Multimodal Embedding-Based Pipeline for Matching Building Components Using Digital Product Passports

Schatz Y., Domer B.
HEPIA, University of Applied Sciences and Arts Western Switzerland
yohann.schatz@hesge.ch

Abstract. Building renovation requires identifying suitable replacement components, but product data are often heterogeneous and poorly structured. Digital Product Passports (DPPs), introduced by EU Regulation 2024/1781, provide standardized, machine-readable product information that can support automated comparison. This paper proposes an embedding-based pipeline for matching building components using data from DPPs. The approach combines multimodal similarity measures with a dual embedding strategy for textual attributes. Attribute similarities are aggregated using a Tversky-based formulation with category-level prioritization. Experiments demonstrate effective semantic similarity capture, robustness to incomplete data, and adaptability to user-defined priorities, showing that the approach is suitable for assisting component replacement decisions.

1. Introduction

A large share of the existing building stock does not meet climate objectives in terms of energy performance and requires renovation (EEA, 2022). Renovation involves decision-making regarding existing building components based on their condition, and when replacement is necessary, suitable alternative products must be identified. In practice, this process remains largely manual and is hindered by the heterogeneity and poor structuring of data describing candidate products.

Digital Product Passports (DPPs), introduced by EU Regulation 2024/1781 (EU, 2024), offer a promising solution. DPPs provide a standardized, machine-readable representation of data on building components or products (Rute Costa and Hoolahan, 2024; Halbach and De Boissieu, 2022). When integrated into BIM workflows, they enable automation and enhance data integrity (buildingSMART Switzerland, 2025; Honic et al., 2024).

The adoption of DPPs is scheduled for 2027 in the EU (EC, 2025). They will be produced by different manufacturers with content that may be inconsistent and thus hinder straightforward equality-based matching. This motivates framing component matching as an entity resolution and similarity assessment problem.

Existing approaches include similarity computation based on embeddings (Huang et al., 2025; Li et al., 2020), optionally combined with rule-based filtering (Forth et al., 2023, 2025), and experimental LLM-based workflows (Petrosa et al., 2025). Despite their potential, existing methods exhibit limitations, including sensitivity to data asymmetry and semantic variation in textual information. In particular, most approaches assume a fixed and homogeneous dataset, whereas DPPs involve heterogeneous information that differs across different types of building products.

This paper proposes a pipeline for matching building components using data from DPPs. The method employs multimodal embeddings to rank candidate components based on their similarity to a user-defined target profile.

Main contributions are a dual embedding strategy for textual attributes that enables robust similarity evaluation for both short labels and longer textual descriptions, and a multi-level, Tversky-based aggregation of feature similarities that supports imbalanced datasets and enhances adaptability.

2. Methodology

The proposed pipeline is illustrated in Figure 1.

