DCPAPERS

DCMI 2026

Proceedings of the International Conference on Dublin Core and Metadata Applications 2026

ISSN: 1939-1366

When Meaning Gets Lost in Translation: Evaluating Semantic Alignment of LLM-Generated Dublin Core Metadata for Korean Cultural Heritage

Yiseul Choi [1]

[1] Sungkyunkwan University

Abstract

Cultural heritage metadata is not merely descriptive — it encodes the values, identity, and memory of communities. As large language models (LLMs) are increasingly applied to automate Dublin Core metadata generation, a critical question emerges: can AI-generated metadata preserve the cultural meaning that human catalogers carefully construct? This poster presents a research-in-progress study investigating the semantic gap between LLM-generated and expert-curated Dublin Core records for Korean cultural heritage objects. We introduce a two-dimensional meaning alignment evaluation framework and a concept we term semantic flattening — the systematic erasure of culturally specific meaning under the pressure of globally dominant classification defaults. A research design is outlined to examine how addressing semantic flattening contributes to realizing the DCMI 2026 vision of Meaning-Driven AI: Using Metadata to Align Systems with Human Values.

1. Introduction and Problem Statement

Cultural heritage collections hold irreplaceable records of human experience — artefacts, manuscripts, oral histories, and objects that embody the knowledge, identity, and values of specific communities. The metadata attached to these objects is not neutral: it shapes how communities are represented, how histories are interpreted, and whose voices are heard in the archive. In this sense, metadata for cultural heritage functions as a site of meaning-making, extending beyond technical data organization.

The rapid adoption of LLMs for automated Dublin Core metadata generation introduces a fundamental tension. While automation offers efficiency gains, it risks imposing dominant cultural frameworks on materials that require nuanced, community-specific description. Foka et al. (2025) demonstrate that every stage of the AI pipeline can inherit and amplify historical biases embedded in existing records [1], and Foka & Griffin (2024) argue that AI applied to cultural heritage may perpetuate colonial or exclusionary worldviews at scale [2]. Birhane (2021) further grounds this concern in a relational ethics framework, showing how algorithmic systems disproportionately harm marginalized communities when designed without adequate cultural consideration [3].

Korean cultural heritage presents a particularly compelling case. Rich in historically specific terminology — dynastic period names, region-based classification systems, and community-defined subject vocabularies encoded in the Korean Subject Headings (KSSH) — Korean heritage objects require metadata that preserves local meaning. When LLMs default to Western-centric classification norms, this meaning is not merely translated: it is flattened and lost. This poster introduces a framework to identify,

Poster

Available Aug 01, 2026

Correspondence to
Yiseul Choi
cis118@g.skku.edu

measure, and mitigate this loss, and presents the research design currently being implemented.

Although the empirical case is Korean, the underlying dynamic is not unique to it: any cultural or linguistic community whose descriptive vocabulary is underrepresented in LLM training distributions faces a comparable risk of dominant-culture flattening. This positions the present study as both a Korea-specific analysis and a transferable diagnostic approach for the wider Dublin Core community's engagement with Meaning-Driven AI.

2. Background and Related Work

Prior work on AI-generated Dublin Core metadata reveals a consistent pattern: LLMs perform well on factual surface elements but poorly on semantically rich fields. Kim et al. (2023) evaluated ChatGPT-generated Dublin Core records for 90 Korean monographs and found satisfactory completeness (0.87) but substantially lower accuracy (0.71), with the weakest performance concentrated in Subject and Description — the very elements most critical for cultural heritage discovery and contextual interpretation [4].

Brzustowicz et al. (2026) extend this line of inquiry to community archives, finding that ChatGPT 4o generated records with approximately 70% element accuracy but significant inconsistencies in schema conformance and source interpretation [5]. Critically, their study introduces qualitative dimensions — bias, meaning, and context — that go beyond standard completeness and accuracy metrics, pointing toward the kind of meaning-level evaluation this poster pursues.

On the ethical side, Foka & Griffin (2024) establish that bias in cultural heritage collections is inherent — rooted in historical practices of digitization, selection, and description — and that AI systems may amplify rather than neutralize it [2]. Zavalina & Burke (2021) further demonstrate the difficulty of producing high-quality Dublin Core metadata even among trained graduate students, reinforcing that metadata creation is a skilled, culturally situated practice that AI cannot yet fully replicate [6].

