Nadelvlies (Teppichboden, 1400 g/m^2)	Decke - HT Labor+Hospitaltechnik GmbH - Decke fuer Hygienebereiche	Flachglas - Guardian - Flachglas unbeschichtet	Fussbodenbelag mehrschichtiges Nadelvlies (Teppichboden, 1400 g/m^2)

Ortbetonestrich	Zementestrich	[EV]	[EV]	Faserzementplatte (Fassade)
Keramische Fliesen	Keramische Fliesen und Platten	Keramische Fliesen und Platten	Keramische Fliesen und Platten	Keramische Fliesen und Platten
Nadelschnittholz - getrocknet	Nadelschnittholz - getrocknet (Durchschnitt DE)	[EV]	Flachglas - Guardian - Flachglas unbeschichtet	Nadelschnittholz - getrocknet (Durchschnitt DE)
Schuettung aus Polystyrol-schaumstoff-Partikeln	Schuettung aus Polystyrol-schaumstoff-Partikeln (ohne Bindemittel)	Schuettung aus Polystyrol-schaumstoff-Partikeln (ohne Bindemittel)	Fernwaerme aus Abfaellen	Schuettung aus Polystyrol-schaumstoff-Partikeln (zementgebunden)
Mehrschichtparkett	Mehrschichtparkett	Mehrschichtparkett	Mehrschichtparkett	Mehrschichtparkett
Transportbeton C20/25	Transportbeton C20/25	Transportbeton C20/25	Transportbeton C20/25	Transportbeton C20/25
Kautschuk-Dichtmasse	Kautschuk-Dichtmasse	Kautschuk-Dichtmasse	Kautschuk-Dichtmasse	Kautschuk-Dichtmasse
Holzwolle-Leichtbauplatte	Holzwolle-Leichtbauplatte	Holzwolle-Leichtbauplatte	Holzwolle-Leichtbauplatte	Holzwolle-Leichtbauplatte
Transparente Platten PC	Transparente Platten PC	Transparente Platten PC	Transparente Platten PC	Transparente Platten PC
Ziegeldach	Dachziegel	Lehmputz	Dachziegel	Vormauerziegel
Silikon-Dichtmasse	Silikon-Dichtmasse	Silikon-Dichtmasse	Silikon-Dichtmasse	Silikon-Dichtmasse
Kesselsand	Kesselsand	Kesselsand	Kesselsand	Kesselsand
Bitumen Emulsion	Bitumen Emulsion (40% Bitumen, 60% Wasser)	Bitumen Emulsion (40% Bitumen, 60% Wasser)	Acrylat Dichtmasse	Extrudierter Polystyrol Daemmstoff (XPS)
Mehrschichtige Massivholzplatte	3- und 5-Schicht Massivholzplatte (Durchschnitt DE)	[EV]	binderholz Massivholzplatte	binderholz Massivholzplatte
Glas - Sicherheitsverglasung	Flachglas - Guardian - Verbundsicherheitsglas (VSG)	Glas - Promat GmbH - Promaglas G30 type 1	Flachglas - Guardian - Flachglas unbeschichtet	Sonnenschutzlammellen Metall
Lack	Lacksysteme Holzfassade deckend (Decklacksystem)	Folie aus Polytetrafluorethylen (PTFE)	Innenfarbe Dispersionsfarbe scheuerfest	Lacksysteme Holzfassade deckend (Decklacksystem)
Glas - Isolierverglasung	Isolierglas 2-Scheiben	Glas - Promat GmbH - Promaglas G30 type 1	Flachglas - Guardian - Flachglas beschichtet	Isolierglas 2-Scheiben