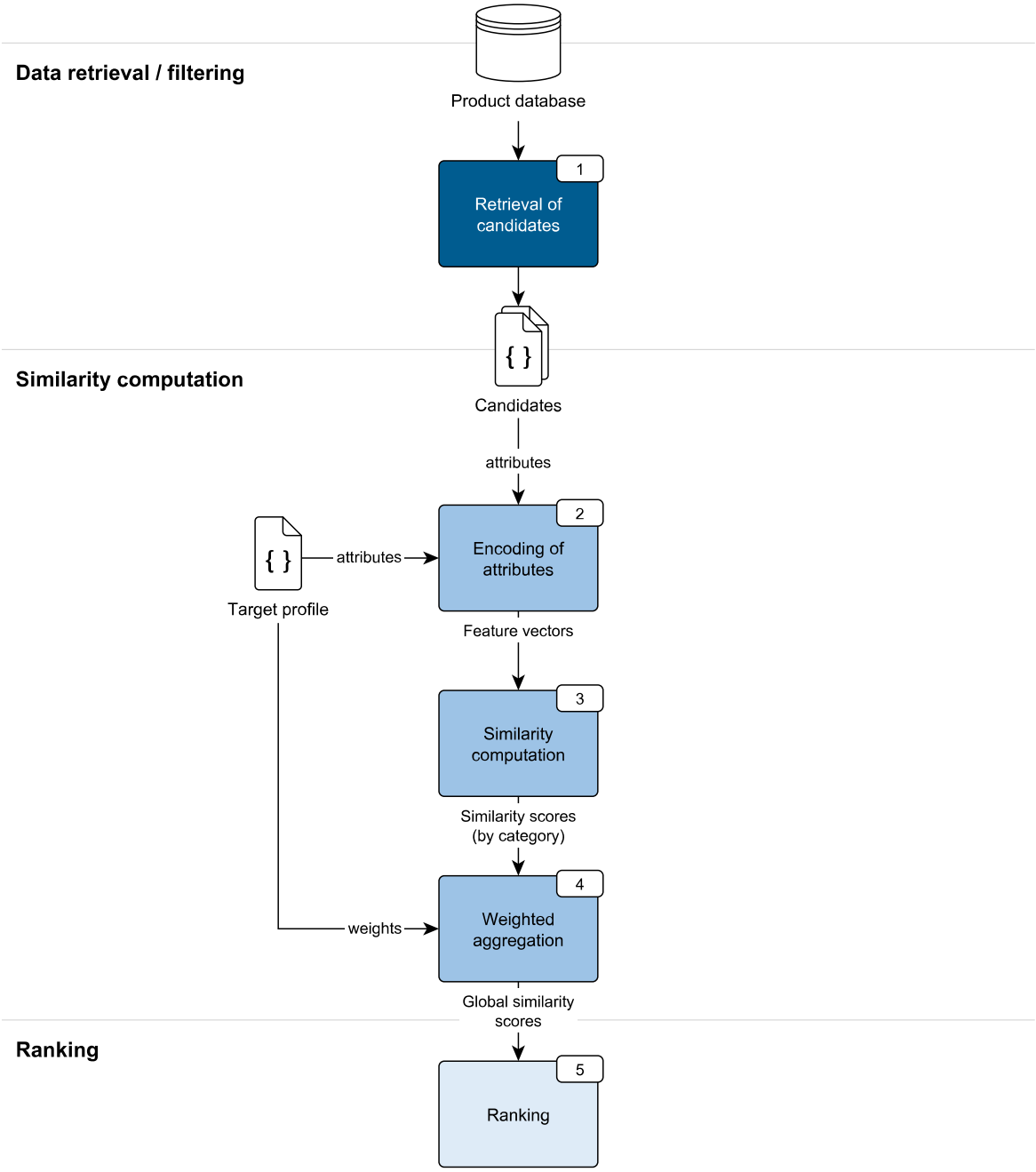

Figure 1: Pipeline for matching building components

2.1 Retrieval of candidates

DPPs of various building components are stored in a database. Referenced products belong to diverse types, such as windows, doors, and hatches, following the categorization used in catalogs such as NBS Source (Hubexo, 2026). During retrieval, user-defined rules are applied to filter products according to their type or classification, thereby restricting the similarity computation to a subset of relevant candidates.

The resulting DPPs are represented as JSON documents. Their structure complies with the DPP schema developed by the United Nations Economic Commission for Europe (UNECE, 2026), which allows issuers to include industry-specific attributes that are not covered by the core model.

In the simplified example below, window-specific attributes are embedded under the object "characteristics", which serves as an extension mechanism :

```
{
  "@context": ["https://example.org/dpp"],
  "type": ["Product"],
  "id": "did:web:manufacturer.com:product:pvc-window-001",
  "name": "PVC window",
  "idGranularity": "item",
  "characteristics": {
    "type": ["Characteristics"],
    "design": {
      "color": "White",
      "material": "PVC",
      "shape": "Rectangular with rounded corners"
    },
    "dimensions": {
      "height": 1.5,
      "width": 1.2
    },
    "performance": {
      "solar_factor_ratio": 0.63,
      "u_value": 1.1
    },
    "durability": {
      "expected_service_life": 30,
      "is_recyclable": true,
      "properties": "UV-resistant and weatherproof"
    },
    "security": {
      "has_safety_glazing": true,
      "lock_type": "Multipoint locking"
    },
    "maintenance": {
      "instructions": "Clean regularly and lubricate once a year"
    }
  }
}
```

Attributes are organized into thematic categories (e.g., design, dimensions, performance), while individual attributes may vary depending on the product type, ensuring both conceptual consistency and flexibility in representation.

