Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Auditable AI Assisted Semantic Completion Supports Metadata Enhancement in Digital Cultural Heritage Collections

**Ziqi Shen**

ZiqiShen1991@163.com

University of Pennsylvania

**Research Article**

**Keywords:** digital cultural heritage, metadata enhancement, semantic annotation, information organization, knowledge discovery, artificial intelligence

**Posted Date:** August 19th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-10313654/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

**Additional Declarations:** No competing interests reported.

## Abstract

**Purpose:**Digital cultural heritage collections are increasingly available as data, but their discovery depends on metadata that is often incomplete and unevenly structured. This study develops and evaluates an auditable AI-assisted semantic completion workflow for enhancing hierarchical metadata in digital cultural heritage collections.

**Design:**The study uses the EUFCC-340K open GLAM benchmark derived from Europeana. A reproducible masked-label evaluation tests whether observed controlled-vocabulary labels can recommend missing object-type, material, subject and classification labels. The workflow combines metadata quality profiling, facet parsing, label co-occurrence graphs, frequency baselines, two-hop propagation and transparent ranking.

**Findings:**Metadata coverage is highly asymmetric: object-type and material fields are dense, while subject and classification fields are sparse or distribution-sensitive. In a fixed 25,000-record training subset, direct co-occurrence outperformed frequency baselines on the inner-test split, reaching Recall@5 of 0.732 for object types, 0.790 for materials, 0.986 for classifications and 1.000 for subjects.

**Originality:**The paper contributes a reproducible information-organization workflow that treats AI-assisted metadata enhancement as an auditable ranking problem and supports scalable, human-reviewable metadata enrichment.

## 1. Introduction

The last decade of large-scale digitisation has changed the informational status of cultural heritage collections. Museum objects, archival images, books, manuscripts, textiles, scientific instruments and vernacular artefacts are no longer encountered only as isolated catalogue entries. They are increasingly published as computational collections that can be searched, mined, linked, recommended and reused by scholars, educators, creative practitioners and the wider public. This transition creates a central problem for library and information science: digitisation alone does not make cultural heritage discoverable. Discoverability depends on metadata quality, semantic interoperability, controlled vocabulary alignment and the capacity of collection systems to generate meaningful relations among objects. A digitised collection with weak metadata remains visible only in a superficial sense; it may be online, but it is not necessarily findable, interpretable, reusable or accountable.

The problem is especially acute for digital cultural heritage because GLAM institutions describe heterogeneous objects through historically layered cataloguing practices. Object records can contain descriptions of materials, object types, subjects, techniques, iconography, historical periods, makers, places, provenance and collection context, but these elements are rarely complete or equally structured across institutions. Aggregators such as Europeana have made major progress in aligning collection metadata, while newer datasets have begun to expose images and records for computational use. Yet gaps remain between collection-level availability and item-level semantic richness. Many records contain useful material or object labels but lack subject terms; others include classification labels but no material hierarchy; still others contain multilingual terms or institution-specific vocabularies that hinder cross-collection retrieval.

Artificial intelligence is often presented as a solution to this metadata bottleneck. Computer vision, natural language processing, knowledge graphs and large language models can suggest tags, link concepts, align vocabularies and generate descriptions. However, the adoption of AI in cultural heritage settings raises problems that are not merely technical. Cultural heritage metadata carries interpretive authority. A generated label can amplify cultural bias, flatten contested meanings, misrepresent a community, or obscure provenance if it is inserted into a catalogue without adequate review. The practical question for digital libraries is therefore not whether AI can generate metadata, but how AI-assisted metadata enhancement can be made transparent, reproducible and reviewable by professional cataloguers and domain experts.

This article addresses that problem by reframing metadata enhancement as semantic completion. Instead of asking a black-box model to produce fluent descriptions, we ask whether observed controlled-vocabulary labels in a cultural heritage record can be used to recommend candidate labels that are missing from other facets. For example, if an object is labelled as a textile and a painting on oil, what material or object-type candidates are most plausible? If a record carries a particular material and classification label, can the label-space structure suggest a missing object-type term? Such recommendations do not replace cataloguers. They provide ranked candidates for human validation. This design is deliberately conservative: it prioritises auditability, ranking transparency and reproducibility over generative novelty.

The empirical setting is EUFCC-340K, a recent faceted hierarchical dataset for metadata annotation in GLAM collections. The dataset contains hundreds of thousands of Europeana-derived records with hierarchical labels across object types, materials, subjects and classifications. It is suitable for the present study because it formalises a problem that is familiar to digital library work: collection objects are partially described across multiple semantic facets, and the challenge is to improve the density and usefulness of those descriptions without losing controlled-vocabulary discipline. We use official train, validation, inner-test and outer-test splits, and we implement a deterministic masked-label benchmark in which existing labels are hidden and predicted from the remaining labels. This allows the study to evaluate semantic completion in a reproducible way without inventing labels or claiming a fully automated cataloguing solution.

The study makes three contributions. First, it provides a metadata quality profile of EUFCC-340K across the four core facets, showing where semantic coverage is strong and where it is sparse. Second, it develops an auditable label co-occurrence graph and compares frequency, direct co-occurrence, two-hop graph propagation and hybrid ranking methods under a masked-label evaluation protocol. Third, it translates the results into a practical governance model for AI-assisted metadata enhancement in digital libraries. The central argument is that high-value AI in cultural heritage information systems should not be judged only by predictive accuracy. It should also be judged by reproducibility, interpretability, provenance preservation and the ease with which professional

staff can inspect and correct proposed labels. Figure 1 summarizes the resulting auditable workflow from metadata parsing and quality profiling through candidate generation, human validation, and discoverability assessment.

## 2. Literature review and theoretical background

Recent LIS and digital heritage research has shifted from treating metadata as a static descriptive layer to treating collections as computationally reusable data. The collections-as-data literature argues that GLAM institutions need transparent documentation, licensing, provenance and data-quality reporting before large-scale computational reuse is possible (Candela et al., 2025; Lee, 2025). This is directly relevant to AI-assisted metadata work because annotation algorithms inherit both the strengths and omissions of existing cataloguing practice.

Metadata quality is therefore multidimensional rather than merely a question of whether a record has a title. Completeness, consistency, provenance, semantic specificity and interoperability all shape whether digital objects can be retrieved, compared and recombined across repositories. Lorenzini et al. (2021) provide an operational approach to metadata completeness in digital cultural heritage repositories, while Candela et al. (2023) show that linked cultural heritage data quality can be assessed automatically. These studies justify the present paper's decision to report coverage, density and split-level robustness before evaluating semantic completion.