Kunststoffprofil SBR	Kunststoffprofil SBR	Kunststoffprofil SBR	Kunststoffprofil SBR	Kunststoffprofil SBR
Polycarbonatplatte	Polycarbonatplatte	Polycarbonatplatte	Polycarbonatplatte	Polycarbonatplatte
Ortbeton - C20/25	Transportbeton C20/25	Zementmoertel	Transportbeton C20/25	Transportbeton C20/25
Huettensteine	Huettensteine	Huettensteine	Huettensteine	Huettensteine
Dachsteine	Dachsteine	Dachsteine	Dachsteine	Dachsteine
Feuerverzinktes Stahlband	Feuerverzinktes Stahlband	Feuerverzinktes Stahlband	Feuerverzinktes Stahlband	Feuerverzinktes Stahlband
Daemmung - XPS	XPS-Daemmstoff	Fussbodenheizung PEX (100 mm Abstand)	Flachglas - Guardian - Flachglas unbeschichtet	XPS-Daemmstoff
Asphalttragschicht	Asphalttragschicht	Asphalttragschicht	Asphalttragschicht	Asphalttragschicht
Lehmplatte	Lehmbauplatte	[EV]	[EV]	Lehmbauplatte
Argeton	Argeton	Argeton	Argeton	Argeton
BIMs Schotter	Bims Schotter	[EV]	Bims Schotter	Bims Schotter
Zementbauplatte	Zementgebundene Spanplatte	[EV]	[EV]	Zementgebundene Spanplatte
Melaminharz-Schaum	Melaminharz-Schaum	Melaminharz-Schaum	Melaminharz-Schaum	Melaminharz-Schaum

Table B-2 Similarities of mapping results computed using different NLP tools

Word Embedding and Similarity Computing through Gensim		Word Embedding and Similarity Computing through spaCy		Word Embedding and Similarity Computing through BERT	
Manual Mapping Result	**Gensim Mapping Result**	**Manual Mapping Result**	**spaCy Mapping Result**	**Manual Mapping Result**	**BERT Mapping Result**
[EV]	[EV]	[EV]	[EV]	1.000000	1.000000
0.563370	0.741386	0.942882	0.942882	0.961197	0.961197
0.682718	0.802028	-0.017330	0.877658	0.886623	0.889431
1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
[EV]	[EV]	0.994589	0.994589	0.985597	0.985597
[EV]	[EV]	0.974647	1.000000	0.983678	0.983678
[EV]	[EV]	[EV]	[EV]	1.000000	1.000000
[EV]	[EV]	[EV]	[EV]	1.000000	1.000000
[EV]	[EV]	1.000000	1.000000	1.000000	1.000000
0.681897	0.800892	0.476967	0.816554	0.892529	0.892529
[EV]	[EV]	[EV]	[EV]	0.822909	0.875407
1.000000	1.000000	0.861328	0.861328	0.965129	0.965129
[EV]	[EV]	0.732391	0.900585	0.935419	0.935419
1.000000	1.000000	0.288122	1.000000	0.973294	0.979074
[EV]	[EV]	1.000000	1.000000	1.000000	1.000000
1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
[EV]	[EV]	[EV]	[EV]	1.000000	1.000000
[EV]	[EV]	[EV]	[EV]	1.000000	1.000000
1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
0.528889	0.627138	0.760532	0.760532	0.841518	0.846852

[EV]	[EV]	[EV]	[EV]	1.000000	1.000000
[EV]	[EV]	[EV]	[EV]	1.000000	1.000000
1.000000	1.000000	0.244121	0.776563	0.870319	0.875211
[EV]	[EV]	0.090671	0.811645	0.867038	0.912543
0.489599	1.000000	0.718521	0.878769	0.862980	0.885223
[EV]	0.656246	0.151196	0.592134	0.796803	0.796803
0.569636	1.000000	0.573187	0.865301	0.918573	0.918573
[EV]	[EV]	1.000000	1.000000	1.000000	1.000000
[EV]	[EV]	[EV]	[EV]	1.000000	1.000000
0.563370	0.741386	0.950391	0.950391	0.965844	0.965844
[EV]	[EV]	[EV]	[EV]	1.000000	1.000000
1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
[EV]	[EV]	1.000000	1.000000	1.000000	1.000000
[EV]	0.766853	[EV]	0.741554	0.939645	0.939645
1.000000	1.000000	[EV]	[EV]	1.000000	1.000000
[EV]	[EV]	[EV]	[EV]	0.953273	0.953273
[EV]	[EV]	[EV]	[EV]	1.000000	1.000000
[EV]	[EV]	0.823791	0.823791	0.906689	0.906689
[EV]	[EV]	[EV]	[EV]	0.926561	0.926561
[EV]	[EV]	[EV]	[EV]	1.000000	1.000000