Individual attributes are represented as key–value pairs with controlled attribute names and values of varying types, including boolean, numeric, short labels (e.g., "PVC"), and longer textual descriptions (e.g., "Rectangular with rounded corners"). All measurement values are expressed in SI units to ensure comparability.

2.2 Encoding of attributes

Candidates are evaluated against a user-defined target profile, specifying the expected attributes and associated values to match in DPPs. To enable scalable and consistent comparison, attributes are encoded as vectors, providing a common representation for similarity computation and aggregation. The encoding strategy depends on the nature of the attribute values, with specific representations defined for each data type :

- Boolean attributes are encoded as binary values.
- Numeric attributes are normalized to ensure scale invariance.
- Short labels are encoded using Numberbatch, a pre-trained word embedding model based on the ConceptNet knowledge graph (Speer et al., 2016) that effectively captures semantic relationships between discrete terms.
- Long textual descriptions are encoded using Sentence-BERT embeddings (Reimers and Gurevych, 2019), with MPNet (Song et al., 2020) used as the underlying transformer backbone. The model is fine-tuned on 28,500 classes and property definitions extracted from the buildingSMART Data Dictionary (buildingSMART International, 2026), enabling the capture of semantic similarity across technical descriptions.

The embedding space is defined separately for each attribute type. Original vector dimensions are preserved for labels (300) and long textual descriptions (768), resulting in a multimodal, multidimensional representation. Similarity comparisons are consistently performed within the same embedding space to guarantee valid and reliable measurements.

2.3 Similarity computation

Similarity is computed separately for each attribute category, enabling high-level prioritization and enhancing interpretability. First, similarities are computed for all available attributes within each category, using cosine similarity for textual attributes, relative difference for numerical attributes, and exact matching for Boolean attributes. The resulting attribute similarities are then aggregated using a Tversky-based formulation :

$$\mathrm{Sim_c} = \frac{\sum_{\mathrm{a}\in \mathrm{A}\cap \mathrm{B}} \mathrm{Sim_a}}{|\,\mathrm{A}\,| + \alpha\,|\,\mathrm{A}\setminus \mathrm{B}\,|} \tag{1}$$

where :

- Sim_a is the similarity of attribute a ;
- Sim_c is the similarity for category c ;
- $A \cap B$ are attributes present in both the target profile (A) and candidate (B) ;
- $|\,A\,|$ is the total number of attributes in A ;
- $|\,A \setminus B\,|$ is the number of attributes present in A but not in B ;
- α is a penalty factor for missing attributes.

This formulation accounts for data asymmetries between the user-defined profile and candidate products, which may differ in completeness. Attributes present in a candidate but absent from the target profile are ignored to prevent artificial inflation, while missing attributes are penalized through the α parameter, reducing the similarity of incomplete candidates.

2.4 Weighted aggregation

Global similarity is obtained by aggregating the similarities of attribute categories using user-defined weights :

$$\mathrm{Sim}_{\mathrm{global}} = \frac{\sum_{c \in C} w_c \cdot \mathrm{Sim}_c}{\sum_{c \in C} w_c} \tag{2}$$

where :

- $\mathrm{Sim}_{\mathrm{global}}$ is the global similarity of the candidate relative to the target profile ;
- Sim_c is the aggregated similarity for category c ;
- w_c is the user-defined weight of category c ;
- C is the set of all attribute categories.

