https://doi.org/10.2352/issn.2168-3204.2025.22.1.22

Enhancing Archival Transparency with AI: A RAG-Based Case Study on Cinematic Heritage

Enea Ahmedhodzic, Andrea Mario Trentini; Dipartimento di Informatica "Giovanni degli Antoni", Università degli Studi di Milano, Milan, Italy

Abstract

Cinematic archives preserve an invaluable heritage, yet accessing their content is often challenging due to data fragmentation, inconsistent standards, and the absence of user-friendly tools. Even when materials are available, consultation may require archival staff or specialized knowledge. This paper introduces Valter, a prototype AI chatbot developed as a case study for the Film Center Sarajevo (FCS), which explores how retrieval-augmented generation (RAG) can support transparency and accessibility in under-resourced archival settings. The system uses semantic search over multilingual embeddings to retrieve relevant information and generate answers in natural language. While still under development, Valter demonstrates the potential to enhance resource discovery and metadata validation, offering a replicable approach that could inform future digital access strategies across similar institutions.

Introduction

Cinematic archives are essential for safeguarding audiovisual memory, yet their consultation is often hindered by technological, organizational and data-related constraints. This situation is especially evident at the **Film Center Sarajevo (FCS)**, which inherited a large and heterogeneous collection under challenging historical and material conditions. The 1992–1996 Siege of Sarajevo caused severe damage to Bosnia-Herzegovina's cultural heritage, including its film archive materials. In the post-war period, preservation became increasingly difficult, as neglect and limited resources led to the progressive deterioration of materials (e.g., fungal growth, shrinkage, acetate decay, and other forms of film degradation) [1].

Despite the existence of a basic internal structure, information is still fragmented across various storage formats and physical supports, ranging from paper registers to external hard drives and LTO tapes, and dispersed in disconnected files. In the absence of a unified archival system, access and verification continue to rely on slow, manual procedures. Metadata quality also varies significantly: some films are fully documented in printed catalogues, while others carry only a handwritten title on the reel box (see Figure 1).

These gaps hinder external collaboration. During the restoration of *Slike iz života udarnika* (1972), for example, missing metadata made it hard to locate suitable film elements; success was possible only through cooperation with several institutions abroad. Situations like this underline the need for interoperable data infrastructures and long-term digital-preservation strategies.

Although Bosnia and Herzegovina has adopted legal instruments such as the "*Zakon o slobodi pristupa informacijama"* [2] (comparable to the Freedom of Information Act), the country still faces major limitations in terms of digital openness and public data infrastructure. According to the 2024 Open Data Maturity Report [3] by the European Data Portal, Bosnia and Herzegovina ranks last among the surveyed countries, with an overall score of just 15%. By comparison, the next-lowest country, Albania, scores 47%, while France leads with a full 100%. Particularly concerning is Bosnia and Herzegovina's zero score for DCAT-AP implementation, indicating a complete lack of compliance with semantic metadata standards. Access to cultural heritage data is still mediated by bureaucratic or analogue processes, and machine-readable formats remain rare.

Most of the Film Centre's documentation is published in PDF format, corresponding to the first (★) level of Tim Berners-Lee's Five-Star Open Data model [4]. While this ensures basic online availability, it does not support machine readability, structured reuse, or integration with other datasets. Transitioning to open formats such as CSV or RDF, and linking to external sources, would significantly enhance transparency and data reusability.

Achieving openness, however, depends on clean and structured metadata, something FCS cannot yet fully guarantee due to limited staff capacity and a lack of adequate tools.

To address these challenges, we developed Valter, an AI-chatbot that retrieves information from existing records and answers questions in natural language. A chatbot is a question-answering (QA) system that can respond in natural language, either by referring to a document collection or without explicit context [5]. Valter is built on a retrieval-augmented generation (RAG) pipeline: a multilingual embedding model finds relevant chunks; a local LLaMA-based language model crafts the response while citing sources. During interaction Valter also flags duplicate IDs and missing fields, helping staff improve data quality.

