What AI Gets to See: Auditing the Loss of Multilingual Community Knowledge in Machine-Facing Metadata on Wikimedia Commons

Anonymous ACL submission

Abstract

AI systems increasingly inherit what public knowledge infrastructures make machine-visible. Yet multilingual NLP still tends to frame data responsibility around new corpora, annotations, and benchmarks, even though public knowledge communities have already produced rich multilingual knowledge that may be lost when AI-facing infrastructures compress it into narrower metadata layers. We study this problem on Wikimedia Commons, where multilingual captions and descriptions coexist with the English-dominant category system and Wikidata-linked Structured Data on Commons. Using 2,500 featured images across five domains, we introduce a representational recovery framework that audits lexical, semantic, and retrieval-based recovery from current machine-facing metadata layers.

Across these tests, we find a consistent asymmetry. After removing the exact query text from the gold image's free-text document, Commons categories plus Wikidata entity labels outperform the remaining free text for English retrieval. For non-English retrieval, however, the remaining multilingual captions and descriptions outperform categories plus Wikidata entity labels rendered in the query language. Entity-label rendering improves non-English retrieval, but does not fully recover the meaning already present in contributor-written multilingual text. These results show that multilingual NLP should not only collect new data, but also audit, identify, and repair recovery gaps in public knowledge infrastructures as they increasingly feed large-scale AI training pipelines.

1 Introduction

Multilingual NLP increasingly invests in new corpora, annotations, and benchmarks. This work is essential, yet it leaves a complementary question underexamined: public knowledge communities have already produced rich multilingual knowledge, often over decades and in forms not designed primarily as machine-learning datasets. As AI systems increasingly draw on these public knowledge infrastructures, and as large-scale training pipelines aggregate public knowledge at scale, NLP must ask what existing multilingual community knowledge is preserved, lost, or made less recoverable when it is transformed into machine-facing representations, rather than focusing only on what multilingual data should be newly collected. We study this as a problem of representational recovery: when rich contributor-written multilingual text is operationalized through narrower metadata layers, how much of its meaning remains recoverable, where do recovery gaps appear, and how can those gaps guide repair?

Wikimedia Commons[1] makes this problem visible at scale. It contains more than 140 million freely usable media files contributed by internet users around the world, including contributors from more than 300 language editions of Wikipedia and its sister projects. For more than twenty years, these contributors have shared, described, explained, and contextualized images of cultural heritage, historical events, art, science, and everyday life in their own languages (Yu and McDonald, 2023). Commons is therefore one of the world's largest free image repositories, and a living multilingual record of how communities make the visual world legible to one another. It is also deeply entangled with multilingual multimodal AI infrastructure: Wikipedia-based resources such as WIT aggregate tens of millions of multilingual image-text examples from the broader Wikimedia ecosystem (Srinivasan et al., 2021). This makes the representational choices embedded in Commons metadata consequential not only for search, but for what AI systems learn to see.

Commons also makes visible a central representational tension. Its most important search-facing

[1] https://commons.wikimedia.org/wiki/Main_Page

metadata layer has long been the category system (Yu and McDonald, 2022): a broad community-maintained folksonomy that organizes images for search, discovery, and reuse. Categories carry substantial contributor knowledge about what an image depicts, how it should be grouped, and how others might find it. Yet this layer is overwhelmingly English-dominant. Commons can therefore contain rich multilingual captions and descriptions while still placing a higher barrier on non-English contributors whose knowledge is less likely to enter its central search-facing representation layer unless it is translated into an English-dominant category system (Yu and McDonald, 2022). Structured Data on Commons (SDC) was introduced in part to move beyond this limitation. By linking media files to Wikidata-linked *Depicts* entities, SDC promised a more machine-readable and semantically interoperable metadata layer whose labels could be rendered across languages (Yu and McDonald, 2026). In principle, this design could make Commons' visual knowledge more accessible to non-English users and make non-English contributors' interpretations more likely to enter machine-facing metadata. Recent community-grounded work further shows that Commons contributors valued these goals of improved interoperability and multilingual support, while questioning whether SDC would genuinely extend Commons' multilingual knowledge infrastructure or reproduce older limitations in a different form (Yu and McDonald, 2026). Whether that promise has been realized, and for whom, remains an open empirical question.

In this paper, we answer one part of that question: how much multilingual community-authored meaning remains recoverable from Commons' current machine-facing metadata layers, and whether that recovery is distributed evenly across languages. We introduce a representational recovery framework and apply it to 2,500 featured Commons images across five domains. We evaluate recovery in three ways: whether words from contributor-written multilingual captions and descriptions appear in Commons categories and Wikidata-linked entity metadata; whether these machine-facing layers remain semantically close to contributor-written text in a multilingual embedding space; and whether each layer can retrieve the correct image when used for metadata-based image search.

