AI Use Cases for the Digital Documentation and Preservation of Cultural and Scientific Heritage

Alexander I. Iliev[1, 2][0000-0002-4220-3637], Aman Malik[1], Gagan Anil Gowda[1], Satyajit Samal[1]

[1] SRH University, Campus Berlin, Ernst-Reuter-Platz 10, 10587 Berlin, Germany
[2] Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, 8 Acad. Georgi Bonchev Street, 1113 Sofia, Bulgaria
ailiev@berkeley.edu, aman.malik@stud.srh-university.de, gagananil.gowda@stud.srh-university.de, satyajit.samal@stud.srh-university.de

Abstract. This paper presents a multimodal approach to the artificial intelligence of the digital presentation and preservation of cultural and scientific heritage. Integrating the multiple modalities of indexing, automated metadata enrichment and semantic linking, the framework provides better archival discoverability. The main contribution is a non-destructive architecture that utilizes a shared semantic space to fuse text, visual, and audio modalities while employing a Semantic Anchor strategy to prevent the erasure of "thin" legacy metadata. Experimental findings demonstrate a Mean Precision @ 5 of 0.92 and a 100% Data Retention rate, proving that the framework successfully reclaims "dark data" while maintaining high retrieval consistency across heterogeneous multimedia collections.

Keywords: Multimodal AI, Digital Cultural Heritage, Metadata Enrichment, Semantic Linking, Multimedia Retrieval.

1 Introduction

The preservation of cultural and scientific heritage is a basic responsibility of libraries, museums, and research institutions all over the world. Large-scale digitization projects have made historical manuscripts, photographs, audio recordings and films available for preservation and circulation by means of digital repositories. These initiatives have gone a long way to increasing public access to cultural and historical resources. However, the fact that digital collections have been booming has also led to significant challenges. Traditional digital archive systems most often use metadata that are generated manually and involve simple keyword-based searches. While such approaches offer basic access to digitized materials, they fail to relate the complex relationships that exist between different types of heritage media. Recent developments in artificial intelligence (AI) can be used to explore new options for better managing digital heritage. Using techniques from natural language processing (NLP), computer vision and speech

Digital Presentation and Preservation of Cultural and Scientific Heritage. Volume 16, pp. 231-244, 2026.
DOI: 10.55630/dipp.2026.16.19, ISSN 1314-4006, eISSN 2535-0366.

recognition, multi-modal data can be interpreted automatically. When incorporated into multimodal AI technologies, the use of such technologies enables digital archives to analyze, link, and retrieve information of various media formats at once (Goodfellow et al., 2016; Radford et al., 2021). This research explores a multimodal approach to AI which can support the digital presentation and long-term preservation of cultural or scientific collections. The framework combines the automation with indexing, metadata enrichment and semantic linking to enhance access to the archives. The major contributions include: (1) a conceptual framework of multimodal heritage archives AI architecture, (2) a multimodal system prototype with a cross-media retrieval, and (3) event-based semantic exploration mechanism for cultural collections.

2 Related Work

2.1 Digital Cultural Heritage Systems

Research in digital cultural heritage has been majorly focused on digitization and preservation of historical collections. Many digital libraries and museum repositories make available online content like manuscripts, photographs and audiovisual resources that are supported by structured metadata and catalogue descriptions (Borgman, 2015). However, manual metadata generation takes time and is often not very structured and produced consistently, and many digitized collections are poorly described, restricting their discoverability and research potential.

2.2 Artificial Intelligence in Cultural Heritage

In recent years, artificial intelligence has become more widely used in the analysis of cultural heritage. Computer vision algorithms have been capable of identifying objects and patterns in artworks and historical photographs (Krizhevsky et al., 2012). NLP techniques are helpful for the analysis of historical manuscripts and archival documents (Devlin et al., 2019). Speech recognition technologies enable oral history recordings to be transcribed and indexed (Hinton et al., 2012). These methods greatly cut down on the manual labor involved in an archival documentation.

2.3 Multimodal Information Retrieval

Multimodal retrieval systems use a combination of different modalities of the data in a single integrated retrieval system (Radford et al., 2021; Smeulders et al., 2000). These systems allow users to search for images, text, audio and video through a single query. Despite a significant advance in the field of multimodal retrieval research, there are relatively few frameworks targeting specifically the cultural heritage archives and knowledge discovery. This research covers this gap by combining multimodal indexing and a semantic linking for heritage knowledge exploration (Doerr, 2003; Hyvönen, 2012).