Categories without explicit weights share the remaining proportion equally, preserving balanced contributions to the global similarity score. Categories defined in the target profile but absent from a candidate are assigned a zero similarity score, ensuring that missing information is penalized.

2.5 Ranking

Candidates are finally ranked based on their global similarity scores. This ranked output facilitates expert decision-making by highlighting the most compatible alternatives and maintaining transparency regarding the contribution of individual attribute categories to the overall similarity.

3. Evaluation

The proposed pipeline is evaluated in the context of a building renovation scenario in which an existing wooden window must be replaced. The objective is to identify suitable candidate components that match the original dimensions while improving performance characteristics.

A target profile representing a modern PVC window is defined, and 50 synthetic DPPs are generated from real-world products, following the structure presented in Section 2.1.

The algorithms were implemented in Python. The fine-tuned S-BERT model was developed using the Sentence Transformers library (Reimers, 2026), while Gensim (Řehůřek, 2026) was used for the integration of ConceptNet. The evaluation focused on four critical aspects: (i) ranking accuracy, (ii) semantic similarity performance, (iii) robustness to data asymmetry, and (iv) sensitivity to weighting parameters.

3.1 Ranking accuracy

This experiment evaluates the framework's ability to correctly rank candidate products according to their similarity to the target profile. Ten different configurations are tested, each with a specific set of candidates and multiple weighting schemes, with averages computed across weights for each configuration.

Results are shown in Figure 2.

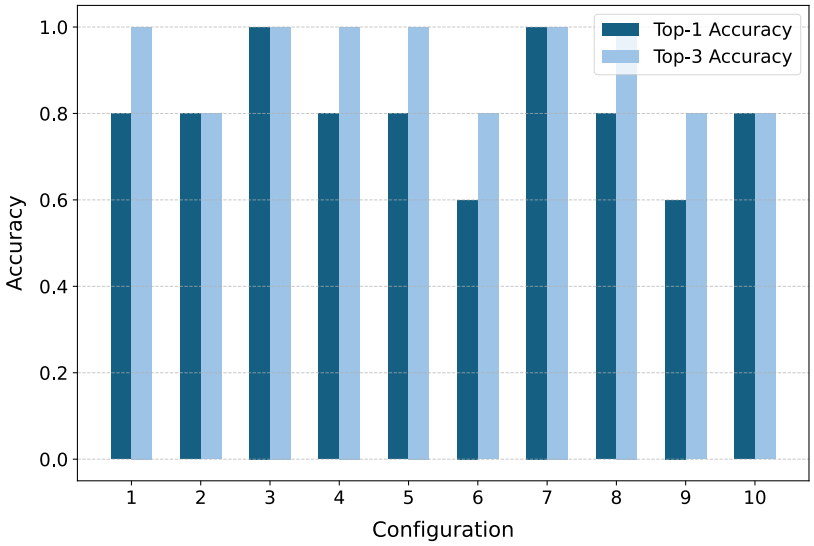

Figure 2: Ranking accuracy across configurations

The proposed method achieves strong ranking performance, with an average Top-1 accuracy of 0.80 and Top-3 accuracy of 0.92, indicating that the most relevant candidates are reliably identified among the top-ranked results. The limited variability across configurations demonstrates robustness to changes in data and weighting schemes.

3.2 Semantic similarity performance

This experiment evaluates the ability to capture semantic equivalence and proximity for both short labels and long textual descriptions. Pairwise similarity is calculated for "PVC" against five other materials, and for "Rectangular with rounded corners" against alternative shape descriptions.

Results are shown in Table 1.

Table 1: Similarity scores for textual attributes

Reference	Candidate	Similarity
PVC	Polyvinyl chloride	0.85
	Polyvinyl	0.66
	Plastic	0.60
	Wood	0.27
	RPVC	0.00
Rectangular with rounded corners	Rectangular rounded	0.89
	Rectangular with smoothed corners	0.86
	Rectangular	0.77
	Trapezoidal with rounded corners	0.75
	Circular with smooth edges	0.55

The Numberbatch model effectively distinguishes related from unrelated labels, assigning high similarity scores to "polyvinyl chloride" (0.85), "polyvinyl" (0.66), and "plastic" (0.60), and a low score to "wood" (0.27). The similarity of "RPVC" (Rigid PVC) is zero, as the term is not defined in ConceptNet, highlighting a critical limitation.