To improve conceptual rigor, semantic flattening is defined here as an unintended, generation-time replacement of culturally or historically specific terminology with a generalized or dominant-culture equivalent, distinguished from three adjacent concepts. Unlike semantic drift, which describes the gradual reinterpretation of a term's meaning through use over time, semantic flattening is a synchronic compression occurring within a single act of AI generation. Unlike cultural bias, which denotes systemic favoritism embedded in training data or model behavior at a structural level, semantic flattening is defined at the level of observable metadata output rather than as a claim about underlying model mechanisms. And unlike metadata homogenization — the deliberate, institution-driven standardization of vocabularies to support interoperability — semantic flattening is an unintended by-product of automated generation rather than

a chosen cataloging strategy, and concerns loss of content-level specificity rather than structural convergence.

3. Research Questions

RQ1. To what extent do LLM-generated Dublin Core records for Korean cultural heritage objects align with expert-curated records in terms of semantic meaning, beyond syntactic accuracy?

RQ2. Which Dublin Core elements exhibit the greatest semantic drift — specifically the loss of culturally specific terms in favor of dominant-culture defaults — in AI-generated outputs?

4. Methodology

This study employs a two-phase comparative evaluation design. Data collection and analysis are currently underway; the research design is described below.

4.1. Phase 1 – Dataset Construction

A sample of 60 cultural heritage records will be drawn from the National Heritage Portal of Korea, covering three domains with high cultural sensitivity: historical artefacts (Joseon Dynasty period), intangible heritage (folk performance, ritual practice), and colonial-era documentary materials. These domains were selected because accurate metadata representation requires period-specific and region-specific knowledge that lies largely outside the training distribution of general-purpose LLMs.

4.2. Phase 2 – LLM Metadata Generation and Evaluation

Dublin Core metadata will be generated for each record using LLMs under zero-shot and few-shot prompting conditions. Output will be collected for ten Dublin Core elements, with primary analytical focus on the four elements most sensitive to cultural meaning: dc:subject, dc:description, dc:coverage, and dc:relation.

Zero-shot prompting establishes a baseline reflecting typical, unprompted institutional use of general-purpose LLMs, while few-shot prompting supplies a small number of expert-curated exemplar records to test whether direct exposure to culturally specific terminology improves preservation of Korean-specific descriptors. The full study (Section 7) extends this comparison with chain-of-thought prompting, which asks the model to articulate an object's historical and cultural context before generating metadata, testing whether explicit contextual reasoning — rather than exemplar exposure alone — reduces reliance on generalized defaults. Comparing these conditions allows the study to distinguish flattening that is inherent to a model's default behavior from flattening that practitioners can mitigate through prompt design alone.

Two evaluation dimensions will be applied:

Semantic Fidelity: Cosine similarity between LLM-generated subject terms and expert-curated terms will be computed using multilingual sentence embeddings (LaBSE).

LaBSE maps text from over 100 languages, including Korean and English, into a shared embedding space, enabling direct comparison between Korean-language expert terms and LLM outputs without an intermediate translation step. For each compared element, embeddings will be generated separately for the original Korean-language term and its English equivalent, allowing flattening that occurs during metadata generation to be distinguished from separate effects introduced by cross-lingual translation. Rather than applying an arbitrary fixed threshold, cutoffs for classifying terms as flattened versus preserved will be derived empirically from the distribution of similarity scores observed during the pilot phase, used jointly with expert judgment.

Cultural Specificity: A panel of three domain experts will assess whether culturally significant terms — including KSSH subject headings, Korean dynastic period names, and regional classifications — are preserved in AI outputs or replaced by generic Western-centric equivalents. This dimension directly operationalizes the concept of semantic flattening.

5. Illustrative Example

To ground the framework in a concrete case ahead of full pilot results, this section presents a real, citable object: the National Treasure Celadon Prunus Vase with Inlaid Cloud and Crane Design (Goryeo dynasty, 12th century; sanggam inlay technique), held at the Kansong Art Museum, Seoul. Its official designation is government-standardized under the Korea Heritage Service's English Designation Rules for Cultural Heritage Names, and can be treated as an expert-curated dc:subject baseline: "Celadon; Goryeo dynasty; Maebyeong (prunus vase); Cloud-and-crane design; Sanggam inlay technique." A preliminary zero-shot generation, produced without access to this official designation, rendered the same object as "a green ceramic vase with bird and cloud motifs, believed to be from East Asia" — retaining general object categorization while omitting the dynastic period, the object-type term (maebyeong), and the technique-specific term (sanggam). This single-item comparison illustrates the intended unit of analysis for the full pilot; the complete dataset will be constructed from structured records obtainable through the e-Museum (emuseum.go.kr) open API of the National Museum of Korea and affiliated institutions.