Appendix C: LCA Results and Errors

Table C-1 Values of LCA indicators obtained through eLCA based on the manual mapping result

LCA Indicator	Unit	$\mathrm{total/m^2NGF^a}$ in life cycle stages					
		Total	A1-A3	C3	C4	B	D
GWP	kg CO2 equiv.	1.96045979E+02	3.27762947E+01	1.90695469E+01	1.75465449E-01	1.44024672E+02	-7.63865027E+00
ODP	kg R11 equiv.	1.05455562E-08	9.77920698E-09	1.23510162E-10	3.76784843E-14	6.42801333E-10	1.25786218E-09
POCP	kg ethene equiv.	3.55110329E-02	7.12199392E-03	2.38573161E-04	3.85789823E-05	2.81118868E-02	-1.04825164E-03
AP	kg SO2 eqv.	2.98992136E-01	6.49894016E-02	2.95424542E-03	4.65354193E-04	2.30583135E-01	-8.93762653E-03
EP	kg PO4 equiv.	4.29325041E-02	1.03533768E-02	6.30046906E-04	1.37976870E-04	3.18111035E-02	-1.17749661E-03
Total PE	MJ	3.52382815E+03	8.61643761E+02	1.04582746E+00	1.47994602E+00	2.65965861E+03	-1.37846483E+02
PENRT	MJ	2.65196168E+03	6.86191867E+02	6.31671809E+00	1.33109496E+00	1.95812200E+03	-1.16275928E+02
PENRM	MJ	6.40764269E-02	1.84980444E+02	-1.84794930E+02	-1.21437909E-01	0.00000000E+00	0.00000000E+00
PENRE	MJ	2.65186593E+03	5.01212066E+02	1.91111403E+02	1.42046119E+00	1.95812200E+03	-1.16114280E+02
PERT	MJ	8.71866466E+02	1.75451894E+02	-5.27089062E+00	1.48851060E-01	7.01536612E+02	-2.15705548E+01
PERM	MJ	1.59285599E+00	2.20993467E+01	-2.05064907E+01	0.00000000E+00	0.00000000E+00	0.00000000E+00
PERE	MJ	8.70203185E+02	1.53352389E+02	1.51679989E+01	1.46185585E-01	7.01536612E+02	-2.15129461E+01
ADP elem.	kg Sb equiv.	1.82914469E-03	1.95595252E-04	1.60683061E-07	1.01262401E-08	1.63337863E-03	-2.53925659E-06
ADP fossil	MJ	2.47545923E+03	6.46177511E+02	5.93968128E+00	1.28874723E+00	1.82205329E+03	-1.06899451E+02

Table C-2 Values of LCA indicators obtained through eLCA based on the BERT automated mapping result