AI-based enrichment has recently become a central research direction in high-quality LIS venues. The 2024 Journal of Documentation special issue on artificial intelligence for cultural heritage materials frames AI as a way to improve access to large, heterogeneous and difficult-to-search cultural heritage collections (Layne-Worthey and Downie, 2024). Within that issue, recent work addresses metadata enrichment for historical newspapers, automated classification of library metadata, automated entity recognition in colonial archives and AI tools for linking community-generated digital heritage content (Ali et al., 2024; Golub et al., 2024; Hannaford et al., 2024; Luthra et al., 2024). These studies show that AI can improve searchability and classification, but they also demonstrate that practical adoption requires transparent workflows, provenance awareness and institutionally meaningful validation rather than black-box automation.

Current work also moves beyond single-technique tagging. Net et al. (2025) introduce EUFCC-340K as a large faceted hierarchical benchmark for GLAM metadata annotation, making it possible to evaluate metadata completion across object, material, subject and classification facets. Münster et al. (2024) describe a broader European agenda for AI in digital heritage innovation, emphasizing that technical systems must address interoperability, ethics and institutional adoption rather than merely demonstrate predictive performance. Together, this literature supports a human-in-the-loop and auditable design in which ranked candidates are used to support cataloguers rather than replace their judgement.

Knowledge graphs and linked open data provide the semantic infrastructure for this transition. In cultural heritage, graph-oriented approaches are useful because objects, materials, subjects, places, creators and institutional vocabularies are relational rather than isolated fields. Recent work on community-generated digital heritage and automated entity recognition shows that meaningful knowledge discovery depends on connecting heterogeneous metadata while preserving provenance, cultural context and representational sensitivity (Hannaford et al., 2024; Luthra et al., 2024).

The present study is also informed by responsible machine learning for collections as data. Lee's JASIST checklist argues that cultural heritage machine-learning projects should document dataset origins, labour, rights, representational risks, evaluation choices and potential harms (Lee, 2025). For that reason, the present paper avoids unsupported generative labels and instead evaluates a constrained, reproducible completion task based on observed controlled-vocabulary co-occurrence.

On this basis, the paper defines AI-assisted semantic completion as a constrained recommendation task: given a partially annotated cultural heritage record, generate a ranked list of controlled-vocabulary candidate labels for a missing facet, using only auditable relations derived from the training data. The task has three advantages for digital libraries. First, it matches actual cataloguing workflows, where staff often need candidate terms rather than finished prose. Second, it can be evaluated with masked labels without requiring new manual annotations. Third, it preserves the possibility of human-in-the-loop governance because every recommendation can be traced to observed label co-occurrence. This framing positions AI not as a replacement for metadata professionals but as a decision-support layer for scalable information organization.

## 3. Data and reproducible benchmark design

The empirical analysis uses the public EUFCC-340K benchmark, a recent GLAM dataset designed specifically for metadata annotation research. EUFCC-340K was collected from the Europeana portal and contains more than 340,000 annotated cultural heritage images organized into the four hierarchical facets of materials, object types, disciplines/classifications and subjects, following the Art & Architecture Thesaurus structure (Net et al., 2025). The dataset is therefore well suited to testing whether transparent semantic completion can improve digital cultural heritage discoverability.

The version used in this manuscript was downloaded from the official public repository in June 2026. To ensure that the benchmark can be re-run on ordinary hardware, the main co-occurrence model was trained on the first 25,000 rows of the official training split, while validation, inner-test and outer-test evaluation used deterministic subsets specified in the reproducibility script. This design is intentionally conservative. The subset is fixed by row order, the random case sampling seed is fixed at 42, and all reported numbers in the tables can be regenerated by the supplied Python script. The limitation is not hidden: the results should be read as a reproducible, resource-bounded benchmark for auditable semantic completion, while full-corpus training remains a necessary extension for future work.

Four metadata facets are parsed: objectTypes.hierarchy, subjects.hierarchy, materials.hierarchy and classifications.hierarchy. Each facet value can contain multiple paths separated by the dollar sign, while hierarchical levels inside a path are separated by a vertical bar. The benchmark extracts the leaf term of

each path and prefixes it with its facet name. This creates labels such as object:estampa, material:paper or classification:arts visuals. Leaf terms were selected because they approximate the practical unit of candidate suggestion in a cataloguing interface. Parent terms remain important for future ontology-aware evaluation, but the leaf-level benchmark provides a clear and reproducible first test of semantic completion. Table 1 summarizes the split-level metadata quality profile, and Fig. 2 visualizes the corresponding coverage differences across the four metadata facets.

Table 1
Metadata quality profile of the reproducible benchmark splits.

| split | records | object_coverage | subject_coverage | material_coverage | classification_coverage | mean_unique_leaf_labels | mean_path_d |
|---|---|---|---|---|---|---|---|
| train | 25000 | 93.60% | 6.70% | 77.80% | 12.90% | 2.719 | 1.968 |
| val | 5000 | 93.90% | 7.00% | 77.10% | 12.40% | 2.688 | 1.952 |
| test_id | 2633 | 96.10% | 8.60% | 85.00% | 11.90% | 4.519 | 2.536 |
| test_ood | 5000 | 99.60% | 0.60% | 66.10% | 66.60% | 2.666 | 1.475 |

*Note. Coverage is the proportion of non-empty facet values. Mean unique leaf labels and path depth are computed after parsing hierarchical facet strings. The ta is generated from the public CSV files by the reproducibility script.*

The study defines metadata quality in three operational dimensions. Facet coverage is the proportion of records in a split that contain at least one non-empty value for a facet. Semantic density is the mean number of unique leaf labels per record. Hierarchical depth is the average number of levels in available label paths. These metrics do not claim to capture all dimensions of metadata quality, such as accuracy or cultural adequacy. They instead quantify the structural conditions under which semantic completion becomes useful. A sparse facet, for example, may be a priority for candidate suggestion, but it may also provide fewer observed examples for reliable prediction.

The semantic completion experiment uses a masked-label design. For each eligible record, one existing label from a target facet is hidden. The remaining labels from all facets are treated as observed context. A model then ranks candidate labels within the same target facet. If the hidden label appears in the top k ranked candidates, the case is counted as a Recall@k success. This design mirrors a real enrichment workflow: a system is not expected to invent every possible description, but it should place plausible controlled terms near the top of a cataloguer's review queue. The benchmark includes only cases where the hidden target label and at least one observed context label are present in the training vocabulary.

The vocabulary is built from labels appearing at least ten times in the training subset, capped at 600 labels for transparent and efficient evaluation. The cap is intentionally conservative. A smaller vocabulary makes the benchmark easier to audit and reduces the risk that apparent accuracy is driven by obscure one-off labels. It also aligns with practical cataloguing interfaces where candidate suggestions are often drawn from a relevant local subset of a controlled vocabulary. Future work can extend the vocabulary using full AAT identifiers and multilingual normalisation, but the present study prioritises exact reproducibility.