6. Work in Progress and Expected Findings

This study is currently in the data collection and pilot design phase. Quantitative evaluation results are not yet available and will be reported in subsequent publications. However, the theoretical basis for anticipating semantic flattening is well grounded in existing literature. Kim et al. (2023) demonstrate that LLM performance is systematically weakest on Subject and Description elements — the fields in which culturally specific Korean terminology is most densely represented [4]. Brzustowicz et al. (2026) similarly find that AI-generated records exhibit meaningful inconsistencies in how primary source material is interpreted, suggesting that culturally embedded meaning is among the first aspects to be lost in automated generation [5].

Based on these patterns in the literature, we anticipate that LLM-generated records for Joseon-era artefacts will show lower semantic fidelity than records for intangible heritage items, given the greater terminological distance between period-specific Korean classifications and the Western-centric training data of general-purpose LLMs. We also expect semantic flattening to be most pronounced in dc:subject and dc:coverage — the elements that carry the heaviest burden of cultural and geographic specificity. A secondary pattern of interest, to be explored as an additional dimension in the full study, concerns whether AI-generated descriptions of materials related to historically marginalized communities reproduce dominant-culture framing — an ethical representativeness dimension that extends the core semantic drift analysis toward the broader meaning alignment goals of DCMI 2026.

7. Research Plan and Future Directions

The full study is planned for completion in 2026–2027. Evaluation will be conducted by a panel of three domain experts to enable inter-rater reliability assessment, addressing a key methodological constraint of single-reviewer designs. LLMs will be evaluated under zero-shot, few-shot, and chain-of-thought prompting conditions, allowing systematic comparison across models and prompting strategies.

The completed study aims to deliver three contributions. First, it will produce an empirically validated, two-dimensional meaning alignment evaluation framework applicable to non-Western cultural heritage contexts beyond Korea. Second, it will generate a comparative analysis of semantic drift patterns across models and prompting conditions, providing evidence-based guidance for practitioners selecting LLMs for heritage metadata workflows. Third, in direct response to the design question raised by RQ2, it will formalize human-in-the-loop metadata curation principles that explicitly prioritize cultural meaning — a concrete methodological step toward metadata systems that align with human values as called for by DCMI 2026.

8. Conclusion

Cultural heritage metadata carries meaning that belongs to communities — it is not a neutral technical artifact. This poster argues that the automation of Dublin Core generation for Korean cultural heritage objects risks producing semantic flattening: the substitution of community-specific cultural meaning with globally dominant classification defaults. This is precisely the failure mode that the DCMI 2026 theme of Meaning-Driven AI seeks to address.

While quantitative results are forthcoming, the theoretical framework and research design presented here offer a replicable foundation for evaluating AI-generated metadata not only for accuracy, but for the human values it encodes or omits. As AI systems become more deeply embedded in the management of cultural memory, ensuring that metadata remains a genuinely meaning-driven practice constitutes a technical and ethical obligation. This research-in-progress presents a structured approach toward that objective.

References

[1] A. Foka, G. Griffin, D. Ortiz Pablo, P. Rajkowska, and S. Badri, "Tracing the bias loop: AI, cultural heritage and bias-mitigating in practice," *AI & Society*, vol. 40, pp. 5835–5847, 2025, doi: 10.1007/s00146-025-02349-z.

[2] A. Foka and G. Griffin, "AI, cultural heritage, and bias: Some key queries that arise from the use of GenAI," *Heritage*, vol. 7, no. 11, pp. 6125–6136, 2024, doi: 10.3390/heritage7110287.

[3] A. Birhane, "Algorithmic injustice: A relational ethics approach," *Patterns*, vol. 2, no. 2, p. 100205, 2021, doi: 10.1016/j.patter.2021.100205.

[4] S. Kim, H. Lee, and Y. Lee, "Quality evaluation of automatically generated metadata using ChatGPT: Focusing on Dublin Core for Korean monographs," *Journal of the Korean Society for Information Management*, vol. 40, no. 2, pp. 183–209, 2023, doi: 10.3743/KOSIM.2023.40.2.183.

[5] R. Brzustowicz, A. Marchetti, and C. Stevenson, "'Appears to be about': An evaluation of AI-generated metadata quality for community archives," *Information Research*, vol. 31, no. 1, 2026, doi: 10.47989/ir3162419.

[6] O. L. Zavalina and M. Burke, "Assessing skill building in metadata instruction: Quality evaluation of Dublin Core metadata records created by graduate students," *Journal of Education for Library and Information Science*, vol. 62, no. 4, pp. 423–442, 2021, doi: 10.3138/jelis.62-4-2020-0083.

[7] UNESCO, "Recommendation on the ethics of artificial intelligence," United Nations Educational, Scientific and Cultural Organization, 2021. [Online]. Available: https://unesdoc.unesco.org/ark:/48223/pf0000381137