Exploring Neural IR in Europeana

Suhaib Basir*[1][0009−0005−9108−5912], Mónica Marrero[2][0000−0002−2359−6340], and Julián Urbano[3][0000−0003−2933−1949]

[1] Infinite Analytics, Boston, United States
[2] Europeana Foundation, The Hague, The Netherlands,
monica.marrero@europeana.eu
[3] Delft University of Technology, Delft, The Netherlands

Abstract. Europeana is the leading digital library of Europe's cultural heritage, providing access to over 60 million items in more than 40 languages. Its search infrastructure relies on Solr and BM25 over the items' metadata, thus depending heavily on keyword matching and resource-intensive treatments such as translation and multilingual metadata enrichment. This paper explores the application of Neural Information Retrieval (NIR) approaches in Europeana, focusing on multilinguality. We created a dataset for comparative evaluation, and show that while NIR demonstrates strong potential for multilingual search, challenges remain regarding its performance, particularly for entity-centric queries. This work also highlights the need for more reliable evaluation data.

Keywords: Digital Library · Cultural Heritage · Neural IR · LLM

1 Introduction

Artificial Intelligence (AI) is increasingly deployed in digital libraries to enhance internal processes (e.g., cataloging and content enhancement) and to improve discoverability and accessibility. Somewhat mature applications include handwritten and optical character recognition, as well as enrichment initiatives like Saint George on a Bike [16], Transcribathon [5], the Cultural Heritage AI Cookbook [14], and the generation of domain-specific models, like those created in the national libraries of Norway [8] and Sweden [6]. Applications of AI in search remain more exploratory. Examples include image-based similarity search at the Bavarian State Library [2], AI-guided query reformulation at Stony Brook University Libraries [18], semantic image search at the National Museum of Norway [17], and Retrieval-Augmented Generation (RAG) efforts at Northwestern University Libraries [11].

The cultural heritage community is particularly alert to risks posed by AI, since its operations depend on public trust regarding the accuracy and reliability of information, as well as the protection of intellectual property. As a result, institutions are slowly developing strategies for responsible AI adoption

* Work as an MSc student at TU Delft [1].

(e.g., [13]). Digital libraries often lack substantial technical capacity and operate under resource constraints, making experimentation difficult. In addition, research on AI for information access rarely reflects the characteristics of CH collections: heterogeneous, dynamic, multilingual, and with a mix of structured and unstructured sparse data. This mismatch reduces the direct applicability of many state-of-the-art methods.

Europeana exemplifies these challenges. It aggregates digitized metadata for CH content from thousands of European libraries, archives and museums, forming a large and diverse metadata collection. Search is currently supported by a keyword-based Solr+BM25 system, which scales well but is limited in handling multilinguality, semantics, and contextual understanding. To mitigate these issues, Europeana enriches metadata with multilingual entity information from external sources like Wikidata, and translates queries and selected metadata fields to English as a pivot language [12, 9]. However, these steps introduce noise, especially for short queries, entities and fields with little contextual information. They are also resource-intensive, costly to maintain, and incomplete, as translation is currently not automatically applied to newly ingested metadata.

This paper summarizes Europeana's first evaluation of NIR models to better accommodate its search needs [1]. We compare them with the current pipeline and assess whether neural methods can provide better multilingual and semantic search while reducing reliance on expensive workflows.

2 Methodology

We compared the current BM25 baseline with three neural models that follow different approaches: Jina ColBERT v2, a multilingual late-interaction model that encodes token-level embeddings [7]; SBERT (Multilingual DistilUSE), a lightweight sentence transformer [15]; and BGE-M3 Hybrid, a dense+sparse model that combines semantic and lexical evidence via rank fusion [3].

Given the absence of a suitable dataset with explicit relevance judgments, we built one from Europeana's user click logs in the period April–May 2024. We started from user queries with associated clicked documents, keeping only those where both query and document languages were among the 20 most represented ones. This resulted in 23K queries and 45K clicked documents. We randomly selected 40% of queries for testing, ensuring that the joint query-document language distribution remained consistent across training and test sets.

From the full Europeana corpus of 60M documents, we retained the 40M documents with enrichment and English translation. To reduce the cost of experiments, we randomly downsampled smaller subsets of 4M and 1M documents, to which we added the 45K clicked ones. Figure 1 shows the language distributions of documents, queries and clicks. The 1M collection was used for testing, while the 4M collection was used to retrieve negatives for fine-tuning: for each training query, clicked documents served as positives, and negatives were sampled by running BM25 and taking one document per language from the bottom of the ranking.