The aim of this paper is to present Valter's design and to evaluate how a lightweight, retrieval-augmented chatbot can improve transparency, accuracy and accessibility in an under-resourced cinematic archive.

The name honours *Valter brani Sarajevo* (1972), an iconic film in the collection and a symbol of resistance still celebrated by the Centre through a dedicated cinema and museum.

Several recent initiatives have explored the use of conversational AI to improve access to cultural heritage. **Europeana**[1]**,** the European platform for digital cultural heritage, describes its mission as sharing and promoting Europe's cultural content "to be used and enjoyed by everyone for learning, for work, or just for fun." One example developed in collaboration with Europeana is **CultureBot**[2] **,** a prototype chatbot designed to allow users to explore cultural collections through natural

[1] Europeana, https://www.europeana.eu

[2] CultureBot, https://culturebot.eu

language dialogue. Users can access the interface via the project website and pose questions directly to the chatbot. During our tests, we asked simple prompts such as "Who are you?", and the system responded with generic phrases like "Not much. What's new with you?". When asked about Europeana itself, the chatbot provided no response at all. The CultureBot website also references other cultural chatbot projects, but as of May 2025, two of the three links provided (e.g., for JHN-AMS and Cultura Italia) are no longer accessible. Although the initiative points to an interest in conversational tools for culture, our brief test returned limited responses, and we were unable to locate clear information on its practical use or current applications.

Another relevant case is **Deep Dive the Museum**[3], a plugin developed by Te Papa Tongarewa, the national museum of Aotearoa New Zealand. As described in the museum's official blog [6], this experimental tool connects the museum's collections to ChatGPT, enabling conversational access. When asked "Who are you?", the system replies: "I'm Deep Dive the Museum, your friendly guide to the rich collections of Te Papa Tongarewa." During testing, the plugin also demonstrated the ability to return direct links to specific items in the collection, offering a more integrated and user-responsive experience.

Methodology

Valter is currently based on a diverse collection of documents, including the official feature film catalog, an upcoming documentary catalog, archival film inventories, legal texts (such as the Statute, Code of Ethics, and the Decision on Establishing the Film Center), as well as descriptive material related to the recent restoration of *Slike iz života udarnika* (1972).

Data Collection and Cleaning

The collected documents were available in various formats (PDF files, Excel spreadsheets, and Word documents) and often contained overlapping, inconsistent, or incomplete information. To enable accurate and meaningful retrieval, a multi-step pipeline was implemented using a series of custom Python scripts designed to clean, align, and normalize the data.

The first phase focused on processing an internal inventory file that combined both titles from the official catalog and additional records identified during on-site inspections of physical film reels. This dataset included uncataloged materials, informal labels, and partial metadata, and therefore required extensive cleaning and verification. Irrelevant entries were manually removed, column headers were standardized, and all codes and titles were validated where possible.

This cleaning phase was implemented using the `pandas`[4] library for data handling and the `fuzzywuzzy`[5] package for string similarity computation. Fuzzy string matching techniques (Levenshtein distance ≥ 90%) were applied to identify potential duplicates. The script generated detailed CSV reports of near-duplicate titles and missing or duplicated codes, supporting the early detection of anomalies such as mislabelled or uncatalogued records prior to integration with the official catalog.

Once cleaned, the data was reconciled with the official FCS catalog by matching entries based on codes, titles, and type labels. This comparison was carried out through a dedicated script that performed record-level alignment across separate catalog sheets (feature films and documentaries). When matches were found, film names were compared using fuzzy logic, and type labels were normalized to a common format (e.g., "D" for documentary, "M" for movie). Results were classified as exact match, similar name, or different name depending on the similarity score, and discrepancies in type were also flagged. The script produced a comprehensive report and a secondary filtered file listing only problematic entries requiring manual review. As part of the data enrichment process, all documentary descriptions originally available only in Bosnian were translated into English to enable multilingual indexing and retrieval.