Across these tests, we find a consequential asymmetry. After removing the exact query text from the gold image's free-text document, categories combined with Wikidata entity labels outperform the remaining free text for English retrieval (73.35 vs. 56.09 Recall@1). For non-English retrieval, however, the remaining multilingual captions and descriptions outperform categories combined with Wikidata entity labels rendered in the query language (51.95 vs. 46.10). Entity-label rendering improves non-English retrieval relative to unlocalized entity labels, but it does not fully recover the meaning already present in contributor-written multilingual text. Supporting analyses show the same pattern of uneven recovery across surface-form coverage, semantic alignment, the largest non-English language slices, paired query-level comparisons, most domains, and a stricter cross-field retrieval probe. What remains underrecovered is not decorative wording. It is community-authored interpretive care expressed through multilingual text: historical, cultural, relational, and explanatory knowledge that contributors encode when they describe why an image matters.

These findings reframe a core responsibility for NLP. The field should not only collect more multilingual data. It must also audit and repair the infrastructural loss of multilingual community knowledge already present in public archives. Representation design is not a downstream implementation detail. It helps determine what AI systems get to see, whose descriptions survive the transition into machine-facing infrastructure, and where contributor-facing repair should begin. We make four contributions: (1) a representational recovery framework for auditing multilingual community knowledge across machine-facing metadata layers; (2) evidence that Commons' current category-and-entity metadata layers recover English-accessible semantics more successfully than non-English semantics; (3) a demonstration that SDC-enabled entity-label rendering helps but does not close the multilingual recovery gap; and (4) a recovery-gap prioritization method for contributor-facing repair, which identifies files where multilingual meaning is already present in free text but remains poorly absorbed by current machine-facing metadata layers.

2 Related Work

Multilingual and multicultural NLP has increasingly expanded beyond English-centric evaluation through multilingual models and benchmarks, cross-lingual evaluation datasets, and culturally

grounded vision-language tasks (Conneau et al., 2020; Hershcovich et al., 2022; Conneau et al., 2018; Ponti et al., 2020; Liu et al., 2021; Nayak et al., 2024; Romero et al., 2024). These efforts are essential, but they primarily treat multilingual and cultural knowledge as something to be newly collected, annotated, or benchmarked. Our work asks a complementary question: whether multilingual knowledge that communities have already built remains recoverable when transformed into machine-facing infrastructure. Prior work shows that image descriptions vary across languages in semantic content and expressive scope (Ye et al., 2025), suggesting that multilingual free text carries meaning that machine-facing metadata layers may not fully capture. Wikidata's multilingual linked-data model supports cross-lingual entity labels (Erxleben et al., 2014; Pellissier Tanon et al., 2016), and Structured Data on Commons extends this model to media files through Wikidata-linked *Depicts* entities. Recent community-grounded work found that Commons contributors valued SDC's goals of interoperability and multilingual support while questioning whether it would genuinely extend Commons' multilingual knowledge infrastructure or reproduce older limitations (Yu and McDonald, 2026). We turn that concern into an empirical audit: how much contributor-written multilingual meaning remains recoverable from Commons' current machine-facing metadata layers, and for whom? Concurrent work studies the complementary data-preparation question of when legacy Commons categories can serve as candidate supervision for future AI-assisted Depicts tagging (Anonymous, 2026).

3 Data

We study 2,500 featured images from Wikimedia Commons. Featured pictures are images that the Wikimedia Commons community has selected as among the highest-quality images on the site, and the Featured Pictures page[2] organizes them into subject galleries such as animals, people, places, plants, and natural phenomena. We sample five content domains from these galleries, namely animals, nature, people, places, and plants, and construct domain-balanced subsets of 500 images each. Featured images provide a comparatively high-quality, curatorially salient slice of the Commons ecosystem. This lets us ask a conservative version of our central question: if multilingual recovery gaps remain visible even among well-described files, they are unlikely to be explained only by poor curation or sparse metadata. For each image, we collect four metadata sources used throughout the analysis: Commons categories, Wikidata-linked *Depicts* entities, multilingual captions, and multilingual descriptions. Table 1 summarizes the resulting dataset. Additional dataset construction and implementation details are provided in Appendix B.

[2] https://commons.wikimedia.org/wiki/Commons:Featured_pictures

Metric	Animals	Nature	People	Places	Plants
Rows	500	500	500	500	500
Avg. caption languages	0.99	3.94	1.55	1.19	1.53
Avg. description languages	2.94	5.95	3.56	4.36	3.68
Avg. categories	4.09	10.43	6.92	6.12	5.73
Avg. Depicts	1.52	4.27	2.56	3.58	2.90
Extra description languages	1.95	2.01	2.01	3.17	2.15
Exact duplicate pairs (%)	17.59	6.23	16.80	13.22	6.93

Table 1: Domain-level summary of the featured-image dataset.

4 Recovery Gap Framework

We frame the problem as a *representational recovery gap*: when a richer human-authored multilingual text layer is operationalized through narrower machine-facing metadata, how much of the original meaning remains recoverable, and for whom? This framing is descriptive rather than information-theoretic. We do not assume that machine-facing metadata should reproduce free text verbatim, or that every contextual detail ought to become a structured statement. Instead, we ask whether machine-facing metadata remains sufficiently accountable to the multilingual knowledge contributors have already supplied, especially in access tasks such as retrieval.