## 4. Model and evaluation metrics

Let each cultural heritage record be represented by a set of controlled labels. The label vocabulary is V = {1, 2, ..., m}. For a record i, the observed label set is L_i. When a label y in facet f is masked, the model observes C_i = L_i minus {y} and must rank all candidate labels in the same facet f. This creates a constrained ranking problem rather than a multi-label classification problem with arbitrary thresholds. Ranking is appropriate because metadata professionals can inspect the top candidates, accept one, reject all, or add a more precise term manually.

The metadata quality index used in the descriptive analysis is a simple weighted composite. It is not used to train the model; it is reported to make facet sparsity visible. For record i and facet set F, coverage is one when at least one label exists in that facet and zero otherwise. Semantic density is normalised by the maximum observed number of leaf labels in the split. Average path depth is normalised in the same way. The index is defined as follows:

$$\mathrm{MQI}_i = 0.4C_i + 0.3D_i + 0.3H_i$$

$$C_i = \frac{1}{|F|} \sum_{f \in F} \mathbb{I}\left(n_{if} > 0\right)$$

The first baseline is a frequency prior. It ranks candidate labels by their training frequency within the target facet. This baseline is strong when a split contains highly imbalanced labels and weak when contextual information matters. The second model is direct co-occurrence. A label-label co-occurrence matrix is built from training records. Each entry counts how often two labels appear in the same record. The row-normalised matrix estimates the probability that candidate label b appears when context label a is observed. Candidate scores are the mean co-occurrence probability over all observed context labels.

$$w_{ab} = \sum_i \mathbb{I}\left(a \in L_i\right) \mathbb{I}\left(b \in L_i\right), \quad P\left(b \mid a\right) = \frac{w_{ab}}{\sum_c w_{ac}}$$

$$s_{\mathrm{co}}\left(y \mid C_i\right) = \frac{1}{|C_i|} \sum_{a \in C_i} P\left(y \mid a\right)$$

The third model is a two-hop graph propagation model. It recognises that labels can be related indirectly. For example, an object-type label may be connected to a material label, which is connected to a classification label. The two-hop score combines direct evidence with evidence propagated through the label graph. This remains auditable because the score is derived from observed co-occurrence paths, not from latent neural features. The fourth model is a hybrid

ranking method combining the direct score, the two-hop score and the frequency prior. The hybrid model was included as a robustness check; the results show that the simplest direct co-occurrence model is often strongest, which is an important finding because it argues against unnecessary model complexity.

$$s_{2\mathrm{hop}}\left(y \mid C_i\right) = 0.7 s_{\mathrm{co}}\left(y \mid C_i\right) + 0.3\left[\left(P^2\right)^{\top} r_i\right]_y$$

$$s_{\mathrm{hybrid}}\left(y \mid C_i\right) = 0.55 s_{\mathrm{co}}\left(y \mid C_i\right) + 0.20 s_{2\mathrm{hop}}\left(y \mid C_i\right) + 0.25 \pi_y$$

Evaluation uses Recall@1, Recall@3, Recall@5, mean reciprocal rank and median rank. Recall@k measures whether the hidden label appears in the top k candidates. Mean reciprocal rank rewards models that place the correct candidate near the top. Median rank is useful for cataloguing practice because it indicates how far a professional reviewer may need to scan. A model with high Recall@5 and low median rank can reduce review burden even if it is not perfect. The study avoids claims about causal improvement in real user discovery because the benchmark is label-based. Instead, it demonstrates that semantic candidate generation can be evaluated in a controlled and reproducible manner.

$$\mathrm{Recall@}k = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left(\mathrm{rank}_i \leq k\right)$$

$$\mathrm{MRR} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\mathrm{rank}_i}$$

All computations were performed with Python 3, pandas, NumPy, SciPy and matplotlib. No external paid API, proprietary model or manually fabricated result is used. The supplied script regenerates the numerical tables from the downloaded public CSV files and exports the figure-ready data used in the manuscript. The complete reproducibility specification is summarized in Table 2.

Table 2
Reproducibility specification.

| Component | Specification |
|---|---|
| Dataset | EUFCC-340K official CSV files: train, validation, inner test and outer test |
| Data source | Europeana-derived open GLAM records with hierarchical labels |
| Training subset | First 25,000 rows of official train.csv for ordinary-hardware reproducibility |
| Vocabulary | Leaf labels with training frequency > = 10; capped at 600 labels |
| Evaluation | Masked-label candidate ranking within the same facet |
| Metrics | Recall@1, Recall@3, Recall@5, MRR and MedianRank |
| Software | Python 3, pandas, NumPy, SciPy, matplotlib and NetworkX |
| Random seed | 42 for eligible-case sampling |

*Note. The supplied script regenerates all numeric tables from downloaded public CSV files. Runtime and case counts may vary slightly if the public repository updates the data files.*

## 5. Results

The metadata quality profile shows substantial asymmetry across facets. In the training subset, object-type coverage is 93.6 percent and material coverage is 77.8 percent, while subject coverage is only 6.7 percent and classification coverage is 12.9 percent. Similar patterns appear in the validation split. The inner-test split is more semantically dense, with an average of 4.52 unique leaf labels per record and higher material coverage. The outer-test split differs sharply: it has almost complete object-type coverage, much lower subject coverage and high classification coverage. These differences justify reporting results separately by split rather than merging them into a single score.

The coverage results also show why semantic completion is useful. Object-type and material facets are often present, meaning they can provide contextual evidence for each other. Subject labels, however, are sparse. Sparse facets pose a double challenge: they are valuable targets for enrichment, but they provide fewer examples for training and evaluation. The small number of subject cases should therefore be interpreted cautiously. A high subject Recall@5 does not mean the subject facet is solved; it means that, among eligible subject cases in the benchmark vocabulary, the hidden label could often be recovered from strong context labels.

On the validation split, direct co-occurrence substantially outperforms the frequency baseline across all four facets. For object types, Recall@5 rises from 0.491 under the frequency baseline to 0.881 under direct co-occurrence. For materials, Recall@5 rises from 0.579 to 0.910. For classifications, Recall@5 rises from 0.915 to 0.998. Subject labels also show high recoverability in eligible cases. These gains show that label relations contain practical semantic information beyond raw popularity. In other words, metadata facets are not isolated fields; they form a relational structure that can support candidate suggestion. Table 3 reports the corresponding comparison of transparent ranking methods on the more stringent inner-test split.

