Domain-specific fine-tuning of LLM for material matching of BIM elements and Material Passports

Kasimir Forth, S.M.ASCE,[1] Patrick Berggold[2], and André Borrmann, Prof. Dr.-Ing.[3]

[1]Chair of Computational Modeling and Simulation, TUM School of Engineering & Design, Technical University of Munich, Germany (corresponding author). e-mail: kasimir.forth@tum.de
[2]Chair of Computational Modeling and Simulation, TUM School of Engineering & Design, Technical University of Munich, Germany. e-mail: patrick.berggold@tum.de
[3]Full Professor, Chair of Computational Modeling and Simulation, TUM School of Engineering & Design, Technical University of Munich, Germany. e-mail: andre.borrmann@tum.de

ABSTRACT

Material Passports (MP) enable a combined assessment of life cycle assessments and circularity assessment of buildings. Semantically rich 3D models, such as Building Information Models (BIM), facilitate deriving consistent and automated creation of MPs. Nevertheless, a time-consuming effort is still needed to manually match material and element information to automate the BIM-based MP. To improve this step, we propose a method of semi-automatically matching BIM materials to the relevant material datasets using Semantic Textual Similarity and fine-tuning pre-trained Large Language Models (LLM). The method matches the semantically most similar environmental material datasets to every BIM material to enrich further information. We are fine-tuning the LLM by proposing different strategies, such as adding domain knowledge, testing different loss functions, applying different labeling, adding negative pairs or filtering, and using manually matched pairs of datasets from 23 real-world case studies. Combining different strategies for fine-tuning a pre-trained LLM significantly increases the accuracy of the proposed method of matching BIM elements and materials to environmental material datasets.

INTRODUCTION

In 2020, buildings and the construction industry were responsible for 36% of the Greenhouse gas (GHG) emissions as well as for ca. 37.5 % of the waste generation within the European Union (European Commission 2020). To tackle the insufficient documentation for realizing the reuse and recycling of buildings, the concept of material passports (MP) is introduced in different scales, such as material, product, or building (Çetin et al. 2023). Building Information Models (BIM) contain geometric and semantic information about buildings and can facilitate MPs (Honic et al. 2019). Besides precise quantity take-offs, further semantic information about the elements' layers, materials, and detachability can be included. However, manual steps are still required to enrich materials from circularity databases to those used in BIM (Honic et al. 2019), as architectural nomenclature differs from the more precise databases. These manual enrichments are expensive in costs and labor. We define this automated enrichment step as the primary technology gap to be addressed by this publication.

To close this gap, we propose a novel method to automatically enrich open BIM models with material information from Life Cycle Assessment (LCA) and circularity databases using Natural Language Processing (NLP) and its subtask Semantic Textual Similarity (STS). Usually, the naming of BIM materials is more generic, e.g., "pre-cast concrete", while datasets for MPs are more specific, e.g., "reinforced concrete" with specific compressive strengths. A previous publication showed a similar approach using a well-structured database and a pre-trained Large Language Model (LLM) for LCA (Forth et al. 2023). In this publication, though, we propose a domain-specific fine-tuning of pre-trained LLM using different strategies for this task. These include domain-specific abbreviations, loss functions, and additional information from the BIM model. Our method is based on open BIM data formats like Industry Foundation Classes (IFC).

BACKGROUND AND RELATED WORKS

BIM for Material Passports. Recently, different researchers have proposed BIM-based methods for material passports (MP). The findings suggest that LCA-based BIM plugins have significant potential for improving circularity in early design stages but emphasize the importance of data accuracy, effective management, clear guidance for modeling, and increased knowledge in implementing LCA and circular economy concepts. Honic et al. introduced a BIM-based MP approach to optimize the recyclability of buildings. However, they identified the manual material matching by a specialist as a significant obstacle (Honic et al. 2019). Atta et al. developed a framework for digital MPs using BIM, considering the deconstructability of elements (Atta et al. 2021). However, their approach is based on the BIM authoring tool Autodesk Revit and is limited to its closed BIM workflows. Gebetsroither et al. compared current BIM-based approaches for building Material passports mainly in the German-speaking market (Gebetsroither et al. 2024). They came to the conclusion that the approach by Madaster and from EPEA is currently practicable, and the BIM integration not only saves time but also supports the documentation and archiving of the building. The discussed approaches lack open BIM data exchange and a fully automated process of matching material datasets from external databases to those of the BIM model. The detachability, deconstructability, or connection types of elements for circularity assessments are out of the scope of this publication and part of future research.