In the Commons setting, the *source layer* is contributor-written captions and descriptions across languages. The *machine-facing metadata layer* consists of Commons categories and Wikidata-linked *Depicts* entities, evaluated both separately and in combination. We compare five core metadata channels: `categories`, `entities`, `localized_entities`, `category_plus_entity`, and `category_plus_localized_entity`. The localization ablation re-renders *Depicts* labels in the query language when a Wikidata-native la-

bel is available, while leaving Commons category strings unchanged. This gives the metadata side a realistic multilingual repair without introducing external translation or redesigning the existing category layer. We therefore treat `category_plus_localized_entity` as our primary localization-repaired comparison channel.

We audit recovery through three distinct but complementary notions. First, *lexical recovery* measures how much of the vocabulary in contributor-written captions and descriptions is directly present in each metadata channel. This captures surface-form accessibility, but it does not estimate full semantic preservation: a token may be absent while its meaning is represented indirectly, and a token may be present without preserving the contributor's intended context. Second, *semantic recovery* tests whether machine-facing representations remain close to contributor-written multilingual text in a shared embedding space. This captures meaning beyond exact token overlap, but it remains an alignment diagnostic rather than a retrieval result: high semantic similarity does not guarantee that an image will be found, and low similarity does not identify which specific meaning was lost. Third, *functional recovery*, operationalized here through retrieval, evaluates whether each metadata channel can retrieve the correct image when contributor-written captions or descriptions are used as queries. This captures whether metadata remains useful for access, which is central to Commons' search-facing and AI-facing roles. For the central retrieval comparison, we pair localized category-and-entity metadata with a leave-one-query-out free-text control that removes direct query reuse, allowing us to ask whether the current category-and-entity metadata layers still underrecover multilingual meaning after its most immediate entity-side language mismatch has been repaired.

This framework is intended to travel beyond Commons. Any platform or dataset pipeline that transforms richer human-authored text into narrower machine-facing structure can be audited through the same question: what meaning remains recoverable, for whom, and where does the gap widen?

5 Results

5.1 Layered Multilingual Metadata

We begin with the metadata layers themselves. Table 1 summarizes the domain-level structure of the featured-image dataset. Across all five domains, descriptions are available in more languages than captions, adding 1.95-3.17 additional languages per image on average. Same-language caption-description pairs are also rarely exact duplicates, ranging from 6.23% in nature to 17.59% in animals. These patterns show that Commons' multilingual text fields are layered rather than redundant: captions and descriptions often contribute distinct language coverage and non-identical content. This motivates treating captions and descriptions as complementary sources of contributor-written multilingual meaning in the recovery analyses that follow.

5.2 Lexical Recovery

We first ask how much contributor-written multilingual free text remains lexically recoverable from narrower metadata layers. For each caption or description, we compare its normalized tokens against the same image's metadata under five channels: categories, entities, localized entities, category plus entity, and category plus localized entity. This should therefore be read as a surface-form recoverability diagnostic, not as a full semantic estimate. Table 2 shows a large lexical recovery gap. The best combined metadata channel covers 0.4544 of English free-text tokens on average, but only 0.2444 of non-English tokens after entity localization. Localization improves non-English entity coverage from 0.0833 to 0.1354 and combined coverage from 0.1875 to 0.2444, but it changes little for English. Category-and-entity metadata is therefore lexically much closer to English than to non-English free text, even after entity-side localization.

5.3 Semantic Recovery

We next ask whether recovery gaps are visible beyond surface overlap. Using LaBSE (Feng et al., 2022), we embed a stratified sample of 1,000 captions and descriptions, stratified by language bucket, domain, and metadata field, including 231 English and 769 non-English queries. We then compare each query against the same image's metadata under several channels. Because the full free-text channel can contain the query itself, we distinguish source-access free text from a stricter held-out condition that removes the exact query string from the gold image's free-text document before encoding. Table 3 shows that held-out free text remains more semantically aligned with contributor-written

Language bucket	Metadata channel	Avg. metadata coverage
English	Categories	0.4077
	Entities	0.1736
	Localized entities	0.1766
	Category plus entity	0.4531
	Category plus localized entity	0.4544
Non-English	Categories	0.1792
	Entities	0.0833
	Localized entities	0.1354
	Category plus entity	0.1875
	Category plus localized entity	0.2444

Table 2: Surface-form recoverability by metadata channel.

Metadata channel	Overall	English	Non-English
Categories	0.519	0.556	0.508
Category plus entity	0.529	0.566	0.518
Category plus localized entity	0.536	0.566	0.527
Held-out free text	0.646	0.673	0.638
Source-access free text	0.697	0.794	0.669

Table 3: Sampled LABSE same-item semantic alignment by metadata channel.

queries than localized category-and-entity metadata. Overall mean similarity is 0.646 for held-out free text versus 0.536 for category-plus-localized-entity metadata. This pattern holds for both English queries (0.673 vs. 0.566) and non-English queries (0.638 vs. 0.527), which indicates that the underrecovery is not only lexical. Because exact query strings are removed from the held-out condition, this gap is not explained by direct query reuse. However, semantic alignment alone does not determine retrieval performance, so we next test item retrieval directly.