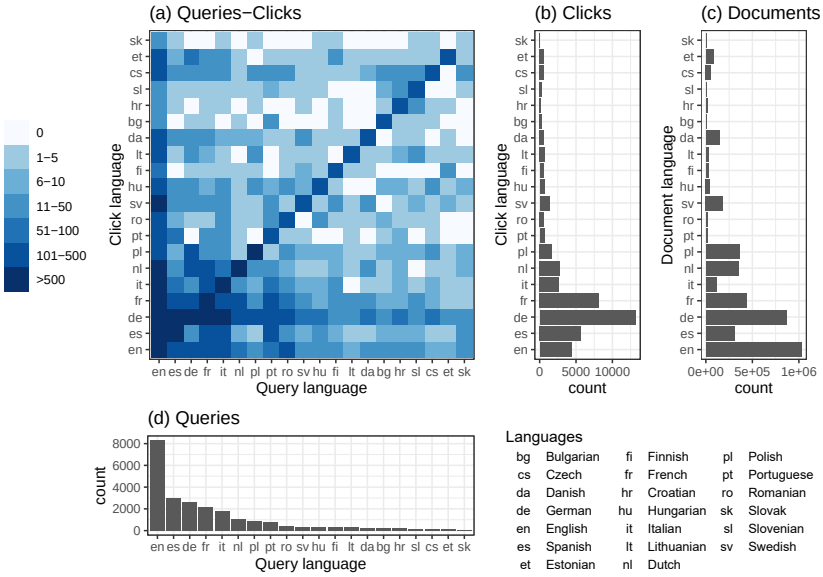

Fig. 1. The language distribution of the documents (c) shows a clear overrepresentation of English and German and, to a lesser extent, French, Polish, Dutch and Spanish. The language of clicked documents (b) follows a similar distribution, though English is much less frequent because clicks were recorded under a BM25 policy with document augmentation (i.e. non-English documents become retrievable for English queries). The language distribution of the queries (d) shows a similar bias towards popular languages, especially English. The joint query-click language distribution (a), where the diagonal reflects monolinguality, shows the expected clustering among the most popular languages in both queries and clicks: English, Spanish, German, French, Italian and Dutch.

For evaluation, the 1M test dataset was indexed in Solr for BM25, FAISS for ColBERT, and Milvus for SBERT and BGE. We indexed all metadata fields mapped to Solr-BM25's default field (e.g., `title`, `description`, `creator`). We contemplated two main factors for comparison: model (and possible fine-tuning) and augmentation (adding English translation to the query, and enrichments and/or translations to the documents). These factors are examined along two dimensions: multilinguality and effectiveness. For multilinguality we calculate the fraction (F) of retrieved documents in a language other than the query's, and the entropy (H, normalized in $[0, 1]$) of their distribution as a measure of (lack of) bias towards specific languages. For effectiveness we compute Average Precision (AP) and Recall (R). Only the top 100 documents per query were analyzed. Since English acts as the pivot language for translation and accounts for 36% of the queries, the metrics are calculated separately for English and non-English queries, and reported here only for the latter; the overall conclusions are the same.

Table 1. Multilinguality (F,H) and effectiveness (AP,R) for a selection of systems (no fine-tuning). Bare for systems without augmentation, Daug for full document augmentation (translations and enrichments), and Daug+Qaug for query augmentation.

	Bare				Daug				Daug+Qaug			
	BM25	ColBERT	SBERT	BGE	BM25	ColBERT	SBERT	BGE	BM25	ColBERT	SBERT	BGE
F	0.542	0.747	0.730	0.608	0.677	0.745	0.729	0.609	0.740	0.755	0.723	0.630
H	0.726	0.807	0.733	0.789	0.743	0.805	0.732	0.789	0.797	0.805	0.728	0.788
AP	0.752	0.323	0.034	0.379	0.775	0.343	0.031	0.365	0.703	0.306	0.039	0.318
R	0.890	0.623	0.144	0.696	0.911	0.649	0.135	0.694	0.900	0.642	0.148	0.694

3 Results

Table 1 shows that bare ColBERT and SBERT models achieve the multilinguality of BM25 when it incorporates all augmentations (F≈0.74). BGE is less multilingual (F=0.608), but still more so than a bare BM25 model (F=0.542). Document augmentation has a clear impact on the multilinguality of BM25 (ΔF=+0.135), where the effect of translations is much larger than the effect of enrichments (ΔF=+0.128 vs. +0.022). Further adding query augmentation substantially improves BM25, especially in balancing the language distribution (ΔF=+0.063, ΔH=+0.054). However, the impact of augmentations is generally limited in the neural models because they are already multilingual. In many instances, and especially with ColBERT, augmentations did even alter the rankings substantially ($\text{RBO}^{w}_{p=0.96}$ [4] below 0.35), but the overall multilinguality and effectiveness barely changed. Only query augmentation led to a mild improvement in BGE (ΔF=+0.021), but made both ColBERT and BGE less effective.