YouTube trailer links were also automatically integrated into the catalog. Used script employed the `yt_dlp`[6] library to extract video metadata from public playlists and used the same fuzzy matching logic to perform approximate string matching between video and film titles. Title normalization via regular expressions and a set of synonym mappings further improved match accuracy. Trailers with a similarity score of 85% or higher were automatically assigned to the corresponding catalog entries.

This rigorous cleaning and integration process resulted in a unified dataset, establishing a robust foundation for semantic indexing and interactive exploration.

A different workflow was adopted for the extraction of text from PDF and Word documents, particularly for legal and descriptive materials. These files were intended for semantic indexing and required conversion into plain text. To ensure high textual quality, PDF files were manually converted into .docx format to allow visual inspection and correction of OCR errors, broken formatting, or inconsistent numbering.

Significant sections (e.g., articles, paragraphs) were manually annotated. A script based on the python-docx library was then used to convert each .docx file into a .txt file while preserving paragraph breaks and document structure. Metadata such as source document, article numbers, and section markers were also extracted, preparing the material for chunking and embedding.

Embedding and Vector Database Construction

To enable semantic retrieval across our sources, we constructed three persistent vector databases using a unified pipeline: one for film collection, one for general textual content, and one for legal documents. Each collection was processed through the same stages as data cleaning, chunking, embedding, and storing metadata-rich representations into ChromaDB [7], an AI-native open-source vector database designed for large-scale semantic search.

The embedding model `intfloat/multilingual-e5-base` [8] was downloaded and executed locally via the `Sentence-Transformers`[7] library. This transformer-based encoder, initialized from `xlm-roberta-base` [9], includes 12 layers, produces 768-dimensional embeddings, and processes inputs up to 512 tokens. Trained on a large corpus of contrastive

[3] Deep Dive the Museum, https://chatgpt.com/g/g-FkPJb2QGR-deep-dive-the-museum.

[4] `Pandas`: https://pypi.org/project/pandas/

[5] `Fuzzywuzzy`: https://pypi.org/project/fuzzywuzzy/

[6] `yt_dlp`: https://pypi.org/project/yt-dlp/

[7] `Sentence-Transformers`: https://sbert.net/

text pairs in multiple languages, including English and Bosnian, the model supports sentence-level semantic comparison and cross-lingual alignment.

This setup enabled fully offline and reproducible embedding generation. The system was developed and tested locally to allow rapid iteration and controlled evaluation.

The vector database contains three distinct collections, each built from a different source: a structured CSV catalog for films, plain text files for general documents, and normative texts for legal references. For the film collection, a synthetic semantic field was created by concatenating titles, descriptions, and key metadata. General documents were segmented into overlapping token-based chunks, using a chunk size of approximately 350 tokens and an overlap of 50 tokens. This approach follows best practices adopted in retrieval-augmented generation (RAG) models, which rely on token-level segmentation to improve dense indexing and retrieval performance [10]. Legal texts were first segmented into logical units such as "Article 1" or "Principle 1" using regular expressions, then processed using the same chunking strategy, preserving the internal coherence of each provision.

All embeddings were normalized and compared using cosine similarity, which was explicitly set as the distance metric in ChromaDB. This measure is commonly used for semantic similarity tasks and reflects the angular proximity of vectors in high-dimensional space [11]. Each vector collection was also exported as a CSV summary to ensure transparency, reproducibility, and integration into downstream applications. This infrastructure forms the foundation for the retrieval-augmented generation workflows developed for the Film Center Sarajevo.

Semantic Retrieval

The system relies exclusively on semantic similarity, measured via cosine distance between embedding vectors. Similarity thresholds are set at 70% for films and 75% for general and legal texts. The slightly lower threshold used for the film archive reflects the greater variability often found in film-related information. Titles, roles, and formats can differ a lot across languages and cataloging practices, so a more flexible threshold helps avoid missing relevant results. On the other hand, general and legal texts are usually more structured and consistent, which allows for a higher threshold to reduce noise and keep results more precise.