For long textual descriptions, the fine-tuned SBERT model produces consistent results, with similarity decreasing as shapes deviate from the reference. Rectangular shapes score highest, with "Rectangular rounded" at 0.89 and "Rectangular" (0.77) higher than "Trapezoidal with rounded corner" (0.75), indicating correct penalization of missing terms. Using the backbone model without fine-tuning yields similar rankings but with reduced separation among the top candidates, confirming the benefit of fine-tuning on bSDD data.

3.3 Data asymmetry robustness

This experiment evaluates the robustness of similarity computation in the presence of missing or additional data, verifying that similarity scores degrade progressively rather than collapsing abruptly. The candidate set is derived from the target profile, with five variants :

- 1 — additional attribute ;
- 2 — additional category ;
- 3 — original profile (perfect match) ;
- 4 — missing attribute ;
- 5 — missing category.

Results are shown in Figure 3.

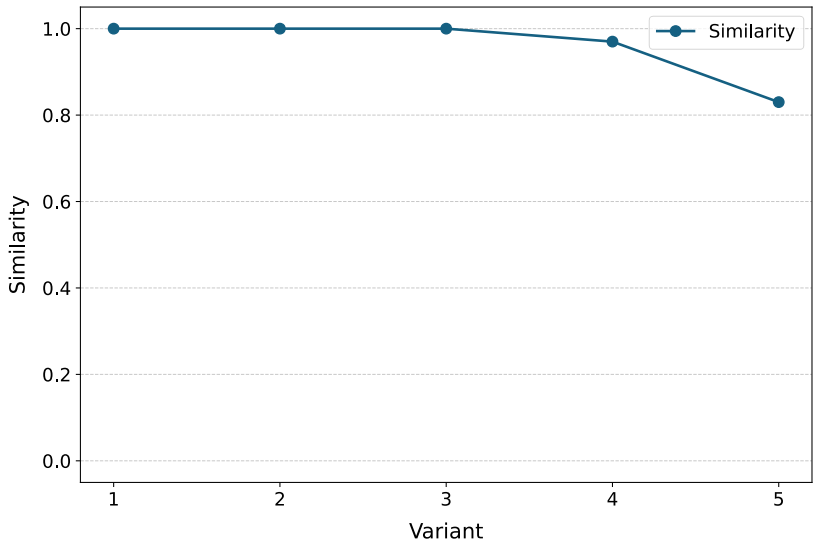

Figure 3: Similarity degradation under data asymmetry

Candidates with additional attributes or categories remain identical to the original profile, with a similarity of 1, showing that extra information does not inflate similarity. A missing attribute reduces similarity slightly (0.97), and a missing category reduces it further (0.83). The gradual decrease reflects that unchanged attributes still contribute, demonstrating stable, predictable behavior and the effectiveness of the asymmetry-handling mechanism.

3.4 Weight sensitivity

This experiment evaluates the sensitivity of the ranking to category-level weights, illustrating how the method adapts to different renovation priorities. Three weighting schemes (balanced, design-focused, performance-focused) are tested on candidates optimized for design (C1), optimized for performance (C2), or reflecting a neutral profile (C3).

Results are shown in Figure 4.