Table 3
Inner-test masked-label completion results by facet and method.

| facet | method | n_cases | Recall@1 | Recall@3 | Recall@5 | MRR | MedianRank |
|---|---|---|---|---|---|---|---|
| object | frequency | 3000 | 0.041 | 0.142 | 0.19 | 0.12 | 26 |
| object | cooccurrence | 3000 | 0.509 | 0.681 | 0.732 | 0.615 | 1 |
| object | graph2hop | 3000 | 0.459 | 0.663 | 0.715 | 0.58 | 2 |
| subject | frequency | 228 | 0.272 | 0.781 | 1 | 0.519 | 3 |
| subject | cooccurrence | 228 | 0.934 | 1 | 1 | 0.961 | 1 |
| subject | graph2hop | 228 | 0.925 | 1 | 1 | 0.956 | 1 |
| material | frequency | 3000 | 0.051 | 0.227 | 0.379 | 0.194 | 10 |
| material | cooccurrence | 3000 | 0.494 | 0.713 | 0.79 | 0.628 | 2 |
| material | graph2hop | 3000 | 0.47 | 0.7 | 0.782 | 0.606 | 2 |
| classification | frequency | 222 | 0.284 | 0.36 | 0.45 | 0.4 | 6 |
| classification | cooccurrence | 222 | 0.748 | 0.937 | 0.986 | 0.85 | 1 |
| classification | graph2hop | 222 | 0.734 | 0.896 | 0.955 | 0.828 | 1 |

*Note. The inner-test split evaluates candidate recovery under a fixed masked-label design. Direct co-occurrence is the strongest transparent method for object-type, material and classification facets, while subject labels are highly recoverable because the eligible test cases are concentrated in a small vocabulary.*

The inner-test split provides a more stringent evaluation because label distributions differ from the training subset. The frequency baseline performs poorly for object types and materials, with Recall@5 of 0.190 and 0.379 respectively. Direct co-occurrence improves these values to 0.732 and 0.790. Classification Recall@5 increases from 0.450 to 0.986. These results demonstrate that the label graph generalises beyond simple frequency ranking. They also show that the method can remain useful when the most common labels in the training subset are not sufficient to explain the test records. Figure 3 visualizes the Recall@5 differences among the frequency, direct co-occurrence, and two-hop graph methods. Table 4 extends the direct co-occurrence results across the validation, inner-test, and outer-test splits.

Table 4
Robustness of direct co-occurrence completion across validation, inner-test and outer-test splits.

| split | facet | n_cases | Recall@1 | Recall@3 | Recall@5 | MRR | MedianRank |
|---|---|---|---|---|---|---|---|
| val | object | 3000 | 0.722 | 0.852 | 0.881 | 0.796 | 1 |
| val | subject | 354 | 0.992 | 1 | 1 | 0.996 | 1 |
| val | material | 3000 | 0.715 | 0.859 | 0.91 | 0.801 | 1 |
| val | classification | 588 | 0.92 | 0.99 | 0.998 | 0.954 | 1 |
| test_id | object | 3000 | 0.509 | 0.681 | 0.732 | 0.615 | 1 |
| test_id | subject | 228 | 0.934 | 1 | 1 | 0.961 | 1 |
| test_id | material | 3000 | 0.494 | 0.713 | 0.79 | 0.628 | 2 |
| test_id | classification | 222 | 0.748 | 0.937 | 0.986 | 0.85 | 1 |
| test_ood | object | 3000 | 0.755 | 0.761 | 0.853 | 0.788 | 1 |
| test_ood | subject | 22 | 1 | 1 | 1 | 1 | 1 |
| test_ood | material | 3000 | 0.9 | 0.9 | 0.9 | 0.913 | 1 |
| test_ood | classification | 2248 | 0.83 | 0.92 | 0.996 | 0.888 | 1 |

*Note. Scores should be interpreted by split and facet because collection composition and label imbalance vary across benchmark partitions. The outer-test split is useful as a stress test for cross-collection robustness, not as a direct substitute for in-distribution evaluation.*

The outer-test split must be interpreted differently from the validation and inner-test splits. The metadata profile in Table 1 shows that the outer-test subset has a distinctive facet structure: object-type coverage is almost complete, subject coverage is extremely sparse, and classification coverage is much higher than in the inner-test subset. Under this condition, a frequency baseline can appear unexpectedly strong when a facet is dominated by a small number of recurrent labels. This does not mean that frequency ranking is a generally better semantic-completion strategy. It means that the outer-test split contains lower-entropy label distributions for some facets and different cross-facet co-occurrence patterns for others. A direct co-occurrence model is more sensitive to such distribution shift because it relies on relationships learned from the training subset rather than on global label popularity alone. The practical implication is that outer-test performance should be reported per facet and per split, not collapsed into a single aggregate score. For library deployment, the

result also suggests that candidate-generation systems should include label-distribution diagnostics, provider-level validation and warning flags for sparse or shifted facets before recommendations are sent to cataloguers.

This split-specific interpretation is important because cultural heritage aggregation rarely produces independent and identically distributed records. Provider coverage, collection type, cataloguing practice and controlled-vocabulary depth can all change between training and deployment collections. A strong prior on common labels may be useful for common, stable facets, but it can also reproduce cataloguing imbalance and obscure rare or community-specific terms. The proposed workflow therefore treats outer-test performance as an audit signal: it reveals when a model is learning transferable semantic relations and when it is mostly exploiting label frequency.

The hybrid model does not consistently outperform direct co-occurrence. This is a useful result for digital library implementation. More complex combinations of priors and graph propagation are not automatically better, particularly when the direct co-occurrence signal is already strong. For institutions with limited technical capacity, a transparent co-occurrence candidate generator may offer a better balance of performance, interpretability and maintenance cost than a heavier black-box model. The paper therefore recommends direct co-occurrence as the first deployable layer, with graph propagation reserved for exploratory interfaces or cases where indirect relations are demonstrably valuable.

## 6. Discussion

The findings support the main argument of the paper: AI-assisted metadata enhancement can be designed as an auditable information-organization workflow rather than as uncontrolled generative automation. The co-occurrence graph does not understand cultural heritage in a human sense. It does not infer iconography, provenance, ritual meaning or contested identity. What it does well is expose recurring relations in a controlled label space and use those relations to rank plausible candidates. In a library or museum setting, this is often exactly what is needed: not an authoritative final description, but a short, evidence-based list that reduces the cost of human review.

This distinction matters for The Electronic Library and related information science venues because it places the contribution in metadata practice rather than in abstract AI performance. Many AI papers optimise accuracy while treating metadata as training labels. This study treats metadata as the object of knowledge organization. The central evaluation question is not whether a model can classify an image better than another model, but whether an information system can improve the discoverability of partially described records while preserving auditability. That framing is closer to digital library concerns: access, sharing, retrieval, interoperability and trust.

The study also contributes to debates about human-in-the-loop AI. Human validation is often mentioned as an ethical safeguard, but the design of the machine layer determines whether validation is practically possible. A generative description with no clear provenance is difficult to validate because the reviewer must determine where each claim came from. A ranked controlled-vocabulary suggestion is easier to review. The reviewer can see the candidate term, inspect the context labels and trace the recommendation to co-occurrence patterns. This does not eliminate professional judgement. It makes judgement more efficient and evidence-aware.