For the film archive, in addition to semantic retrieval, the system also applies structured exact matching based on normalized metadata equivalences to capture direct correspondences. These matches are processed and logged alongside semantic results, and deduplication is performed to avoid redundant outputs, especially in cases where multiple entries refer to the same film.

The system returns up to 10 results for films and up to 5 for general and legal documents, depending on their similarity to the input query. Retrieval for general and legal texts operates at the chunk level, with each chunk enriched by metadata such as document source, position, and similarity score. All queries are logged in a CSV file, capturing match type, similarity score, and retrieved content to support evaluation and reproducibility.

Retrieval-Augmented Generation and LLM Integration

The system adopts a Retrieval-Augmented Generation (RAG) architecture [12] to enhance user interaction by combining semantic retrieval with large language models (LLMs). When a user submits a query, it is first converted into an embedding and compared against a vector database of pre-processed archival content. The most semantically relevant documents are retrieved and passed to the LLM, which generates a natural language response grounded in the retrieved content rather than relying solely on pre-trained knowledge.

Thanks to the multilingual capabilities of the embedding model, queries written in Bosnian can still yield accurate and contextually relevant results during the retrieval phase. However, due to the current limitations of the selected LLM, which does not support Bosnian, the generated responses are provided in English. While this restricts full multilingual support at the response level, the system remains functional and consistent in bilingual contexts where users are familiar with both languages.

Unlike traditional keyword-based search, semantic retrieval captures the meaning and intent of a query, improving the relevance of returned results even when users use different expressions from the source text [13]. This allows the system to respond effectively to complex and context-sensitive questions, especially in domains like archival access, where users may not be familiar with controlled vocabularies or internal metadata structures.

To facilitate natural language interaction with the archival system, we integrated a locally hosted large language model (LLM) using the Ollama[8] framework. Ollama provides a lightweight infrastructure for running quantized LLMs on local machines, offering efficient model management and inference capabilities without requiring specialized hardware. For this prototype, we selected the instruction-tuned variant of Meta's LLaMA 3[9] model, specifically the 8B parameter version, referred to as "`llama3:instruct`[10]". This model was chosen due to its relatively small size (approximately 4.7 GB), which allows for deployment on standard hardware without the need for specialized computational resources. Despite its compactness, the model has shown promising performance during initial tests.

The backend architecture routes each user query to the most appropriate document collection, such as film records, general texts, or legal sources, based on keyword-level semantic analysis. Retrieved text chunks are passed to the LLM with a structured prompt that includes detailed instructions to avoid hallucinations, prevent assumptions, and ensure factual responses.

User Interface and Backend Workflow

User interaction is facilitated through a web interface built with Streamlit, which allows for real-time query submission and displays both the generated answer and the corresponding retrieved context. The interface presents responses in natural language alongside the archival excerpts most relevant to the query.

On the backend, each query is semantically routed to the most appropriate collection: film, legal, or general, by comparing its

[8] Ollama: https://ollama.com

[9] LLama: https://www.llama.com

[10] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

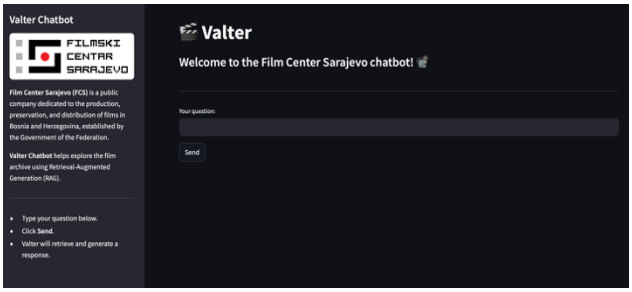

Figure 1 Streamlit-based interface of the Valter chatbot. Users can submit natural language queries and receive LLM-generated answers grounded in retrieved archival content.

embedding to pre-encoded reference examples. This classification mechanism leverages the multilingual embedding model to ensure consistency across languages.

Once routed, the system performs semantic retrieval from the relevant vector database, applying exact matching on structured metadata fields (such as catalog codes) when dealing with film queries. Retrieved results are combined with descriptive metadata and formatted into coherent textual chunks. If available, YouTube trailer links are automatically extracted and included.