5.4 Language-Stratified Retrieval

We next test whether these representational differences affect item retrieval. Our primary retrieval analysis uses TF-IDF as an audit instrument rather than as a state-of-the-art retrieval model. This choice is deliberate. Metadata on Commons does not only support search: resources such as WIT (Srinivasan et al., 2021) aggregate raw Wikimedia image-text metadata directly as training signal for multimodal models. A dense retriever that compensates for missing cross-lingual or semantic correspondences through model-side inference answers a different question from whether the metadata layer itself preserves contributor-written multilingual meaning. We therefore ask not how well a modern retriever can compensate for metadata loss, but how much contributor-written multilingual meaning remains directly recoverable from Commons' current machine-facing metadata layers.

Each non-empty caption or description is used as a query, and the task is to rank the correct source image among all 2,500 items using one metadata channel at a time. Recall@1 reports whether the source image is ranked first. To avoid giving free text an unfair advantage from direct query reuse, our main comparison pairs `free_text_leave_query_out`, which removes the exact query string from the gold image's free-text document, with `category_plus_localized_entity`, which keeps Commons categories while rendering *Depicts* labels in the query language when available. We report source-access free text only as an upper bound.

Table 4 reveals the central crossover. For English queries, localized category-and-entity metadata substantially outperforms held-out free text, reaching 73.35 Recall@1 versus 56.09. For non-English queries, the ordering reverses: held-out multilingual free text reaches 51.95 Recall@1, compared with 46.10 for localized category-and-entity metadata. This shows that the current machine-facing metadata layers, after entity-side localization, are already highly effective for English access, but still underrecovers non-English contributor-written meaning relative to held-out free text. The localization ablation clarifies what is and is not repaired. For non-English queries, entity-only retrieval rises from 23.04 to 31.24 Recall@1 after localization, and category-plus-entity retrieval rises from 40.05 to 46.10. The English side changes very little. Entity-label rendering therefore repairs part of the metadata layers' surface-language bottleneck outside English, but it does not close the residual gap to held-out multilingual free text. Because the localized category-and-entity channel still includes English-dominant Commons categories, we interpret this result as an audit of the current metadata layers after entity-side repair, not as evidence that the remaining gap is independent of category-language asymmetry. The

Metadata channel	All queries	English	Non-English
Categories	43.50	67.14	36.43
Entities	28.12	45.11	23.04
Localized entities	34.52	45.49	31.24
Category plus entity	47.70	73.29	40.05
Category plus localized entity	52.38	73.35	46.10
Held-out free text	52.90	56.09	51.95
Source-access free text	89.31	91.81	88.57

Table 4: Recall@1 retrieval performance by metadata channel.

crossover is statistically stable under query-level bootstrap resampling. For English queries, `category_plus_localized_entity` reaches 73.35 Recall@1 with a 95% confidence interval of [71.85, 75.04], while held-out free text reaches 56.09 with [54.45, 57.64]. For non-English queries, localized category-and-entity metadata reaches 46.10 with [45.26, 46.89], while held-out free text reaches 51.95 with [50.95, 52.79]. The intervals remain separated in both language buckets, preserving the same asymmetry.

5.5 Residual Non-English Gaps

We next ask whether the residual non-English advantage of held-out free text is driven by a single large slice or appears across multiple checks. All analyses in this subsection use the same localization-repaired held-out comparison from Section 5.4: `free_text_leave_query_out` versus `category_plus_localized_entity`.

Table 5 shows that the crossover is not driven by a single non-English language. English remains the opposite case, with localized category-and-entity metadata outperforming held-out free text by 17.26 Recall@1 points. In the three largest non-English slices, however, held-out free text remains stronger: by 3.44 points in French, 2.72 in German, and 11.13 in Spanish. The smaller Vietnamese, Arabic, and Korean slices are reported as descriptive sanity checks rather than standalone population estimates. They vary in direction and should not be overinterpreted.

Table 6 shows the same pattern at the query level. Within the non-English bucket, held-out free text assigns a better gold-image rank than localized category-and-entity metadata for 32.47% of queries, while localized category-and-entity metadata does better for 24.63%. The remaining 42.90% are ties. The large tie mass does not hide widespread joint failure: among ties, most

Language	N	Localized meta.	Held-out free text	Source-access free text	Held-out gap
English	3319	73.35	56.09	91.81	-17.26
French	1208	55.12	58.56	92.48	+3.44
German	1056	57.84	60.56	90.57	+2.72
Spanish	1073	54.94	66.07	90.38	+11.13
Vietnamese	293	54.55	48.82	93.94	-5.73
Arabic	181	18.46	16.92	85.64	-1.54
Korean	158	27.44	34.15	92.07	+6.71

Table 5: Language-specific Recall@1 after entity localization and query holdout.

Comparison	Held-out better (%)	Metadata better (%)	Tie (%)
vs. category plus localized entity	32.47	24.63	42.90
vs. localized entities	50.53	20.38	29.09
vs. category plus entity	39.04	20.36	40.60

Table 6: Paired non-English query-level rank comparisons.

are shared successes, including 81.17% shared Recall@1 outcomes and 91.12% shared top-five outcomes.

Finally, Table 7 shows that the residual gap is broad across image domains but not universal. Animals is the exception: localized category-and-entity metadata remains stronger than held-out free text, consistent with more object-centric and taxonomy-friendly metadata. In nature, people, places, and plants, held-out multilingual free text remains stronger. The non-English recovery gap is therefore uneven, but it extends across multiple domains where descriptions plausibly carry contextual, social, environmental, or place-specific meaning beyond object identity.