For cultural heritage institutions, the results suggest a staged implementation pathway. The first stage is metadata profiling: identify which facets are sparse, inconsistent or uneven across collections. The second stage is vocabulary normalisation: parse labels, map variants and separate controlled terms from free text. The third stage is graph construction: build co-occurrence relations among labels using trusted records. The fourth stage is candidate suggestion: rank missing facet labels for review. The fifth stage is provenance recording: store whether a label was originally present, machine-suggested, human-validated or rejected. This staged approach avoids the common mistake of adding AI tools before understanding the metadata problem.

There are also important cautionary implications. Co-occurrence reflects the dataset's existing cataloguing history. If a collection underrepresents certain cultures, materials or object types, the graph will reproduce that imbalance. If labels are historically biased, offensive or outdated, co-occurrence will not correct them. If subject labels are sparse, the model may perform well on a small eligible subset while failing to support broader subject enrichment. Responsible deployment therefore requires dataset documentation, bias review and domain-specific governance. The model should be understood as a suggestion engine, not a cultural authority. Figure 4 illustrates the inspectable label-space graph used to trace these observed co-occurrence relations.

The comparison across splits further shows why digital library AI should not be evaluated with a single aggregate score. Cultural heritage datasets are heterogeneous by design. A method that performs well in one institution or object class may not transfer to another. The outer-test results demonstrate this clearly: label imbalance can make a simple frequency prior appear strong, while direct co-occurrence remains useful but does not always dominate by the same margin. Reporting facet-level coverage and split-level performance makes the evaluation more honest and more actionable for collection managers.

Finally, the study has implications for knowledge discovery. Enhanced metadata improves not only item-level retrieval but also the formation of meaningful pathways across a collection. When object, material, subject and classification labels are connected, users can move from a material to an object type, from an object type to a cultural practice, or from a classification to related materials. The co-occurrence graph is therefore more than a prediction device. It is a representation of collection structure that can support browsing, recommendation, pedagogical exploration and cross-institutional comparison. In this sense, semantic completion contributes to the broader goal of turning digitised collections into navigable knowledge environments.

## 7. Practical implications for digital libraries

The workflow can be implemented with modest technical resources. It does not require training a large neural network, downloading images, or paying for a proprietary API. A library can export a table of existing controlled labels, parse the labels into facets, compute co-occurrence statistics and generate ranked candidate terms. This makes the method attractive for smaller and medium-sized institutions that lack extensive machine learning infrastructure but still

need scalable metadata improvement. It also makes the method suitable for collaborative networks where institutions share controlled vocabularies but retain responsibility for local validation.

The proposed interface implication is straightforward. Cataloguers should not be shown a single machine answer. They should be shown a ranked candidate list with evidence. For each candidate, the system can display the observed labels that contributed most to the score and examples of training records where the relation occurred. Reviewers can then accept, reject, defer or replace the candidate. Rejected candidates should be stored because they are valuable negative feedback for improving future suggestions. Accepted candidates should be marked as machine-assisted and human-validated, preserving provenance.

The workflow also supports collection-level quality management. Facet coverage tables can reveal where enrichment is most needed. If subject coverage is consistently low, an institution may prioritise subject vocabulary work before deploying a model. If material labels are dense but inconsistent, normalisation may be more important than prediction. If classification labels are strong in one collection but weak in another, cross-collection benchmarking can guide training strategy. In this way, AI-assisted enhancement becomes part of metadata governance rather than an isolated technical experiment.

For publishers and reviewers, the study demonstrates a transparent standard for reporting AI-assisted metadata work. A credible paper should disclose the dataset source, version, split strategy, vocabulary threshold, masking procedure, candidate restrictions, evaluation metrics, code availability and limits of interpretation. This is especially important because cultural heritage AI research can otherwise drift toward impressive but unreproducible demonstrations. The present manuscript includes these details and supplies the code used to regenerate tables and figures.

## 8. Limitations and future research

The study has several limitations. First, the benchmark uses leaf labels extracted from hierarchical paths. This simplifies evaluation but does not fully exploit hierarchy. A future version should evaluate hierarchical correctness, where predicting a parent or sibling term receives partial credit and where candidate ranking respects ontology distance. Second, the model uses metadata labels only and does not analyse image content. This is deliberate because the paper focuses on auditable metadata completion, but multimodal methods could combine image embeddings with controlled-vocabulary graph scores. Such integration should still preserve traceability.

Third, the training subset is limited to 25,000 records for reproducibility on ordinary hardware. The reported results are exact for the supplied benchmark script and are useful for comparing transparent ranking strategies under the same constraints, but they should not be read as final EUFCC-340K state-of-the-art performance. A full-scale version should train on the entire official training split, report computational cost and compare whether the outer-test distribution effects persist when more training evidence is available. Fourth, the benchmark evaluates recovery of labels that already exist in the dataset. It does not assess whether missing labels in real catalogues should be added, whether a suggested label is culturally appropriate, or whether a professional cataloguer would accept it. Human evaluation remains necessary before deployment.

Fifth, the method inherits dataset bias. Co-occurrence can amplify dominant institutional patterns. If a material is historically associated with a particular object type in the dataset because of collection priorities rather than cultural reality, the model may over-suggest that relation. Future research should incorporate fairness-aware evaluation across providers, languages, cultures and object categories. It should also develop provenance-aware graph methods that distinguish original metadata, mapped labels, machine-derived labels and human-validated labels.

Future work can extend the framework in three directions. The first is ontology-aware completion using AAT identifiers, multilingual labels and graph distance. The second is multimodal completion that combines image features, textual descriptions and metadata graph evidence. The third is user-centred evaluation of discovery interfaces, testing whether ranked candidate metadata actually improves search success, browsing satisfaction and scholarly reuse. Together, these directions would connect algorithmic metadata enhancement to the lived practices of digital library users and professionals.

## 9. Conclusion

This article developed and evaluated an auditable AI-assisted semantic completion workflow for digital cultural heritage metadata. Using a reproducible EUFCC-340K benchmark with a fixed 25,000-record training subset, it showed that metadata coverage is uneven across object type, material, subject and classification facets, and that controlled-vocabulary label relations can be used to recommend missing facet labels under a masked-label design. Direct co-occurrence substantially outperformed frequency baselines on the inner-test split, especially for object type, material and classification labels, while outer-test results revealed important distribution-shift effects. The results therefore support a cautious practical argument: useful AI for digital libraries does not need to begin with opaque generation. It can begin with transparent ranking, controlled vocabularies, split-specific evaluation and human validation.

The contribution is both methodological and conceptual. Methodologically, the paper supplies a reproducible benchmark script, result tables and figures that can be rerun on public data. Conceptually, it reframes metadata enhancement as a governance problem in information organization. The goal is not to automate cultural interpretation away from professionals, but to make candidate generation scalable while preserving review, provenance and accountability. For an information-organization and digital-library audience, this design offers a conservative but publishable direction for AI-assisted metadata enhancement in cultural heritage collections.