The impact of fine-tuning is generally inconsistent across models. SBERT is the exception: fine-tuning consistently increases multilinguality (ΔF=+0.118) and slightly improves language-balance; this balance weakens only in case of extreme multilinguality (H≈0.42 when F≈0.98). In addition, fine-tuning moderates the effect of document augmentation on the rankings, evidenced by a dramatic change in RBO scores in BGE (0.393 with fine-tuning vs. 0.742 without) and especially in SBERT (0.018 with vs. 0.812 without). Fine-tuning amplifies the effect of document translation in BGE (ΔF=+0.062 with fine-tuning vs. ΔF=+0.012 without). It also amplifies the effect of enrichment in SBERT (ΔF=+0.189 with fine-tuning vs. ΔF=-0.004 without), but at the cost of a dramatic drop in balance (ΔH=-0.333 with vs. ΔH=-0.003 without).

Regarding effectiveness, BM25 appears to perform much better than the neural models, with an average AP of 0.729. ColBERT and BGE average ≈0.327, and SBERT stays far below at 0.034 when fine-tuned and 0.001 when not. Recall leads to similar conclusions, with an average of 0.895 with BM25, 0.646 with ColBERT and BGE, and 0.142 with SBERT when fine-tuned and 0.001 when not. But we should take these scores with a grain of salt because of the obvious bias in the query logs: as mentioned, user clicks were recorded under a BM25 policy with document augmentation but (mostly) without query translation. Indeed, translating queries in BM25 was expected to perform better, but it showed a 9% drop in AP. To further investigate the possibility of bias, we manually inspected

a few queries in different languages where BM25 appeared to perform much better than the neural systems. We made relevance judgments and found that ColBERT and BGE actually outperformed the best BM25 by 0.08 in AP. This manual inspection also revealed that neural models seem to struggle with entity queries. This is especially concerning for Europeana because more than 50% of the user queries are entity-centric [10, 9].

4 Conclusion and Future work

This study shows that neural models can substantially enhance Europeana's capability to support multilingual and semantically nuanced search without relying on translation or metadata enrichment. Further investigation is still required before deploying neural systems in our production environment. First, our results suggest that these models do not match the effectiveness of pure lexical retrieval for entity-centric queries. Second, the system must maintain acceptable response times within reasonable infrastructure constraints, even as the collection grows dynamically. Third, it is necessary to ensure that the system does not introduce language bias by disproportionately favoring certain languages over others, especially those less represented. For instance, our experiments showed that SBERT can be highly multilingual, but at the cost of favoring some languages. The issue of language bias and fairness is particularly critical for Europeana, because our collection contains items in more than 40 languages.

Moreover, we would like to explore the broader spectrum of possibilities that AI and LLMs offer for enhancing the search process, such as retrieval-augmented generation or LLM-driven query and document augmentation. LLM-generated relevance judgments, in particular, may have a very large impact in the Cultural Heritage domain. As exemplified by this paper, one of the largest limitations is the lack of resources, both for development and evaluation. In our case, the lack of explicit judgments limits the confidence we can have when evaluating model effectiveness. In this line, Europeana will explore ways to better support experimentation, pursuing collaborative research frameworks that ensure the required user privacy while enabling research that can benefit the broader community.

5 Speaker and Company

Mónica Marrero works in search at Europeana since 2018. She holds a PhD on Named Entity Recognition, and has been involved in several projects related to knowledge representation and information retrieval with large collections.