Fine-tuning Large Language Models. As most large language models (LLM) were trained on generic text, they do not always fit well into domain-specific tasks. Accordingly, domain adaptation needs to be applied for domain-specific use cases. Usually, domain adaptation is fine-tuning a pre-trained language model (PLM) on a domain-specific, new dataset. This fine-tuning process adjusts the original model's weights, aligning them with the specific attributes of the domain data and the targeted task. Reimers and Gurevych present Sentence-BERT (SBERT), “a modification of the pre-trained BERT network that uses Siamese and triplet network structures to derive semantically meaningful sentence embeddings that can be compared using cosine-similarity" (Reimers and Gurevych 2019). Their approach focuses on semantic textual similarity (STS) and outperforms other sentence embedding methods. For improving the performance of fine-tuning BERT in a multitask domain, such as sentiment analysis, paraphrase detection, and STS, Jadwin and Huang employed an in-domain pre-training and Multiple Negative Ranking Loss Learning (MNRL) (Jadwin and Huang 2023). They concluded that MNRL fine-tuning has the highest impact on performance optimization. Sachidananda introduced adaptive tokenization (AT), a method for efficiently adapting PLMs to new domains by expanding the tokenization vocabulary with domain-specific token sequences (Sachidananda et al. 15.09.2021). AT achieves significant performance improvements without requiring further language model pre-training, offering a promising approach for domain adaptation in natural language processing tasks. Generally, these methods show different approaches for domain adaptation and fine-tuning of pre-trained language models, which will be further discussed later.

METHOD

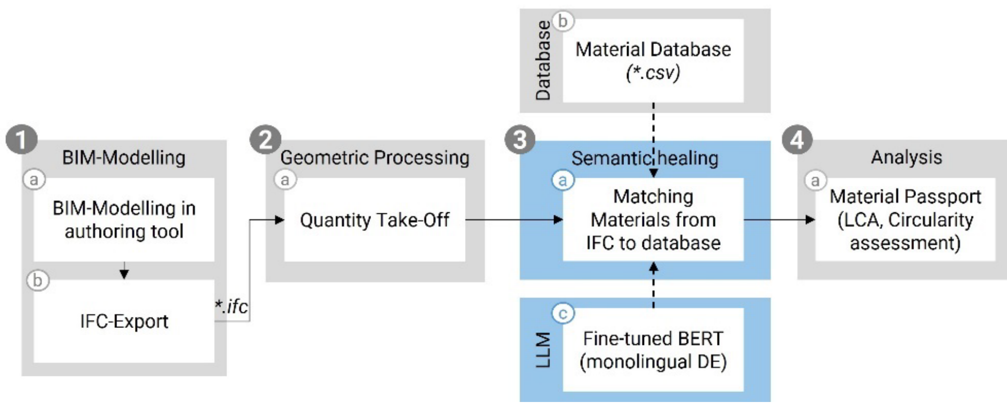

Figure 1. The general workflow of semantic enrichment of IFC models for Material passports.

General Workflow. As shown in Figure 1, the general workflow consists of four main steps. The first step includes the BIM modeling in the authoring tool (1.a) and the IFC export (1.b). The detailed requirements for the IFC export are described in the implementation section 4.2. The next step 2.a contains a quantity take-off of all relevant elements and materials using the base quantities of each element and layer. Step 3.a describes the main part of the proposed method, called semantic model enrichment. The quantity take-off derived in the previous step is used to automatically match the corresponding datasets from the material database (3.b). For this process of semantically enriching the IFC model, the highest semantic similarity of the material datasets with each material of each IFC element is used. We fine-tune a monolingual LLM domain specifically for this task based on the German language (3.c). In the final step 4.a, we can upload the semantically enriched and enriched IFC model to a Material Passport platform for further analysis, such as life cycle assessments (LCA) or circularity assessments.