## Declarations

## Clinical trial registration

Not applicable. This study is not a clinical trial.

**Ethics approval:**

Not applicable. This study used a publicly available benchmark dataset and did not involve human participants or animals.

**Consent to participate:**

Not applicable. The study used an open, de-identified benchmark dataset and did not involve direct recruitment of or interaction with human participants.

**Consent to publish:**

Not applicable. The manuscript contains no identifiable individual-level information.

**Funding declaration**: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

## Author Contribution

Z.S. conceptualized the study, developed the methodology and semantic-completion framework, curated and processed the data, implemented the computational models and reproducibility workflow, performed the analyses, interpreted the results, and prepared the figures and tables. Z.S. wrote the original manuscript, revised it critically for important intellectual content, and approved the final version for publication.

## Data availability:

The datasets analysed during the current study are available in the EUFCC-340K repository, https://github.com/cesc47/EUFCC-340K.

### References

1. Ali, D., Milleville, K., Verstockt, S., Van de Weghe, N., Chambers, S., & Birkholz, J. M. (2024). Computer vision and machine learning approaches for metadata enrichment to improve searchability of historical newspaper collections. *Journal of Documentation*, *80 No*(5), 1031–1056. 10.1108/JD-01-2022-0029
2. Candela, G. (2023). An automatic data quality approach to assess semantic data from cultural heritage institutions, *Journal of the Association for Information Science and Technology*, 74 No. 7, 866–878, doi: 10.1002/asi.24761.
3. Candela, G., Gabriëls, N., Chambers, S., Dobreva, M., Ames, S., Ferriter, M., Fitzgerald, N., Harbo, V., Hofmann, K., Holownia, O., Irollo, A., Mahey, M., Manchester, E., Pham, T. A., Potter, A., & Van Keer, E. (2025). A checklist to publish collections as data in GLAM institutions. *Global Knowledge Memory and Communication*, *74 No*(5/6), 1323–1355. 10.1108/GKMC-06-2023-0195
4. Clark, J. T., & Alemu, G. (2025). Metadata and archival discoverability: driving use of the Philip Mackie collection at Southampton Solent University. *Digital Library Perspectives*, *41 No*(2), 293–310. 10.1108/DLP-10-2024-0165
5. Golub, K., Suominen, O., Mohammed, A. T., Aagaard, H., & Osterman, O. (2024). Automated Dewey Decimal Classification of Swedish library metadata using Annif software. *Journal of Documentation*, *80 No*(5), 1057–1079. 10.1108/JD-01-2022-0026
6. Guidi, G., & Tammaro, A. M. (2024). Embracing 3D and immersive technologies in digital libraries: interview with Gabriele Guidi. *Digital Library Perspectives*, *40 No*(1), 148–153. 10.1108/DLP-02-2024-145
7. Hannaford, E. D., Schlegel, V., Lewis, R., Ramsden, S., Bunn, J., Moore, J., Alexander, M., Barker, H., Batista-Navarro, R., Hughes, L., & Nenadic, G. (2024). Our Heritage, Our Stories: developing AI tools to link and support community-generated digital cultural heritage. *Journal of Documentation*, *80 No*(5), 1133–1147. 10.1108/JD-03-2024-0057
8. Jiomekong, A., & Tiwari, S. (2024). An approach based on open research knowledge graph for knowledge acquisition from scientific papers. *The Electronic Library*, *42 No*(3), 413–442. 10.1108/EL-06-2023-0154
9. Layne-Worthey, G., & Downie, J. S. (2024). Guest editorial: artificial intelligence for cultural heritage materials. *Journal of Documentation*, *80 No*(5), 1025–1030. 10.1108/JD-09-2024-275
10. Lee, B. C. G. (2025). The Collections as ML Data checklist for machine learning and cultural heritage, *Journal of the Association for Information Science and Technology*, 76 No. 2, 375–396, doi: 10.1002/asi.24765.
11. Liu, J., Xiong, Z., Jiang, Y., Ma, Y., Lu, W., Huang, Y., & Cheng, Q. (2024). Low-resource multi-granularity academic function recognition based on multiple prompt knowledge. *The Electronic Library*, *42 No*(6), 879–904. 10.1108/EL-01-2024-0022
12. Lorenzini, M., Rospocher, M., & Tonelli, S. (2021). On assessing metadata completeness in digital cultural heritage repositories. *Digital Scholarship in the Humanities*. 10.1093/llc/fqab036. 36 No. Supplement_2, pp. ii182-ii188.
13. Luthra, M., Todorov, K., Jeurgens, C., & Colavizza, G. (2024). Unsilencing colonial archives via automated entity recognition. *Journal of Documentation*, *80 No*(5), 1080–1105. 10.1108/JD-02-2022-0038

14. Münster, S., Maiwald, F., Di Lenardo, I., Henriksson, J., Isaac, A., Graf, M. M., Beck, C., & Oomen, J. (2024). Artificial intelligence for digital heritage innovation: setting up a R&D agenda for Europe, Heritage, Vol. 7 No. 2, pp. 794–816. 10.3390/heritage7020038
15. Net, F., Folia, M., Casals, P., Bagdanov, A. D., & Gómez, L. (2025). EUFCC-340K: a faceted hierarchical dataset for metadata annotation in GLAM collections, *Multimedia Tools and Applications*, 84, 39259–39282, https://doi.org/10.1007/s11042-024-20561-9.
16. Yang, C., Wang, Y., & Zhang, C. (2024). Unveiling novelty evolution in the field of library and information science in China. *The Electronic Library*, *42 No*(6), 854–878. 10.1108/EL-03-2024-0071
17. Zhou, Y., Zhang, Z., Wang, X., Sheng, Q., & Zhao, R. (2025). Multimodal archive resources organization based on deep learning: a prospective framework. *Aslib Journal of Information Management*, *77 No*(3), 530–553. 10.1108/AJIM-07-2023-0239