5.6 Cross-Field Retrieval

The leave-one-query-out control removes exact query reuse, but the remaining free-text document may still contain closely related material from the same field. We therefore run a stricter cross-field probe: caption queries retrieve only against description documents, and description queries retrieve only against caption documents. Category-and-entity channels remain unchanged. Table 8 shows that cross-field free text remains highly informative for caption queries, especially outside English. For non-English caption queries, held-out description text reaches 63.14 Recall@1, compared with 41.09 for localized category-and-entity metadata. English caption queries show a smaller version of the same

Domain	Local. meta.	Held-out	Source	Gap
Animals	64.01	57.81	89.77	-6.20
Nature	36.90	47.11	83.68	+10.21
People	53.25	60.18	97.59	+6.93
Places	46.61	52.74	90.75	+6.13
Plants	46.66	49.62	86.98	+2.96

Table 7: Domain-level non-English Recall@1 after entity localization and query holdout. Local. meta. denotes category-plus-localized-entity metadata.

Query setting	Raw meta.	Local. meta.	Cross-field free text
Caption, En.	72.54	72.72	77.87
Caption, Non-En.	36.68	41.09	63.14
Description, En.	73.65	73.65	42.23
Description, Non-En.	41.53	48.31	39.18

Table 8: Cross-field Recall@1 after removing direct same-field text matching.

pattern, with cross-field free text at 77.87 versus 72.72 for localized category-and-entity metadata. The reverse direction is weaker: for description queries, category-and-entity metadata outperforms caption-only free text in both language buckets. We therefore do not claim that cross-field free text uniformly dominates category-and-entity metadata. The narrower point is that, even after removing exact query reuse and same-field matching, human-authored multilingual text can retain substantial retrieval value, especially when descriptions provide contextual information for non-English caption queries.

5.7 Dense-Retrieval Sanity Check

Our main retrieval claim relies on the localization-repaired held-out TF-IDF comparison above, because that setup most transparently audits what retrieval signal is directly recoverable from each metadata layer. We additionally run a dense-retrieval sanity check with LABSE (Feng et al., 2022), a cross-lingual sentence embedding model, to ask a narrower question: whether the broader source-access ordering, in which full multilingual free text outperforms category-and-entity metadata, is specific to sparse lexical retrieval. Because this dense rerun does not implement the leave-one-query-out control, we do not use it as evidence for the central held-out crossover in Section 5.4. Its role is only to test whether the broader source-access free-text advantage remains visible under a multilingual vector retriever. Table 9 shows that source-access free text remains

Metadata channel	**English**	**Non-English**
Categories	47.76	38.67
Category plus entity	53.40	43.20
Source-access free text	86.40	71.61

Table 9: LABSE dense source-access retrieval sanity check.

stronger than category-plus-entity metadata, which in turn remains stronger than categories alone, for both English and non-English queries. We interpret this result conservatively: it does not replace the localization-repaired leave-one-query-out analysis or establish the held-out English/non-English crossover under dense retrieval. It only shows that the broader source-access ordering is not unique to sparse TF-IDF retrieval.

6 Discussion

6.1 Partially Realized Multilinguality

Our results clarify what SDC has achieved, and where its promise remains incomplete. SDC was introduced in part to move beyond an English-dominant category system by creating a more machine-readable, Wikidata-linked metadata layer whose labels could be rendered across languages. Prior community-grounded work found that Commons contributors valued these goals, while questioning whether SDC would genuinely broaden multilingual participation and representation or reproduce older limitations in a different form (Yu and McDonald, 2026). Our audit shows that SDC-enabled entity-label rendering has accomplished something real. Category-plus-localized-entity metadata is highly effective for English retrieval and improves non-English retrieval relative to unlocalized entity labels. Yet this success is uneven: in the localization-repaired held-out comparison, localized category-and-entity metadata outperforms held-out free text for English queries, while held-out multilingual free text remains stronger for non-English queries. SDC has therefore reduced part of the language barrier created by English-dominant search-facing metadata, but it has not fully absorbed the multilingual knowledge already present in contributor-written text. The issue is not whether entity-linked metadata is useful, but whether Commons' machine-facing metadata layers have become equally accountable to the multilingual knowledge contributors already provide.

6.2 Interpretive Care

The residual gap should not be understood as the loss of decorative wording or merely longer prose. What remains less recoverable is often community-authored interpretive care: the historical, cultural, relational, and explanatory work contributors perform when they describe why an image matters. We make this gap concrete through high-residual-gap cases: images for which category-plus-entity metadata and localized category-plus-entity metadata rank the gold image poorly, while held-out multilingual free text retrieves it much more successfully. In these cases, useful multilingual meaning is already present in contributor-written text but is not carried forward into the machine-facing metadata layers used for retrieval. Appendix A illustrates three forms of underrecovery: historical-cultural contextualization in Cappadocia, religious and historical distinctions in Jerusalem, and causal environmental explanation in Levikha. These cases differ in retrieval behavior, but they point to the same limitation: categories and entities remain valuable for normalization and interoperability, yet they do not fully preserve the interpretive work already embedded in multilingual free text.