Strategies for fine-tuning domain-specific LLM and improving matching performance. We propose five different strategies for the domain-specific LLM fine-tuning to improve the STS and matching performance: Strategy 1 - Adding domain-specific abbreviations, Strategy 2 - Applying different loss functions for fine-tuning, Strategy 3 - Adding different/ multiple labels for further context information (Element name, classification, IfcType, etc.), Strategy 4 - Adding negative/ contradicting word pairs, and Strategy 5 - Filtering word pairs according to IfcType.

As shown in previous studies with similar model enrichment tasks but for different analysis types (Forth et al. 2023), domain-specific abbreviations were a big challenge for the matching approach. Therefore, our first fine-tuning strategy is to train LLM with these AEC- and BIM-specific abbreviations.

For the second strategy, the following suitable loss functions are proposed for our application of fine-tuning using manually matched word pairs. In brackets, the typical labels for positive or negative word pairs are shown according to (Reimers 2023):

a) Cosine Similarity Loss: manual positive matches (0.8), negative matches (0.3)
b) Contrastive Loss: positive (1), negative (0)
c) MNRL – Multiple Negatives Ranking Loss: no labels needed

Another strategy for improving the fine-tuning performance is to add further knowledge of the BIM models using different labels for each type of information in the training process. For every material pair, we also know the IFC element name, the IfcType, and usually the classification. With the Softmax loss function, we can use different labels for fine-tuning, including this additional information. Therefore, we propose the following labels: Abbreviations (0), IFC material – positive material dataset (1), IFC element name – positive material dataset (2), classification – material dataset (3), IFC Type – material dataset (4), IFC material – contradicting material dataset (5).

As shown for the last label, the fourth strategy for improving the fine-tuning and matching performance is to include negative pairs. The MNRL has only positive word pairs with anchor ai, and pi being positive. But it assumes all other positives pj are the negative pairs, so ai and pj for i!=j are negative pairs. We can manually create negative pairs for all other loss functions according to the same logic but also check that pi!=pj. The negative labels have already been introduced in paragraphs of the previous strategies. This strategy can be realized the Cosine Similarity Loss (4a) as well as Contrastive Loss (4b).

The fifth strategy includes a filtering step of the material database. As the used material database is unstructured, we add a filter structure using different IfcTypes, such as IfcWall, IfcSlab, IfcCovering, IfcColumn, IfcDoor, IfcWindow, and IfcRoof. To enrich only applicable material datasets per IfcType, we check for all positive word pairs for their related IfcType and save the material dataset. Instead of comparing all 387 material datasets for each IfcMaterial, we can limit the material datasets to 82 for IfcSlab, 79 for IfcWall, 51 for IfcCovering, 21 for IfcColumn, etc.

This strategy does not improve the fine-tuning process but improves the matching performance afterward.

Combination of different strategies. We briefly describe how different strategies can be combined with each other to improve the fine-tuning process and match performance further. Adding abbreviations (strategy 1) and the filtering process (strategy 5) can be combined with all different strategies. Multiple labels (strategy 3) can only be realized with the Softmax Loss function, as the other functions don't allow multiple labels. Nevertheless, negative pairs can be realized with the Cosine Similarity Loss and Contrastive Loss functions. MNRL already incorporates the negative pairs, as described in the previous subsection. Different Loss functions could be combined in case more model context was fine-tuned with multiple labels and Softmax, and this LLM is used afterward as the base model for another fine-tuning process with Cosine Similarity, Contrastive, or MNR Loss function or in the opposite order.

Matching materials by highest Semantic Textual Similarity (STS). Figure 2 shows the general matching workflow of matching the semantically most similar material of the material database to each IFC element layer's IFC material. To this end, first (1), all IFC elements are iterated (1.a) and, next, filtered according to their IfcType if the filtering strategy (S5) is applied. If not, we go to the following step (1.c) of iterating for each element and its material layers. These materials are then compared with the whole or the filtered material database, so the material datasets are iterated (1.d). Each IfcMaterial (2.a) and each material dataset from the database (2.b) are encoded in the next step using different fine-tuned LLM. Next, the cosine similarity with each material from the database is calculated for STS (3.a). The material from the database with the highest cosine similarity is selected as the matched material for each IFC material (3.b).