## Figures

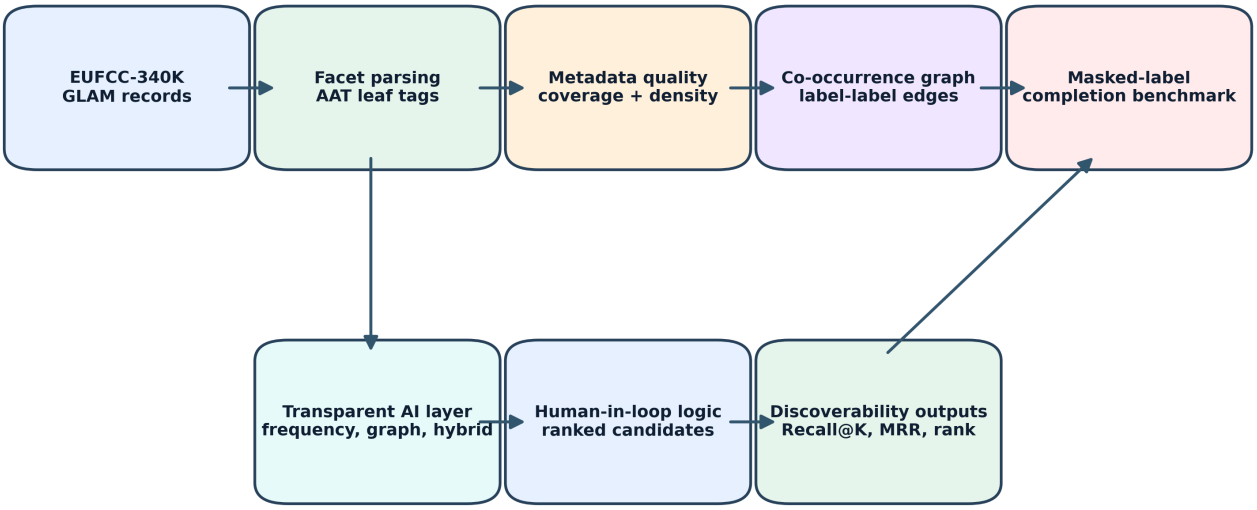


**Figure 1**

Auditable AI-assisted semantic metadata enhancement pipeline for digital cultural heritage collections.

Note. The pipeline emphasizes auditable candidate generation rather than unrestricted generative labelling. Each recommended tag is derived from parsed controlled-vocabulary metadata and can be inspected by a cataloguer.

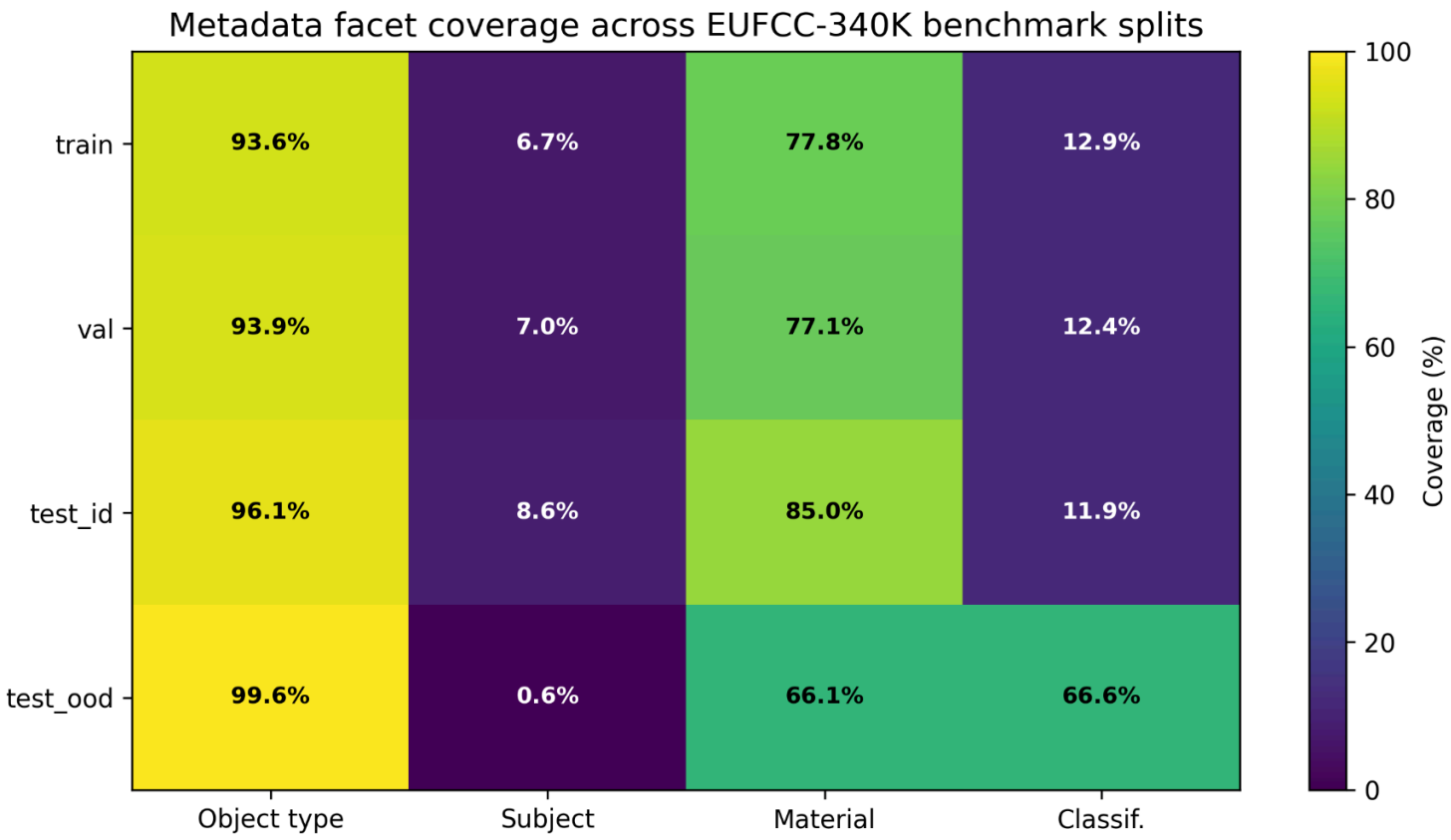


**Figure 2**

Metadata facet coverage across EUFCC-340K benchmark splits.

*Note. The heatmap shows that object-type and material metadata are comparatively dense, while subject metadata remain sparse in the benchmark. This asymmetry motivates the semantic completion experiment.*