LCA Indicator	Unit	total/m^2NGF[a] in life cycle stages					
		Total	A1-A3	C3	C4	B	D
GWP	kg CO2 equiv.	1.96656422E+02	3.26872302E+01	1.93670680E+01	1.33733022E-01	1.44468391E+02	-7.77465718E+00
ODP	kg R11 equiv.	1.05405134E-08	9.77421571E-09	1.23467272E-10	3.01101490E-14	6.42800258E-10	1.25870927E-09
POCP	kg ethene equiv.	3.61436753E-02	7.52924601E-03	2.35225701E-04	2.37165394E-05	2.83554870E-02	-1.06201165E-03
AP	kg SO2 eqv.	2.98994851E-01	6.50985432E-02	2.93566725E-03	3.14920569E-04	2.30645720E-01	-8.96071485E-03
EP	kg PO4 equiv.	4.28098388E-02	1.04000119E-02	6.25828385E-04	4.00007794E-05	3.17439978E-02	-1.18347207E-03
Total PE	MJ	3.52073134E+03	8.61241365E+02	1.16092322E+00	7.82816477E-01	2.65754623E+03	-1.43224419E+02
PENRT	MJ	2.64582828E+03	6.82668726E+02	6.26653708E+00	6.87985746E-01	1.95620503E+03	-1.17946428E+02
PENRM	MJ	6.83304088E-02	1.80988785E+02	-1.80799016E+02	-1.21437909E-01	0.00000000E+00	0.00000000E+00
PENRE	MJ	2.64572828E+03	5.01680585E+02	1.87065309E+02	7.77351967E-01	1.95620503E+03	-1.17784780E+02
PERT	MJ	8.74903055E+02	1.78572639E+02	-5.10561386E+00	9.48307310E-02	7.01341199E+02	-2.52779906E+01
PERM	MJ	1.59504764E+00	2.19387952E+01	-2.03437476E+01	0.00000000E+00	0.00000000E+00	0.00000000E+00
PERE	MJ	8.73237582E+02	1.56633685E+02	1.51705325E+01	9.21652557E-02	7.01341199E+02	-2.52203820E+01
ADP elem.	kg Sb equiv.	2.21211896E-03	5.78595489E-04	1.53609678E-07	6.55727308E-09	1.63336330E-03	-2.54167504E-06
ADP fossil	MJ	2.46951850E+03	6.42807075E+02	5.86680382E+00	6.65148897E-01	1.82017948E+03	-1.08628049E+02

Table C-3 Errors of the LCA result based on BERT mapping compared with the manual result

LCA Indicator	Error					
	Total	A1-A3	C3	C4	B	D
GWP	0.00311377	-0.00271735	0.01560189	-0.23783843	0.00308085	0.01780510
ODP	-0.00047819	-0.00051040	-0.00034726	-0.20086624	-0.00000167	0.00067344
POCP	0.01781538	0.05718231	-0.01403117	-0.38524715	0.00866538	0.01312663
AP	0.00000908	0.00167938	-0.00628863	-0.32326694	0.00027142	0.00258327
EP	-0.00285717	0.00450434	-0.00669557	-0.71009069	-0.00210951	0.00507471
Total PE	-0.00087882	-0.00046701	0.11005233	-0.47105066	-0.00079423	0.03901395
PENRT	-0.00231278	-0.00513434	-0.00794416	-0.48314300	-0.00097898	0.01436669
PENRM	0.06638919	-0.02157882	-0.02162350	0.00000000	0.00000000	0.00000000
PENRE	-0.00231447	0.00093477	-0.02117139	-0.45274677	-0.00097898	0.01438669
PERT	0.00348286	0.01778690	-0.03135652	-0.36291531	-0.00027855	0.17187485
PERM	0.00137592	-0.00726499	-0.00793618	0.00000000	0.00000000	0.00000000
PERE	0.00348700	0.02139710	0.00016704	-0.36953253	-0.00027855	0.17233511
ADP elem.	0.20937342	1.95812646	-0.04402071	-0.35244741	-0.00000938	0.00095243
ADP fossil	-0.00239985	-0.00521596	-0.01226959	-0.48387947	-0.00102841	0.01617032

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Bachelor-Thesis selbstständig angefertigt habe. Es wurden nur die in der Arbeit ausdrücklich benannten Quellen und Hilfsmittel benutzt. Wörtlich oder sinngemäß übernommenes Gedankengut habe ich als solches kenntlich gemacht.

Ich versichere außerdem, dass die vorliegende Arbeit noch nicht einem anderen Prüfungsverfahren zugrunde gelegen hat.

[redacted] [redacted]

Vorname Nachname

[redacted]