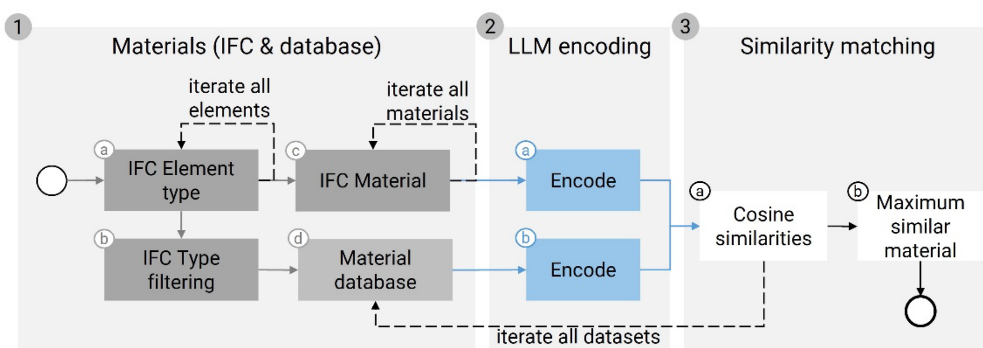

Figure 2. Automated material matching process of IFC materials to MP datasets.

CASE STUDIES AND IMPEMENTATION

Case studies and datasets. For the domain-specific abbreviations, we used 571 general AEC abbreviations and their descriptions (Bundesamt für Bauwesen und Raumordnung 2021) and 155 BIM-specific abbreviations (Helmus et al. 2021), such as construction types or material abbreviations. Both abbreviation datasets are in the German language. We employed 23 IFC models as case studies, where the IFC materials had already manually enriched PSets for the Madaster Platform (Frank 2021). Besides the availability of the provided case studies by LIST Eco, the Madaster platform is one of the few Building Material Passport providers using open BIM data format and the only commercial platform that embeds and manages a portfolio of MPs of several buildings (Gebetsroither et al. 2024). The case study projects are a mix of logistic, residential, and office buildings from different designers and clients. This assures a high diversity in the data and real-world adaptability. The matched materials from the material database mainly include 387 EPEA datasets (Frank 2021), but some were customized and added to the overall database. We extracted the matches of IFC materials and MP dataset following Madaster-specific PropertySet called "MaterialOrProductName". Based on these case studies, we derived 245 unique material matches and split them into 75% training, 184 positive word pairs for training, and 25% test samples (61 test pairs).

Prototypical Implementation. To implement the proposed method of fine-tuning domain-specific LLM, we used the cased version of the German BERT model ('bert-base-german-cased') as a base model for training (Chan et al. 2020). All IFC models and the EPEA database are provided in German language. For the prototypical implementation of the training pipeline, we used SentenceTransformers packages based on the SBERT method by Reimers and Gurevych (Reimers and Gurevych 2019). These packages incorporate all mentioned loss functions from Subsection 3.2. The different labeling for the additional domain knowledge from the IFC models was pre-processed accordingly after parsing all quantity take-offs.

RESULTS

Results of the overall matching approach using different fine-tuned LLM strategies. Figure 3 depicts the achieved matching accuracy, so the ratio of correct and total matches/ predictions, when applying the different strategies. Instead of only showing the correct matches of the most similar solution, we add a continuous solution space of the maximum ten most similar matches. This is because the initial results would not have a significant difference, and a deeper analysis would not be possible. The results show that the base model ('bert-base-german-cased') has only 44,26% correct matches, taking the most similar match into account, but increases up to 60,66% of correct matches considering the top 10 maximum similar matches. Different individual matching accuracies exist for domain-specific information considering AEC-overall and BIM-specific abbreviations. Each abbreviation source slightly increases the matching accuracy. However, by combining both abbreviations, the matching accuracy significantly increases to 67,21% for the top 10 matches. Adding multiple labels (strategy 3) increases the matching accuracy for the top four matches, but it even underperforms the base model for the following matches. The highest increase in the matching performance is using the filter strategy. The filtering is applied to the base model and reaches up to 75,41% correct matches. The loss function with the highest matching accuracy is Multiple Negative Ranking Loss (MNRL), which already considers negative pairs, followed by Contrastive Loss. Cosine Similarity Loss even underperforms compared to the base model. Nevertheless, adding negative pairs significantly increases the matching performance. Overall, the loss function with the highest matching accuracy is Contrastive Loss, including negative pairs, although it's computationally more expensive than MNRL by a factor of ca. 100.