The retrieved content is then used as the basis for answer generation. A structured prompt is dynamically composed, incorporating the selected chunks along with detailed instructions to reduce hallucinations and ensure grounded, non-speculative responses. If no relevant content is retrieved, fallback messages are automatically generated.

All user queries and corresponding retrieval results, including similarity scores, match types, and associated metadata, are logged in a structured CSV file. This logging supports reproducibility, facilitates evaluation, and helps identify potential cataloging inconsistencies through systematic review

Results

To explore the semantic structure of the film collection stored in a vector database, Principal Component Analysis (PCA) was applied to the cosine-normalized embeddings of the titles. The resulting two-dimensional projection (Figure 1) revealed four main clusters: Official FCS Catalogue, Archive Films, Film Reels with Unknown or Missing Titles, and Other Films. K-Means clustering was used to identify these clusters, offering an

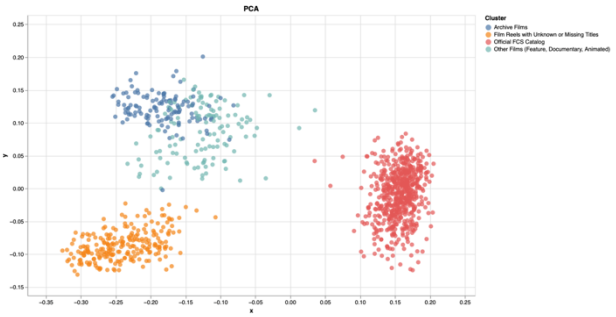

Figure 2 PCA of film embeddings

interpretable representation of the collection's organization within the embedding space.

The PCA projection provided an approximate visual mapping of the dataset and highlighted its internal structure. The visualization also revealed specific anomalies, such as points associated with titles from the film collection that appeared in the plot but lacked any meaningful semantic content. This observation prompted a manual inspection of the original data sources, which revealed that some titles were duplicated across different unique identifiers. As a result, additional functions for automatic cleaning and deduplication were integrated into the dataset processing pipeline.

Structurally, the map shows a clearly defined cluster on the right-hand side, representing the official Film Center Sarajevo catalogue. These records are characterized by rich and complete metadata, including director, synopsis, duration, and year of production. On the left, three separate clusters can be observed, corresponding to archival fragments, feature film or documentary, and reels with missing or unknown titles. These items were recovered during the cataloging process, but lack sufficient information for accurate categorization [1].

Beyond its technical utility, this type of visual analysis also serves as a strategic communication tool. It illustrates the need for a systematic reconstruction of archival holdings. Over time, this representation could serve as a transparency and progress-tracking mechanism: as missing metadata are recovered or added, records currently located in the left-hand clusters are expected to gradually shift toward the well-structured cluster on the right, visually reflecting the enrichment of the collection.

To evaluate the effectiveness of the language model compared to semantic retrieval, we conducted a two-phase

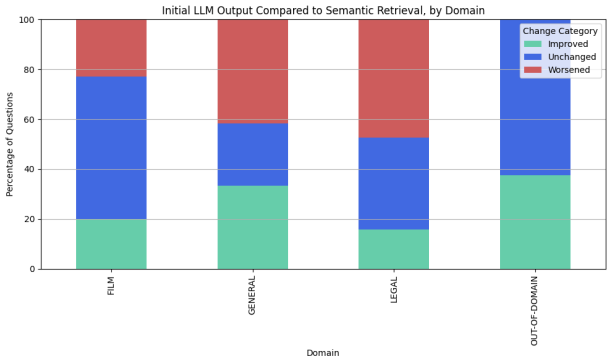

Figure 3 Initial LLM Output vs. retrieval performance by domain

assessment using a fixed set of 100 queries. In each phase, the same questions were posed to both systems, and the LLM's output was evaluated in relation to the retrieved content corresponding to the source collection within the vector database. Each query was assigned to one of four domains:

- FILM (film catalog): structured metadata descriptions from the film archive.
- LEGAL (normative texts): articles and provisions from laws and regulations.
- GENERAL (text documents): documents from institutional or archival sources.
- OUT-OF-DOMAIN: queries not directly aligned with any of the predefined collections

Each response was first evaluated independently and categorized as good, acceptable, or poor, based on its overall quality, clarity, and relevance to the query. After this qualitative assessment, we compared the LLM outputs to the corresponding retrieved content and assigned one of three labels: Improved, Unchanged, or Worsened, to indicate whether the LLM provided added value, maintained the same level of informativeness, or introduced errors relative to the retrieved material.