6.3 From Audit to Repair

The representational recovery gap can also become a system-level repair signal. Wikimedia communities have long developed practices for identifying and addressing knowledge gaps across gender, topic, language, region, participation, and representation (Redi et al., 2020; Wagner et al., 2015; Graells-Garrido et al., 2015; Collier and Bear, 2012; Hood and Littlejohn, 2018; Menking et al., 2019; Worku et al., 2020; Miquel-Ribé and Laniado, 2018; Kaffee et al., 2018; Wulczyn et al., 2016). These efforts are grounded in a central commitment of the Wikimedia movement: knowledge should be contributed and used by everyone. Our findings identify a related gap in the visual part of this knowledge infrastructure. In many cases, multilingual knowledge is not absent. It has already been contributed through captions and descriptions by people writing in different languages. The gap is that this knowledge has not been carried forward into the machine-facing metadata layers that support search, reuse, and AI training. This makes recovery-gap auditing useful for contributor-facing repair. Future computer-aided tagging systems could prioritize files where localized category-and-entity metadata remains weak but held-out multilingual free text remains highly informative. Rather than generating tags indiscriminately, such systems could surface these files for human review, help contributors interpret relevant multilingual captions and descriptions, and support decisions about whether missing meanings should enter machine-facing metadata, be preserved as prose, or be routed to relevant language communities.

6.4 Representational Accountability

The broader implication extends beyond Commons. Multilingual and multicultural NLP has rightly invested in new corpora, annotations, and benchmarks, but this agenda is incomplete if it treats multilingual knowledge primarily as something to be newly collected for AI. Public knowledge communities have already spent decades producing richly contextualized multilingual descriptions of the world. When AI-facing infrastructures preserve those descriptions unevenly, the problem is not only data scarcity, but infrastructural underrecovery. Commons makes this problem measurable. In the featured-image slice we study, non-English contributor-written meaning remains less recoverable from current machine-facing metadata layers even after entity-label rendering, despite those layers' real success for English access. Representation design should therefore be treated as part of NLP's data responsibility: the field should audit whether community-authored multilingual knowledge remains visible, retrievable, and repairable as it is transformed into machine-facing infrastructure.

7 Conclusion

This paper introduced a representational recovery framework for auditing how much community-authored multilingual knowledge remains recoverable after public knowledge infrastructures become machine-facing metadata. On Wikimedia Commons, current category-and-entity metadata, especially after entity-label rendering, supports English access well but still underrecovers non-English contributor-written meaning relative to held-out multilingual free text. This asymmetry shows that multilingual NLP's data responsibility includes auditing, identifying, and repairing recovery gaps in public knowledge infrastructures.

8 Limitations

This study examines a high-quality featured-image slice of Wikimedia Commons rather than the full repository. This makes our setting conservative for detecting recovery gaps, but the magnitude and language distribution of those gaps may differ in less curated parts of Commons. Our primary retrieval analysis uses TF–IDF intentionally as an audit instrument for directly recoverable metadata signal, not as a claim about best possible multilingual retrieval. Stronger dense or multimodal retrievers may compensate for some missing correspondences through model-side inference. Such compensation is valuable in deployment, but it answers a different question from whether current machine-facing metadata itself preserves contributor-written multilingual meaning. Finally, our LABSE dense retrieval rerun is a source-access sanity check rather than a leave-one-query-out replication of the central crossover. Future work should test whether the same recovery asymmetry persists under stronger multilingual and multimodal retrievers while preserving controls against direct query reuse.

9 Ethical Considerations

This study analyzes publicly available Wikimedia Commons records, including categories, Wikidata-linked *Depicts* entities, multilingual captions, and multilingual descriptions. The analysis does not involve private personal information, interviews, human subjects, crowdworkers, annotators, or non-public contributor data. The image filenames discussed in the illustrative cases are already public and searchable Commons records.

Public availability, however, does not remove ethical responsibility. Commons metadata is produced through volunteer labor and may describe culturally, historically, or politically sensitive images. We therefore focus on aggregate recovery patterns and system-level metadata behavior rather than evaluating individual contributors, languages, or communities. In the illustrative cases, our goal is to explain recovery-gap mechanisms, not to assign blame for missing or incomplete metadata.

Wikimedia Commons media and metadata are publicly available under the licenses and terms specified on the corresponding Commons file pages and Wikimedia project pages. Our released materials will include code, processed metadata tables, derived retrieval outputs, and page identifiers rather than redistributing media files. Our use of Commons records is consistent with their role as public open-knowledge infrastructure: we analyze publicly available metadata to audit aggregate recovery patterns and to support more accountable metadata enrichment.

Our repair framing is contributor-facing rather than automation-first. We do not argue that multilingual captions and descriptions should be automatically converted into categories or Wikidata-linked entity metadata. Some meanings may be better preserved as prose, some may require community or language-specific review, and some may not fit existing metadata schemas. Recovery-gap auditing should therefore support human review, community judgment, and targeted metadata enrichment rather than indiscriminate conversion of multilingual text into machine-facing metadata. The derived materials we release are intended for research, auditing, and community-facing repair workflows, not for automated replacement of community judgment or downstream profiling of contributors, languages, or communities.