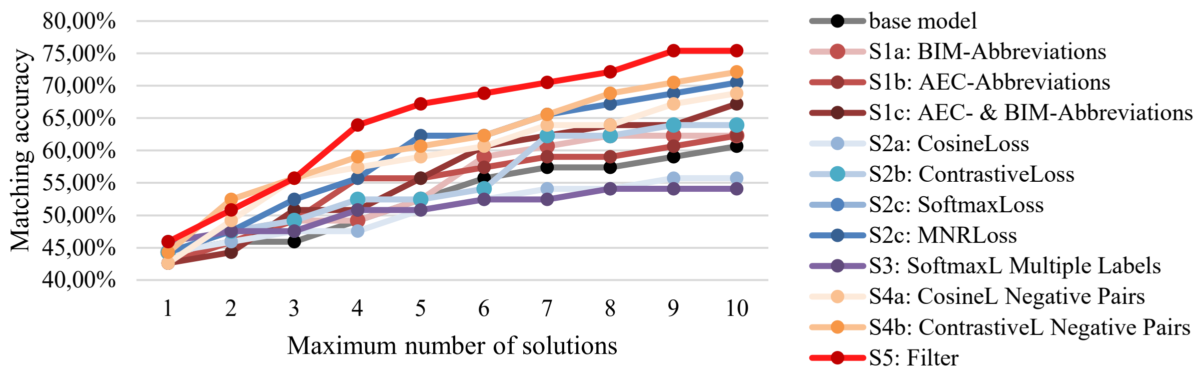

Figure 3. c.

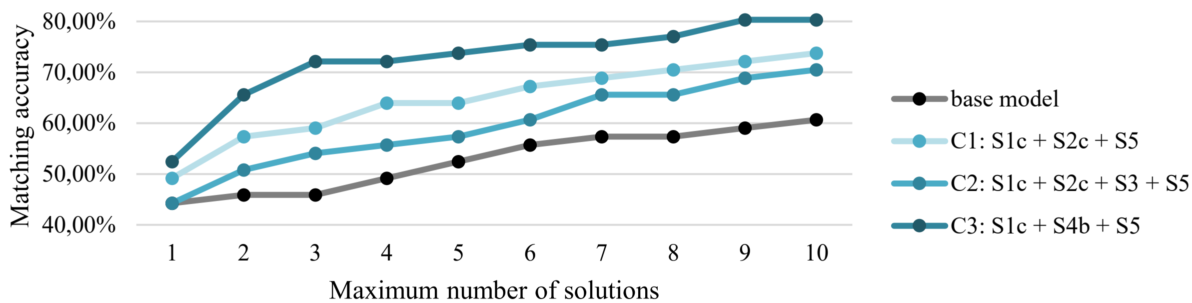

Figure 4. Matching accuracy for combined strategies compared to the base model.

Results of the overall matching approach using different combinations of the strategies. We defined a combination set of different strategies as follows:

- C1: Training AEC-, BIM-Abbreviations (S1c), Material Datasets with MNRL (S2c) and filtering (S5)
- C2: Training AEC-, BIM-Abbreviations (S1c), Material datasets with MNRL (S2c), multiple labels with SoftmaxL (S3) as base model, and filtering (S5)
- C3: Material Datasets with ContrastiveLoss including negative pairs (S4b), AEC- and BIM abbreviations using MNRL (S1c) as base model, and filtering (S5)

Figure 4 shows the results of the matching accuracies of combining different strategies compared to the base model. Generally, the results indicate that combining the individual strategies increases the matching accuracy even more. Nevertheless, adding more context with multiple labels (S3) did not improve the overall performance (see Section 5.1). Adding this strategy to the combination of C1 lowers the accuracy. The best-performing combination of strategies is C3, reaching up to 80,33% matching accuracy for the top ten matches. The following section analyzes the results of the best-performing individual strategies and the best-performing combination.