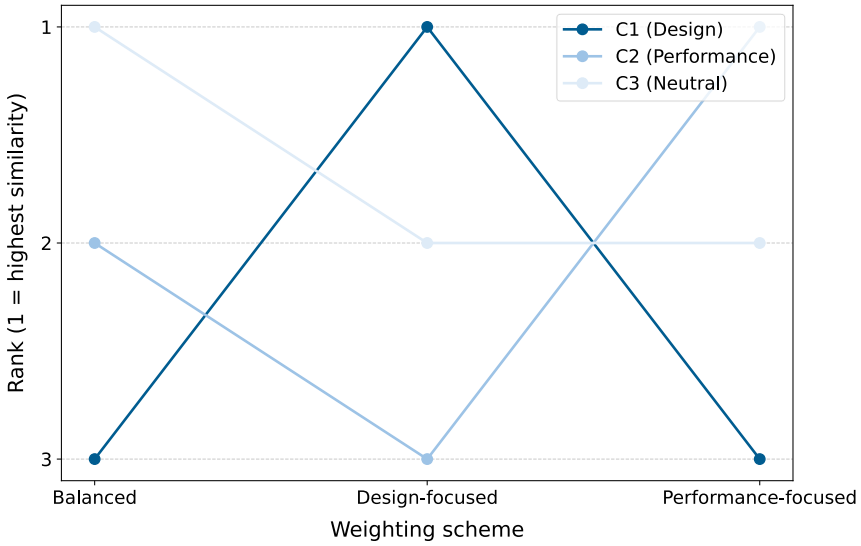

Figure 4: Candidate rankings across weighting schemes

Candidate rankings are influenced by attribute category weighting. Kendall's tau correlations across schemes range from −1 to 0.33, highlighting that emphasizing specific categories can substantially alter similarity-based rankings, whereas balanced weighting tends to maintain partial consistency. These results confirm that the proposed mechanism effectively incorporates user-defined priorities, supporting flexible and context-sensitive decision-making.

4. Discussion

The evaluation highlights several strengths of the proposed approach. First, the multimodal similarity framework enables the comparison of heterogeneous attribute types, including numerical, Boolean, and textual data. Second, the hybrid embedding strategy, combining ConceptNet Numberbatch for labels and a fine-tuned SBERT model for long textual descriptions, effectively captures semantic similarity regardless of text length. Third, the Tversky-based formulation demonstrates robustness to missing data, ensuring stable similarity scores in the presence of asymmetric information. Finally, the use of category-level prioritization allows domain experts to express preferences at a higher level, facilitating the adaptation of rankings to specific renovation objectives.

However, several limitations must be acknowledged. A key limitation is that the evaluation was conducted on synthetically generated DPPs, which may affect real-world applicability. Another limitation is the reliance on ConceptNet Numberbatch, which introduces a dependency on knowledge graph coverage, since labels not defined in the graph result in null similarity scores, potentially biasing comparisons.

In addition, the quality of textual similarity depends on the embedding model and the training data, making domain-specific fine-tuning important for achieving reliable results. The numerical similarity formulation relies on relative differences, assuming that proportional deviations are meaningful, which may not always align with engineering tolerances.

From a computational perspective, the current similarity computation has a complexity of O(nm), where n represents the number of candidate products and m the number of attributes considered for comparison. For each candidate, similarities are computed across all relevant attributes, including vector comparisons for textual embeddings. In large candidate sets, this repeated encoding and comparison may become computationally expensive. In practical implementations, candidate attributes should therefore be pre-encoded, allowing their vector representations to be stored in advance.

5. Conclusion and future work

This paper presented a similarity-based pipeline for matching building components using data from DPPs. The approach leverages multimodal embeddings and a Tversky-based aggregation mechanism to enable the comparison of heterogeneous product attributes and support flexible prioritization of decision criteria. The evaluation demonstrates that the method can capture semantic similarity for both short labels and longer textual descriptions, while remaining robust to incomplete or asymmetric data. These properties make the proposed approach suitable for assisting component replacement decisions in renovation workflows.

Future work will focus on improving scalability and semantic coverage. Specifically, integrating alternative knowledge graphs or fallback embedding strategies could reduce the dependency on ConceptNet and improve the handling of previously unseen labels. Additional research is also needed to better account for dependencies between attributes, for instance, through graph-based similarity models or ontology-aware representations. Finally, the approach should be validated on larger real-world DPP datasets and integrated into BIM-based decision-support tools to assess its performance in practical renovation scenarios.