References

Anonymous. 2026. How ai gets to see: Preparing community categories for ai-assisted tagging on wikimedia commons. Under review.

Benjamin Collier and Julia Bear. 2012. Conflict, criticism, or confidence: An empirical examination of the gender gap in wikipedia contributions. In *Proceedings of the ACM 2012 conference on computer supported cooperative work*, pages 383–392.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Fredo Erxleben, Michael Günther, Markus Krötzsch, Julian Mendez, and Denny Vrandečić. 2014. Introducing wikidata to the linked data web. In *International semantic web conference*, pages 50–65. Springer.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic bert sentence embedding. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 878–891.

Eduardo Graells-Garrido, Mounia Lalmas, and Filippo Menczer. 2015. First women, second sex: Gender bias in wikipedia. In *Proceedings of the 26th ACM conference on hypertext & social media*, pages 165–174.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Nina Hood and Allison Littlejohn. 2018. Hacking history: Redressing gender inequities on wikipedia through an editathon. *International Review of Research in Open and Distributed Learning*, 19(5).

Lucie-Aimée Kaffee, Hady ElSahar, Pavlos Vougiouklis, Christophe Gravier, Frédérique Laforest, Jonathon Hare, and Elena Simperl. 2018. Mind the (language) gap: Generation of multilingual wikipedia summaries from wikidata for articleplaceholders. In *European Semantic Web Conference*, pages 319–334. Springer.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Amanda Menking, Ingrid Erickson, and Wanda Pratt. 2019. People who can take it: How women wikipedians negotiate and navigate safety. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–14.

Marc Miquel-Ribé and David Laniado. 2018. Wikipedia culture gap: quantifying content imbalances across 40 language editions. *Frontiers in physics*, 6:54.

Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd Van Steenkiste, Lisa Anne Hendricks, Karolina Stanczak, and Aishwarya Agrawal. 2024. Benchmarking vision language models for cultural understanding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5769–5790, Miami, Florida, USA. Association for Computational Linguistics.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, and 1 others. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Thomas Pellissier Tanon, Denny Vrandečić, Sebastian Schaffert, Thomas Steiner, and Lydia Pintscher. 2016. From freebase to wikidata: The great migration. In *Proceedings of the 25th international conference on world wide web*, pages 1419–1428.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Miriam Redi, Martin Gerlach, Isaac Johnson, Jonathan Morgan, and Leila Zia. 2020. A taxonomy of knowledge gaps for wikimedia projects (second draft). *arXiv preprint arXiv:2008.12314*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 3982–3992.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, and 1 others. 2024. Cvqa: Culturally-diverse multilingual visual question answering benchmark. *arXiv preprint arXiv:2406.05967*.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 2443–2449.

Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier. 2015. It's a man's wikipedia? assessing gender inequality in an online encyclopedia. In *Proceedings of the international AAAI conference on web and social media*, volume 9, pages 454–463.

Zena Worku, Taryn Bipat, David W McDonald, and Mark Zachry. 2020. Exploring systematic bias through article deletions on wikipedia from a behavioral perspective. In *Proceedings of the 16th international symposium on open collaboration*, pages 1–22.

Ellery Wulczyn, Robert West, Leila Zia, and Jure Leskovec. 2016. Growing wikipedia across languages via recommendation. In *Proceedings of the 25th International Conference on World Wide Web*, pages 975–985.

Andre Ye, Sebastin Santy, Jena D Hwang, Amy X Zhang, and Ranjay Krishna. 2025. Semantic and expressive variations in image captions across languages. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29667–29679.

Yihan Yu and David W McDonald. 2022. Unpacking stitching between wikipedia and wikimedia commons: Barriers to cross-platform collaboration. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–35.

Yihan Yu and David W McDonald. 2023. " why do you need 400 photographs of 400 different lockheed constellation?": Value expressions by contributors and users of wikimedia commons. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):1–34.

Yihan Yu and David W. McDonald. 2026. Computer-aided tagging on wikimedia commons: Designing for human-ai collaboration in open knowledge work. Accepted to CSCW 2026, to appear.

A Illustrative Recovery-Gap Cases for Contributor-Facing Audit

Table 10 summarizes three illustrative non-English recovery-gap cases. Together, they show that representational underrecovery is not a single failure mode. In one case, entity localization sharply improves retrieval while still leaving historically and culturally specific multilingual meaning outside the category-and-entity metadata. In the other two, even localized category-and-entity metadata remains unable to recover the image effectively, while held-out non-English free text retrieves it strongly. These cases arise through different mechanisms, historical-cultural contextualization, culturally consequential religious and historical distinctions, and causal environmental explanation. What is being lost is not decorative wording. It is community-authored interpretive care.

Case	Cat.+ entity	Localized meta.	Held-out free text	Source-access free text
Cappadocia	2398	17	60	1
Jerusalem	2369	2424	5	1
Levikha	1001	1009	7	1

Table 10: Gold-item retrieval ranks for three illustrative non-English recovery-gap cases. Lower is better. Cat.+entity denotes category-plus-entity metadata; localized meta. denotes category-plus-localized-entity metadata.