Analysis of correct and wrong matches. As shown on the left side of Fig. 5, the base model has approximately 44% correct matches but ca. 66% wrong matches. The reason for false matches was classified according to the main material group of the IfcMaterials. Most of the wrong matches (31%) are related to reinforced concrete. Accurate matching is challenging, as most of the IfcMaterials are named "Stahlbeton" (Engl. "Reinforced Concrete"), but the Material Datasets are more diverse, including specific compression strength classes and reinforcement ratios. The reason for wrong matches with aluminum, steel, and others (mainly asphalt and Larch wood) is primarily that in the IfcMaterial, more than one material is included, while in the ground truth, only one Material Dataset is matched. For the insulation materials, there are in the ground truth matching instead of "XPS" other Material Datasets matched, such as "EPS", making direct matching impossible. This can be avoided by including multiple similar materials.

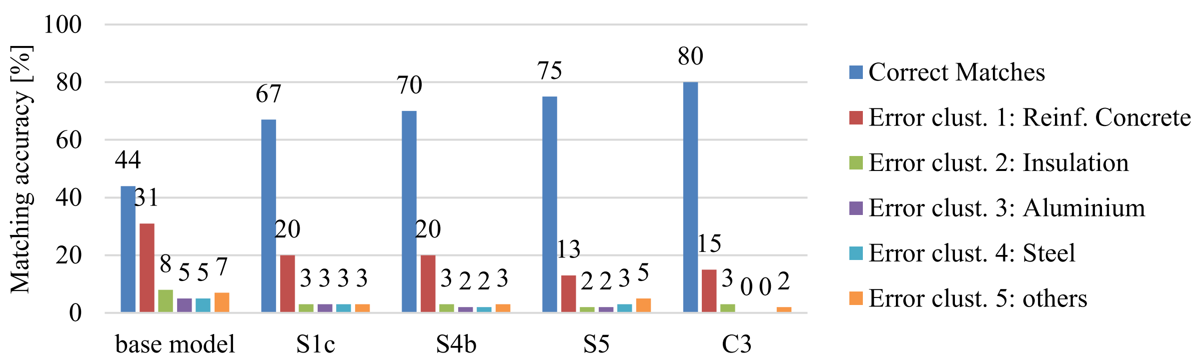

Figure 5. Matching analysis of correct and reasons for wrong matches comparing the base model and strategies S1c, S4b, S5, and C3.

Compared to the base model, the strategies S1c and S4b decrease the error of wrongly matching reinforced concrete by 11% (error cluster 1) and the error cluster 2-5 to 10-12%. Furthermore, only adding the filter reduces the error cluster of reinforced concrete by 18% and the other error clusters by 13%. Combining these three strategies solves the errors with steel and aluminum, and only the insulation error remains at 3% due to wrong classification. Furthermore, error cluster 5 still includes the error with the specific wood material. The error cluster 1 about reinforced concrete remains and can't be further solved. This is mainly because, for the wrong matches, different reinforcement ratios are added, which can not be predicted by the IfcMaterial alone. There are 23 different material datasets for reinforced concrete with varying ratios of reinforcement, priorities, or user-specific datasets. This issue could be handled by adding this information before automatically matching to increase matching accuracy.

Limitations. The most significant limitation of this publication is the limited number of matching samples in the dataset. From 23 real-world BIM models, only 245 unique matching samples were extracted, so the fine-tuning process took place with 186 samples. Also, the test datasets were limited to 61 matches. However, besides the limited number of samples, their quality also limits the accuracy of the matching. As previously analyzed, having multiple IfcMaterials matched to only one input limits the overall performance. Another limitation is that in this study, we only took one LLM network architecture into account. As we have German material expressions, we used the German version of BERT ('bert-base-german-cased') as the base model (Chan et al. 2020). Finally, there is no 100% matching accuracy possible. This means this approach is a support tool than fully automating the process. For this reason, we included the Material Dataset with the highest similarity and extended it to the top 10 most similar samples.

CONCLUSION AND OUTLOOK

In this paper, we introduce a method of (semi-)automatically matching BIM materials to the relevant material datasets using Semantic Textual Similarity (STS) and different strategies of domain-specific fine-tuning pre-trained Large Language Models (LLM). The method matches the semantically most similar material datasets to every BIM material for further analysis. We used the German BERT LLM and sentence embeddings using Siamese BERT-Networks for fine-tuning. The five strategies and their combination increase the matching accuracy from 44,26% to 80,33% by extending the solution space to ten material datasets with the highest semantic similarity. Therefore, the low matching accuracy of the most similar match leads to using this method as a support tool instead of a fully automated approach. Although we had 23 real-world case studies, the 245 material samples with different data quality are still limited.