References

buildingSMART International (2026). buildingSMART Data Dictionary. Available at: https://bsdd.buildingsmart.org/, accessed March 2026.

buildingSMART Switzerland (2025). Livre blanc passeport numérique des produits (PNP).

EC (2025). Ecodesign for Sustainable Products and Energy Labelling Working Plan 2025–2030. Available at: https://environment.ec.europa.eu/document/download/5f7ff5e2-ebe9-4bd4-a139-db881bd6398f_en?filename=FAQ-UPDATE-4th-Iteration_clean.pdf, accessed January 2026.

EEA (2022). Building renovation: where circular economy and climate meet. Available at: https://www.eea.europa.eu/en/analysis/publications/building-renovation-where-circular-economy-and-climate-meet/, accessed January 2026.

EU (2024). Regulation (EU) 2024/1781 (Document 32024R1781). Available at: https://eur-lex.europa.eu/eli/reg/2024/1781/oj, accessed January 2026.

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

Forth, K., Kaltenegger, J., Petrova, E., De Wolf, C. (2025). BIM-based material property enrichment using text-based and ontology-based semantic similarity matching 10 p. DOI: https://doi.org/10.3929/ETHZ-B-000733251.

Halbach, A., De Boissieu, A. (2022). Quel écosystème de données pour un passeport matériau BIM ? Revue de la littérature et perspectives pour de futures recherches. SHS Web Conf. 147, 02001. DOI: https://doi.org/10.1051/shsconf/202214702001.

Honic, M., Magalhães, P.M., Van Den Bosch, P. (2024). From Data Templates to Material Passports and Digital Product Passports, in: De Wolf, C., Çetin, S., Bocken, N.M.P. (Eds.), A Circular Built Environment in the Digital Age, Circular Economy and Sustainability. Springer International Publishing, Cham, pp. 79–94. DOI: https://doi.org/10.1007/978-3-031-39675-5_5.

Huang, P.-H., Song, S.-Y., Xu, Z., Hu, Z.-Z., Lin, J.-R. (2025). Intelligent BIM Searching via Deep Embedding of Geometric, Semantic, and Topological Features. Buildings 15, 951. DOI: https://doi.org/10.3390/buildings15060951.

Hubexo (2026). NBS Source. Available at: https://source.thenbs.com/, accessed March 2026.

Li, N., Li, Q., Liu, Y.-S., Lu, W., Wang, W. (2020). BIMSeek++: Retrieving BIM components using similarity measurement of attributes. Comput. Ind. 116, 103186. DOI: https://doi.org/10.1016/j.compind.2020.103186.

Petrosa, D., Haverkamp, P., Backes, J.G., Crampen, D., Blankenbach, J., Traverso, M. (2025). Development of a BIM-based AI-driven matching tool for LCA datasets. Discov. Sustain. 6, 1237. DOI: https://doi.org/10.1007/s43621-025-02203-8.

Řehůřek, R. (2026). Gensim Python library. Available at: https://pypi.org/project/gensim/, accessed March 2026.

Reimers, N. (2026). Sentence Transformers Python library. Available at: https://pypi.org/project/sentence-transformers/, accessed March 2026.

Reimers, N., Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. DOI: https://doi.org/10.48550/arXiv.1908.10084.

Rute Costa, A., Hoolahan, R. (2024). Material Passports : Accelerating Material Reuse in Construction.

Song, K., Tan, X., Qin, T., Lu, J., Liu, T.-Y. (2020). MPNet: Masked and Permuted Pre-training for Language Understanding. DOI: https://doi.org/10.48550/arXiv.2004.09297.

Speer, R., Chin, J., Havasi, C. (2018). ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. DOI: https://doi.org/10.48550/arXiv.1612.03975.

UNECE (2026). Digital Product Passport. Available at: https://untp.unece.org/docs/0.6.0/specification/DigitalProductPassport, accessed January 2026.