Figure 3 presents the results of the initial phase, where the LLM operated with minimal prompting. The comparison revealed

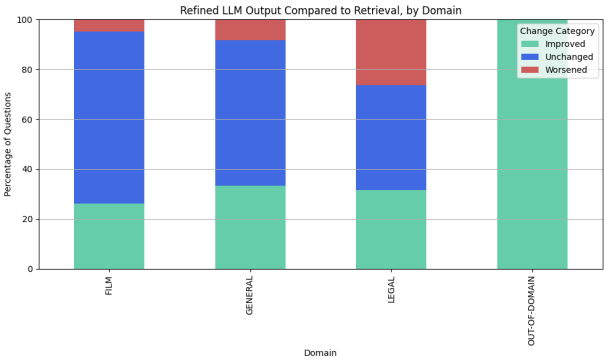

Figure 4 Refined LLM vs. retrieval performance by domain (after prompt optimization).

significant domain-dependent variability. In the LEGAL and GENERAL domains, the model frequently introduced factual drift of 47.4% and 41.7% of responses, respectively, were classified as worsened. The FILM domain showed greater stability, with 57.4% of responses unchanged and only 23.0% worsened. Notably, OUT-OF-DOMAIN queries, such as "What is the average height of a giraffe?", were handled with restraint: although the retrieval returned no content, the LLM provided polite refusals without hallucinating, leading to 100% of those responses being classified as improved.

In the second phase (Figure 2), the same questions were re-tested using more refined and guided prompts. The impact of prompt optimization is clearly visible: degradation rates decreased across all domains, most notably in FILM (from 23.0% to 4.9%) and GENERAL (from 41.7% to 8.3%). The improvement rate also rose in LEGAL (from 15.8% to 31.6%), indicating the model's increased ability to handle complex textual material when properly instructed.

Phase	Improved	Unchanged	Worsened
Initial prompting	22%	50%	28%
Refined prompting	34%	57%	9%

Table 1: Global comparison of LLM outputs vs. retrieval across both evaluation phases

To offer a more synthetic overview, Table 1 summarizes the global results across both phases, aggregating all domains. Although the share of improved responses rose from 22% to 34%, the most significant shift is the sharp reduction in degraded outputs, from 28% to just 9%. This suggests that the refined prompts contributed to greater consistency in the system's behavior. Since the retrieval remained unchanged across both phases, the improved results appear to reflect a better fit between the model's generation and the available content.

The OUT-OF-DOMAIN category offered an interesting case. These queries, such as "What is the average height of a giraffe?", did not correspond to any document in the database. As expected, the retrieval system returned no results. However, the LLM responded appropriately, politely indicating that such questions were outside the scope of available data. This was considered an improvement, as the model correctly handled unanswerable queries without generating misleading responses.

Importantly, this evaluation also revealed problems in the cataloguing information. For instance, one query revealed that two completely different films, both titled So (Salt), had been archived under the same catalog ID (0380). One is a short documentary[11] from 1967 (directed by Vefik Hadžismajlović), while the other is a feature film[12] from 1973 (directed by Gojko Šipovac). This error directly affected the retrieval results, and its discovery illustrates how side-by-side analysis of LLM outputs and archival sources can reveal issues that might otherwise be missed during routine cataloging. In this case, the misclassification of two distinct films under the same identifier could mislead users and compromise retrieval accuracy. Such findings suggest that deploying LLM-based systems, when used internally, not only as public-facing tools, can help staff identify data inconsistencies, minor cataloging oversights, or even structural gaps in incomplete collections. This expands the role of AI from passive retrieval to active support in curatorial validation and quality assurance within digital heritage infrastructures. However, such systems should be seen as complementary instruments, not as definitive solutions.