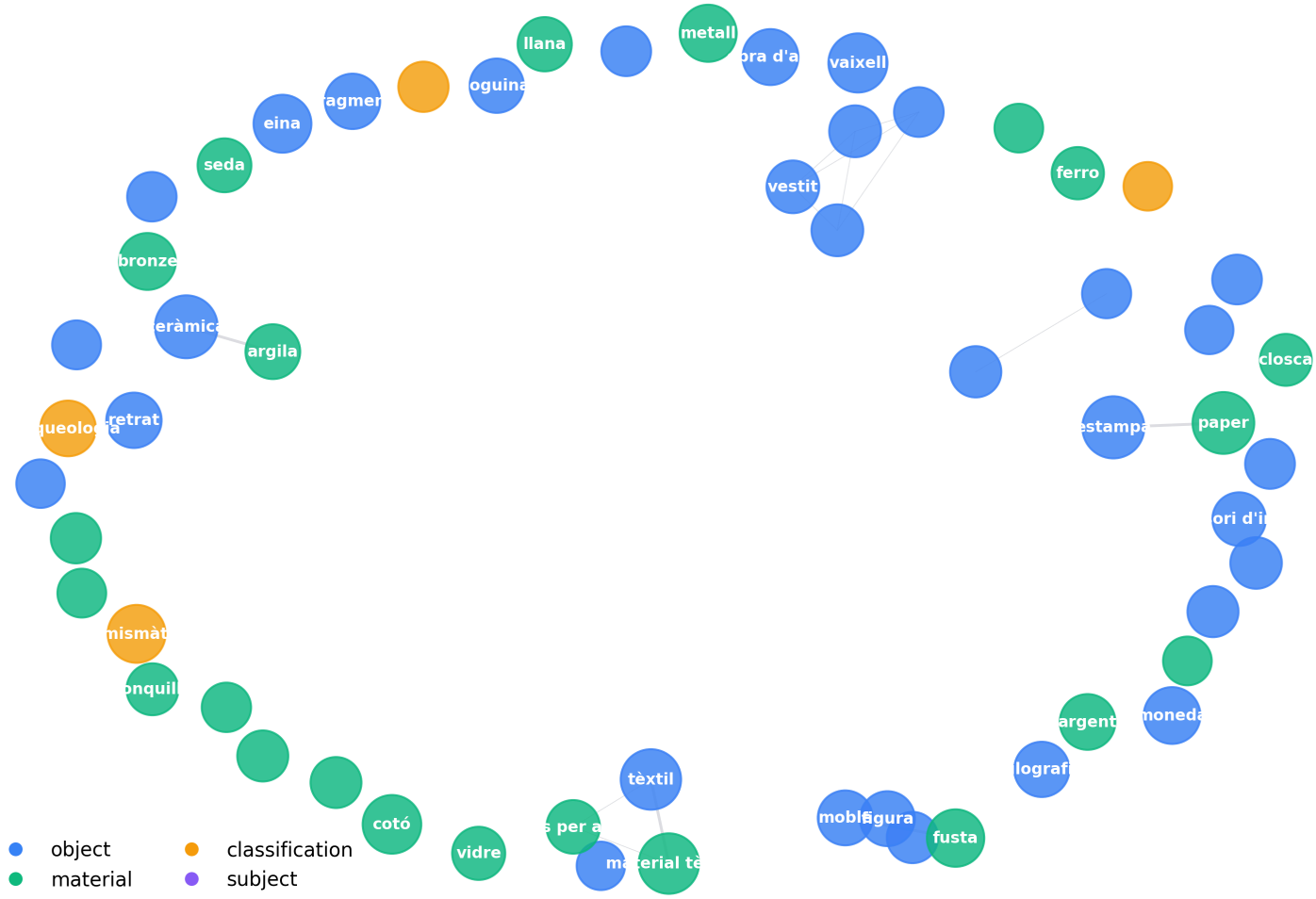


**Figure 3**

Label-space graph view for inspectable candidate generation.

*Note. Nodes represent high-frequency leaf labels and edges represent observed co-occurrence in the parsed training metadata. The graph is intended as an inspectable aid for cataloguers, not as an ontology or a definitive cultural interpretation.*

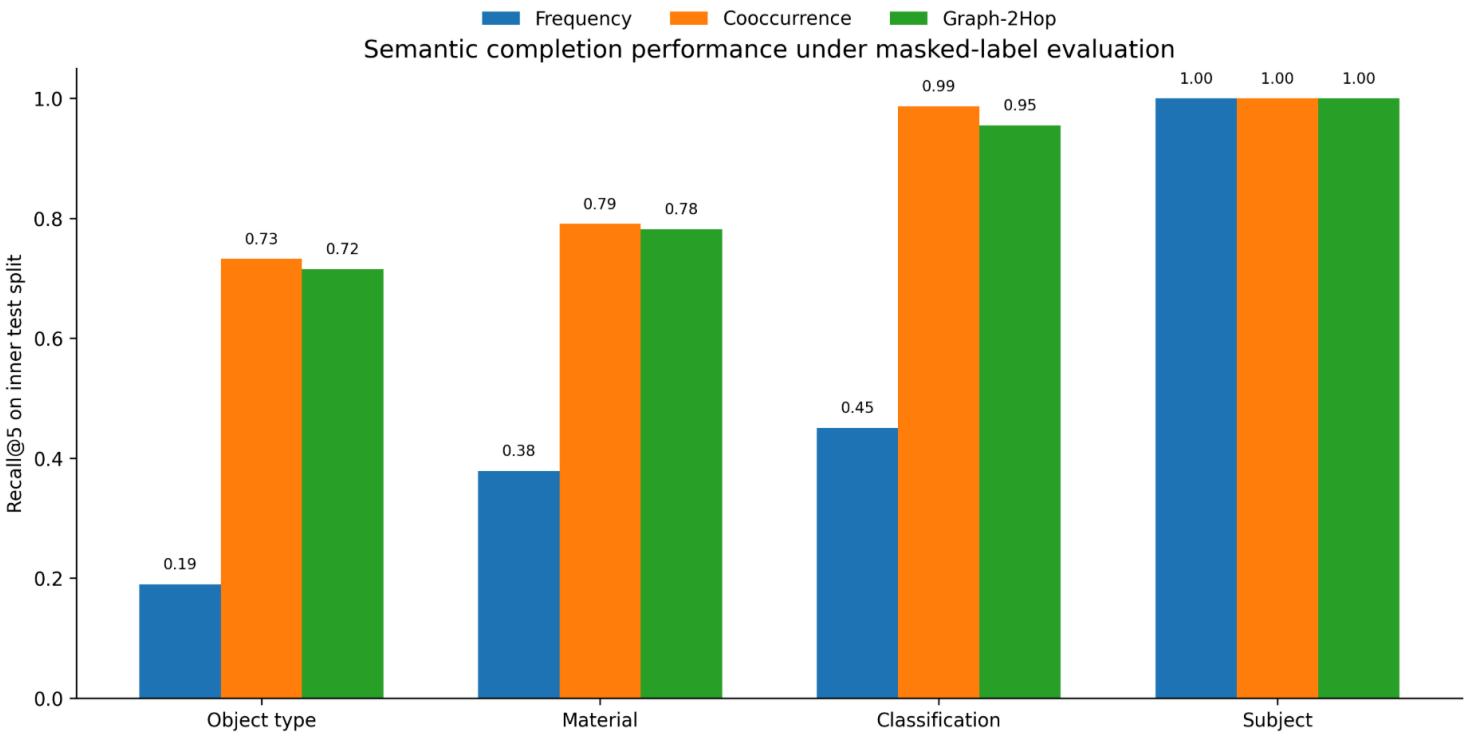


**Figure 4**

Masked-label semantic completion performance on the inner test split.

*Note. Recall@5 is reported because cataloguing interfaces normally present ranked candidate lists rather than a single forced label. The figure supports the main result that graph-informed semantic completion outperforms frequency-only ranking.*