Cappadocia: localization repairs retrieval, but not semantic preservation. For `File:Cappadocia March 2006.jpg`, the Arabic description concerns Ortahisar in Cappadocia. The existing category-and-entity metadata already contains location- and object-level cues such as Ortahisar, Cappadocia, hill, building, and categories related to tuff rock formations and views of Ortahisar. Entity localization dramatically improves the gold image's retrieval rank, from 2398 to 17. Yet the Arabic description carries substantially richer meaning than this category-and-entity metadata preserves: Göreme National Park and UNESCO World Heritage status; Roman-era and early Christian settlement; volcanic tuff formations and their erosion over time; and the region's rock-cut dwellings, shops, churches, and Byzantine artistic heritage. This case shows that retrieval access can be repaired without preserving the historical and cultural semantics contributors have already supplied.

Jerusalem: localization does not recover culturally consequential contextualization. For `File:Jerusalem-2013(2)-Aerial-Temple Mount-(south exposure).jpg`, the current category-and-entity metadata captures an aerial view of Jerusalem, the Temple Mount, al-Qibli, and the Dome of the Chain. Yet localized metadata still ranks the gold image at 2424, while held-out Arabic free text retrieves it at rank 5. The Arabic description contributes layered historical and religious contextualization that the category-and-entity metadata does not recover. It distinguishes the foreground *al-Qibli mosque* from the broader *al-Aqsa compound*; identifies al-Aqsa as the third holiest place in Islam; names the Dome of the Rock and Dome of the Chain; situates the Dome of the Chain in the Umayyad period; and includes a reference to Solomon's Temple in its historical framing of the site. This is not merely additional description. It is interpretive work that preserves distinctions central to how the image is understood.

Levikha: localization does not recover causal and explanatory meaning. For `File:Environmental disaster in Levikha Village, Sverdlovsk region of Russia DJI.jpg`, the category-and-entity metadata captures the broad topic of mining and environmental disaster but remains sparse. Localized metadata ranks the gold image at 1009, while held-out Russian free text retrieves it at rank 7. The Russian description supplies a causal account: the Levikhin-

sky copper-pyrite underground mine entered crisis, was conserved and flooded, and acidic mine waters later rose to the surface, producing an ecological disaster. It further describes neutralization efforts, discharge into the Tagil River, and unsuccessful attempts to commission an environmental harm assessment. The meaning underrecovered here is process-level explanation: not simply what the image depicts, but why the depicted scene matters.

Taken together, these cases suggest two contributor-facing uses of recovery-gap auditing. First, it can identify files where retrieval appears repaired but *semantic enrichment* is still needed, as in Cappadocia. Second, it can prioritize *high-residual-gap* cases for human review, as in Jerusalem and Levikha, where multilingual community knowledge is already present in free text but remains poorly absorbed by the current machine-facing metadata layers.

B Implementation Details

Dataset snapshot. The underlying Wikimedia Commons metadata snapshot was collected on April 20, 2026, between 23:27:54 and 23:49:42 America/Los_Angeles time. We deduplicated the collected records and retained only featured-image items with both Commons categories and Wikidata-linked *depicts* metadata, since these two metadata layers are required for the paper's recovery comparisons. The analyzed dataset was constructed from this snapshot.

Entity-label localization. For the localized entity channels, we render each Wikidata-linked *depicts* item using its existing label in the query language when available. If no query-language label is available, we fall back to the English Wikidata label and then to the original depicts label. We do not use machine translation. Commons category strings remain unchanged in all localized metadata channels.

Sparse retrieval. All TF–IDF retrieval analyses use scikit-learn's `TfidfVectorizer` (Pedregosa et al., 2011) with `token_pattern=r"(?u)\b\w+\b"`. For each metadata channel, we fit a separate vectorizer over that channel's 2,500 image documents and transform the same caption or description queries into the resulting channel-specific vector space. Text is normalized through HTML unescaping, simple wiki-markup cleanup, underscore replacement, URL removal, lowercasing, punctuation removal, and whitespace normalization. We do not apply English-specific stemming, lemmatization, or retrieval-time English stopword removal. Retrieval scores are computed from TF–IDF document and query vectors within each channel-specific space.

Leave-one-query-out retrieval. For `free_text_leave_query_out`, we normalize both the query string and the gold image's free-text document, remove exact occurrences of the normalized query from that gold document, and recompute the gold document's retrieval score against the unchanged query representation. All non-gold image documents remain unchanged. This control removes direct query reuse from the free-text side while preserving the remaining multilingual text associated with the same image.

Dense embedding retrieval. The multilingual dense-retrieval sanity check uses the LaBSE checkpoint `sentence-transformers/LaBSE`, loaded with `SentenceTransformer` (Reimers and Gurevych, 2019) through the script's `-dense-model` argument. Experiments were run with `sentence-transformers` version 5.4.1. We encode query and document texts with normalized embeddings and compute cosine-equivalent dot-product scores for retrieval. Because this rerun does not implement the leave-one-query-out control, we use it only as a source-access directional sanity check rather than as evidence for the central held-out retrieval crossover.

Materials availability. The underlying Wikimedia Commons media pages and associated public metadata remain accessible through the platform itself. We do not redistribute media files. To support reproducibility after review, we will release the full code, processed metadata tables, derived retrieval outputs, and reconstruction materials upon publication, including page identifiers for the analyzed Commons files.