Discussion

Valter's deployment demonstrates the potential of retrieval-augmented generation (RAG) in the context of the Film Center Sarajevo archive, while also exposing persistent challenges related to metadata consistency, data fragmentation, and infrastructural limitations.

The system has shown that semantic retrieval and language models can support access to fragmented collections and assist in curatorial work. However, its performance remains closely tied to the structure and quality of the underlying data. In domains such as legal and institutional texts, where formatting is inconsistent or metadata is sparse, maintaining response accuracy remains challenging and requires further tuning.

A key outcome of this prototype phase has been the recognition of Valter's value as an internal tool. By comparing retrieved documents with generated responses, staff will be able to detect inconsistencies, duplicates, and gaps in cataloguing, issues not always visible during manual review. This highlights the system's dual role: not only as an access interface. It also as a diagnostic aid for archival validation and quality control. To fully realize this potential, several technical and design limitations must be addressed. The current version lacks conversational memory, which limits multi-turn interaction and reduces contextual continuity. Moreover, there is currently no built-in feedback mechanism to capture user corrections or flag errors, restricting the system's capacity for iterative improvement. These aspects are particularly relevant as Valter is introduced to internal staff, whose feedback could help refine both metadata quality and system behavior.

Another open challenge concerns the dynamic nature of archival content. As new materials are digitized or reclassified, the system will need to support incremental updates to the vector database and embedding layers. Ensuring that the retrieved

[11] So, short documentary: https://www.imdb.com/title/tt0876547/

[12] So, feature film: https://www.imdb.com/title/tt0181826/

content remains aligned with the most current state of the archive will be essential for long-term sustainability.

Once the data cleaning process is complete, the resulting catalog will be prepared in open formats to support future integration with Linked Open Data (LOD) initiatives, promoting transparency and interoperability. In parallel, the chatbot itself could be further developed and made available to the public, provided that the underlying data is accurate and the system's outputs are deemed reliable. At that stage, more powerful embedding models and LLMs may also be considered, depending on the available resources and institutional priorities.

Valter represents a promising compromise for advancing institutional transparency. By offering a simple and intuitive interface, it lowers the barrier for users to explore archival content, regardless of their technical background. However, this conversational access must be accompanied by the publication of open and machine-readable data, ensuring that the information made available through the chatbot is also fully accessible, verifiable, and reusable by citizens and digital systems alike.

Conclusion

This paper presented Valter, an AI-powered system designed to support both access and internal validation of cinematic archival content in contexts marked by structural and technological fragility. Using the Film Center Sarajevo as a case study, the project explored how retrieval-augmented generation (RAG) can improve transparency, accessibility, and curatorial workflows in under-resourced heritage institutions.

The evaluation conducted across 100 queries showed that prompt refinement can significantly improve output quality, with degraded responses dropping from 28% to 9% between phases. In addition to enhancing access, the system revealed cataloging anomalies and duplicated identifiers, demonstrating its utility as a support tool for archival consistency and data quality. These results validate Valter's dual role as both a user-facing interface and an internal diagnostic instrument.

Future work will focus on three key directions. First, we plan to integrate conversational memory and structured feedback mechanisms to support multi-turn interaction and adaptive system behavior. Second, dynamic update pipelines will be developed to ensure that newly processed materials are continuously integrated into the vector database. Third, once the data cleaning phase is complete, the catalog will be published in open and machine-readable formats, facilitating reuse and alignment with Linked Open Data principles.

In addition, we intend to release the full pipeline and chatbot interface on GitHub, enabling other small institutions and archives to replicate the model or adapt it to their own collections. By documenting the process in an accessible and transparent way, the project aims to contribute to the broader ecosystem of open-source tools for cultural heritage.