In our future research, we will use more case studies and material samples for training and testing with cross-validation for more robust solutions. Furthermore, a more structured database, rather than only differentiating by IfcTypes, could increase the accuracy of the matching in the filtering step. As we identified too many similar material datasets for reinforced concrete, we suggest an interim step of adding more information about reinforcement ratio and priority. Additionally, these fine-tuning strategies shall be transferred to multilingual training for enriching building energy models for building performance simulations (Forth 2023). Last, more information is missing, such as the connection type of different elements to derive the detachability index to enable fully automated circularity assessments in early design stages.

ACKNOWLEDGEMENT.

We thank LIST Eco GmbH & Co. KG for providing real-world case studies.

REFERENCES

Atta, I., Bakhoum, E. S., and Marzouk, M. M. (2021). "Digitizing material passport for sustainable construction projects using BIM." *Journal of Building Engineering,* 43. 103233. https://doi.org/10.1016/j.jobe.2021.103233.

Bundesamt für Bauwesen und Raumordnung (2021). *Abkürzungsverzeichnis* <https://bauarchivddr.bbr-server.de/bauarchivddr/archiv/dokumente/pb-abkuerzungsverzeichnis.pdf> (Oct. 31, 2023).

Çetin, S., Raghu, D., Honic, M., Straub, A., and Gruis, V. (2023). "Data requirements and availabilities for material passports: A digitally enabled framework for improving the circularity of existing buildings." *Sustainable Production and Consumption,* 40. 422–437. https://doi.org/10.1016/j.spc.2023.07.011.

Chan, B., Schweter, S., and Möller, T. (2020). "German's Next Language Model." In *Proc., Proceedings of the 28th International Conference on Computational Linguistics,* edited by D. Scott, N. Bel, and C. Zong6788–6796., Stroudsburg, PA, USA.

European Commission (2020). "In focus: Energy efficiency in buildings." <https://commission.europa.eu/news/focus-energy-efficiency-buildings-2020-02-17_en>

Forth, K. (2023). *Multilingual semantic enrichment of room-specific load profiles using BIM models for whole building energy simulation*: Ruhr-Universität Bochum.

Forth, K., Abualdenien, J., and Borrmann, A. (2023). "Calculation of embodied GHG emissions in early building design stages using BIM and NLP-based semantic model healing." *Energy and Buildings,* 284. 112837. https://doi.org/10.1016/j.enbuild.2023.112837.

Frank, C. (2021). "IFC-Export für die Madaster-Plattform: Version 3.0.".

Gebetsroither, M., M. Honic, I. Kovacic, C. Löffler, K. Marx, R. Pamminger, S. Robbi, C. Sustr, S. Schützenhofer, and G. Weber, eds. (2024). *Paradigmenwechsel in Bau- und Immobilienwirtschaft*, Berlin, Heidelberg: Springer Berlin Heidelberg.

Helmus, M., A. Meins-Becker, D. J. Feller, G. Hort, H. Hoffmann, C. Motzo, and A. Norrdine (2021). *Leitfaden für die Erstellung eines Bauwerksdatenmodells: Anlage 2: Bezeichnungsvorschläge*.

Honic, M., Kovacic, I., and Rechberger, H. (2019). "BIM-Based Material Passport (MP) as an Optimization Tool for Increasing the Recyclability of Buildings." *AMM,* 887. 327–334. https://doi.org/10.4028/www.scientific.net/AMM.887.327.

Jadwin, A., and Huang, C. (2023). "Improving minBERT Performance on Multiple Tasks through In-domain Pretraining, Negatives Ranking Loss Learning, and Hyperparameter Optimization." *Stanford University.*

Reimers, N. (2023). "Sentence-Transformers Losses." <https://www.sbert.net/docs/package_reference/losses.html> (Nov. 14, 2023).

Reimers, N., and Gurevych, I. (2019). "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks." In EMNLP 2019.

Sachidananda, V., J. S. Kessler, and Y. Lai (2021). *Efficient Domain Adaptation of Language Models via Adaptive Tokenization* <http://arxiv.org/pdf/2109.07460v1>.