The Europeana Foundation is an independent non-profit organisation driving the digital transformation of the cultural heritage sector and promoting the accessibility and reuse of Europe's cultural data. It leads the deployment of the common European data space for cultural heritage, an EU-funded initiative that enables open and trustworthy data sharing across Europe. Europeana also participates in EU projects spanning digitization, data aggregation, technological innovation, reuse and capacity building, artificial intelligence, and 3D.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Basir, S.: Exploring Neural IR Approaches in Europeana. Master's thesis, Delft University of Technology (2025), https://resolver.tudelft.nl/uuid:eee447b8-857a-4493-8679-1eac1d86345e
2. Ceynowa, K., Sommer, D., Hermann, M.: Digital libraries in germany: Federalism, funding, and the bayerische staatsbibliothek. In: Digital Libraries Across Continents, pp. 28–48. Routledge (2025)
3. Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., Liu, Z.: M3-embedding: Multilinguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In: Findings of the Association for Computational Linguistics: ACL 2024. pp. 2318–2335 (2024)
4. Corsi, M., Urbano, J.: The treatment of ties in rank-biased overlap. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 251–260 (2024)
5. Gordea, S., Andresel, M., Drauschke, F., Kahle, P.: Transcribathon. eu: Ai supporting collaborative transcription and enrichment of historical documents. In: Proceedings of the 2024 International Conference on Advanced Visual Interfaces. pp. 1–3 (2024)
6. Haffenden, C., Fano, E., Malmsten, M., Börjeson, L.: Making and using ai in the library: Creating a bert model at the national library of sweden. College & research libraries **84**(1) (2023)
7. Jha, R., Wang, B., Günther, M., Mastrapas, G., Sturua, S., Mohr, I., Koukounas, A., Wang, M.K., Wang, N., Xiao, H.: Jina-ColBERT-v2: A general-purpose multilingual late interaction retriever. In: Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024). pp. 159–166 (2024)
8. Kummervold, P.E., De la Rosa, J., Wetjen, F., Brygfjeld, S.A.: Operationalizing a national digital library: The case for a norwegian transformer model. arXiv preprint arXiv:2104.09617 (2021)
9. Marrero, M., Isaac, A.: Implementation and evaluation of a multilingual search pilot in the europeana digital library. In: Silvello, G., Corcho, O., Manghi, P., Di Nunzio, G.M., Golub, K., Ferro, N., Poggi, A. (eds.) Linking Theory and Practice of Digital Libraries. pp. 93–106. Springer International Publishing, Cham (2022)
10. Marrero, M., Isaac, A., Freire, N.: Automatic translation and multilingual cultural heritage retrieval: A case study with transcriptions in europeana. In: Berget, G., Hall, M.M., Brenn, D., Kumpulainen, S. (eds.) Linking Theory and Practice of Digital Libraries. pp. 133–138. Springer International Publishing, Cham (2021)
11. National Leadership Grants - Libraries. Evanston.IL: Lg-256703-ols-24 (2024), https://www.imls.gov/grants/awarded/lg-256703-ols-24
12. Neale, A., Isaac, A., Manguinhas, H., Moskalenko, D., Marrero, M.: Multilingual strategy. Tech. rep., Europeana Foundation (2020), https://pro.europeana.eu/files/Europeana_Professional/Publications/Europeana%20DSI-4%20Multilingual%20Strategy.pdf
13. O'Hare, E., Bocyte, R., Oomen, J., Aldana, L., Isaac, A., Verwayen, H., Charles, V.: Alignment paper on ai and the data space for cultural heritage:

Boundary-setters and opportunity-seekers working towards responsible ai futures. Tech. rep., Common European data space for cultural heritage (2025). https://doi.org/10.5281/zenodo.17252598
14. Rees, G., Bosse, A., Damiano, R., Isaksen, L., Yousef, T., Barker, E., Al Khatib, K., Chen, A., Daga, E., Gadd, S., Mattingly, W., Maynard, D., Palladino, C., Peeters, S., Rastinger, N.C., Ridge, M., Romanello, M., Sanderson, R., Stranisci, M.A., Thorne, W., Tjong Kim Sang, E., van Wissen, L., Marrero, M., Fantoli, M.: The cultural heritage ai cookbook (2025), https://pelagios.org/llm-lod-enriching-heritage/intro.html
15. Reimers, N., Gurevych, I.: Making monolingual sentence embeddings multilingual using knowledge distillation. In: Webber, B., Cohn, T., He, Y., Liu, Y. (eds.) Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 4512–4525 (2020)
16. Reshetnikov, A., Marinescu, M.C., More Lopez, J., Mendoza, S., Freire, N., Marrero, M., Tsoupra, E., Isaac, A.: Deart: Building and evaluating a dataset for object detection and pose classification for european art. Journal of Cultural Heritage **75**, 258–266 (2025). https://doi.org/https://doi.org/10.1016/j.culher.2025.07.022, https://www.sciencedirect.com/science/article/pii/S1296207425001542
17. Roald, M., Birkenes, M.B., Johnsen, L.G.B.: Visual navigation of digital libraries: Retrieval and classification of images in the national library of norway's digitised book collection. arXiv preprint arXiv:2410.14969 (2024)
18. Stony Brook University Libraries: Search AI, https://library.stonybrook.edu/about-search-ai/