The long-term vision is to make Valter publicly available, expanding access to cinematic heritage while ensuring that responses remain grounded in curated, transparent sources. In parallel, more powerful embedding models and LLMs may be adopted to further improve response accuracy. Ultimately, Valter represents a scalable and transferable approach to modernizing archival access, combining semantic technologies with open data practices to support a more inclusive and accountable cultural infrastructure.

References

[1] E. Ahmedhodzic and E. Muftarevic, 'Preserving Bosnia's Cinematic Heritage: Challenges and Achievements of the Film Centre Sarajevo in the Post-War Era and Beyond', RESEARCH CULTURE AND SCIENCE BOOKS, vol. 13, pp. 26–33, 2025.

[2] Zakon o slobodi pristupa informacijama na nivou institucija Bosne i Hercegovine, vol. 61/23. 2023. [Online]. Available: http://www.mpr.gov.ba/

[3] M. Page, A. Behrooz, and M. Moro, '2024 Open Data Maturity Report'. [Online]. Available: https://data.europa.eu

[4] T. Berners-Lee, 'Linked Data - Design Issues', W3C. Accessed: May 15, 2025. [Online]. Available: https://www.w3.org/DesignIssues/LinkedData.html

[5] M. S. Salim, S. I. Hossain, T. Jalal, D. K. Bose, and M. J. I. Basher, 'LLM based QA chatbot builder: A generative AI-based chatbot builder for question answering', SoftwareX, vol. 29, p. 102029, 2025.

[6] G. Watkins, 'Experiment to plug Te Papa collections into ChatGPT', Te Papa's Blog. Accessed: May 10, 2025. [Online]. Available: https://blog.tepapa.govt.nz/2024/08/22/experiment-to-plug-te-papa-collections-into-chatgpt/

[7] C. Community, 'Chroma: The AI-native open-source embedding database'. 2024. [Online]. Available: https://www.trychroma.com

[8] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, and F. Wei, 'Multilingual E5 Text Embeddings: A Technical Report', Feb. 08, 2024, arXiv: arXiv:2402.05672. doi: 10.48550/arXiv.2402.05672.

[9] A. Conneau et al., 'Unsupervised Cross-lingual Representation Learning at Scale', Apr. 08, 2020, arXiv: arXiv:1911.02116. doi: 10.48550/arXiv.1911.02116.

[10] P. Lewis et al., 'Retrieval-augmented generation for knowledge-intensive nlp tasks', Advances in neural information processing systems, vol. 33, pp. 9459–9474, 2020.

[11] N. Reimers and I. Gurevych, 'Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks', Aug. 27, 2019, arXiv: arXiv:1908.10084. doi: 10.48550/arXiv.1908.10084.

[12] K. Olawore, M. McTear, and Y. Bi, 'Development and Evaluation of a University Chatbot Using Deep Learning: A RAG-Based Approach'

[13] S. Pokhrel, S. Ganesan, T. Akther, and L. Karunarathne, 'Building Customized Chatbots for Document Summarization and Question Answering using Large Language Models using a Framework with OpenAI, Lang chain, and Streamlit', Journal of Information Technology and Digital World, vol. 6, no. 1, pp. 70–86, 2024.

Author Biography

Enea Ahmedhodzic is a PhD student in Computer Science, specializing in IT systems for transparency in public administration. She holds a Master's degree in Conservation Science from the University of Milan and has collaborated with cultural institutions to integrate emerging technologies in the preservation and promotion of film heritage. Her research focuses on data curation, open knowledge, and enhancing public access to cultural collections.

Andrea Trentini is an assistant professor at the Università degli Studi di Milano, teaching Programming 101, Embedded Systems, and Digital Citizenship. He co-founded I&T Mentor (2002-2006),

specializing in object-oriented software consultancy. He has taught professional courses on Unix/Linux, Java, and UML. A Computer Science graduate (1996), he also attended the CEFRIEL Master in IT. He supports Free Software, serves on the Maison de l'Italie board, and is passionate about reading, traveling, and various activities.