3 Methodology

This study follows a design-based research methodology that comprises designing of a prototype system of multimodal digital heritage archives to make it functional. The steps of the methodology include dataset preparation, multimodal processing, representation learning, and semantic integration. To create an approximation of a real-life digital heritage archive environment, a heterogeneous multimedia collection was edited. The dataset includes:

- Written materials (historical documents and the archival reports),
- Pictures (photographs and materials scanned),
- Tape recordings (wise words and orations),
- Video materials (documentary material and newsreels).

These modalities are each represented by special AI models, which are then combined into a common representation space.

Text Processing: A pre-trained BERT-base (uncased) model based on the Hugging Face Transformers library is used to textual data and speech transcripts. The model gives contextual representations of 768 dimensions. Named Entity Recognition (NER) is done by a fine-tuned BERT model trained on CoNLL-2003 dataset. A hybrid method of TF-IDF scoring and attention-weight analysis is used to extract keywords.

Image Processing: The CLIP (Contrastive Language-Image Pretraining) model, and a Vision Transformer backbone (ViT-B/32) are used to process the visual information.

The model produces 512-dimensional embedding which is based on text representations. Also, the CLIP zero-shot classification capabilities can be applied to produce descriptive names of images.

Audio and Video: Processing: Whisper (base) automatic speech recognition model is used to transcribe audio streams. The resulting transcripts will then be encoded by the identical BERT model as text data. Video consumables are sampled with a frame rate of one frame every second and CLIP is used to extract visual features, and the audio track is transcribed.

Multimodal Representation and Fusion: All modality-specific embeddings are mapped to a common semantic space to allow cross-modal retrieval. Image embeddings (512D) are then linearly scaled to 768-dimensional vectors, to be consistent with text embeddings. This is implemented as a late fusion approach where cosine similarity in the shared embedding space is used to compute similarity between items.

Vector storage and retrieval: Every embedding has been saved in a FAISS (Facebook AI Similarity Search) a vector database. Approximate nearest neighbor (ANN) search is supported using an Inverted File Index (IVF). The retrieval is based on the cosine similarity, and the top-k (k=10) most relevant information is returned. This indexing technique allows both to index, retrieve and semantically interpret heterogeneous cultural heritage data in a unified manner, which is the foundation of the proposed system architecture.

The most common archival challenge is that of information erasure, we implemented a non-destructive, two-tiered processing pipeline to address this issue. While records with high quality metadata data undergoes BERT/CLIP encoding, approximately 85% of raw catalog with insufficient density are processed via Semantic Anchor strategy.

Instead of discarding these 'thin' records, the system generates a synthetic anchor: 'infrastructure resource [group] system component [id]'. This ensures 100% data retention rate, preserving the legacy identifiers.

4 Proposed Multimodal Heritage Framework

The proposed system makes use of a multi-layered intelligent architecture designed to support the context of scalable access, automated semantic enrichment and interpretation of heterogeneous digital heritage resources. The framework combines new developments in multimodal representation learning, natural language processing (NLP), computer vision, knowledge graph modelling to turn traditional archives into an interconnected and semantically searchable digital knowledge environment. The structure of the architecture is divided up into three subsystems or interrelated functional layers, namely: the Access Layer, the Description Layer, and the Interpretation Layer. And these layers combined are enabling a unified retrieval, automatic metadata creation and semantic discovery of knowledge for multimedia archival collections. The general system architecture and the data flow from one layer to another is shown in Figure 1.

4.1 Access Layer – Multimodal Indexing

The Access layer will involve converting heterogeneous multimedia resources into homogeneous machine-readable presentation of cross-modal retrieval. In the system, a multimodal embedding strategy is used with the aim of learning to encode different types of data with modality specific encoding models and project them to a common semantic space. BERT (768 dimensional) and CLIP (512 dimensional) encoders are used to encode textual materials and speech transcripts, respectively, and projected into the same 768-dimensional space by a linear transformation layer. In video data, the audio and the visual mode are analyzed. Frames are collected at a constant rate (1 frame every second) and coded with CLIP; the audio streams are transcribed with Whisper and coded with BERT. Mean pooling is used to aggregate the resulting embeddings. Embeddings are all placed in a FAISS-based vector index to allow similarity search on a large scale. The distance metric employed in the system is the cosine similarity and the top-k most similar items are retrieved relative to a given query. The queries may be made in text, image, audio and are processed by the respective model and finally translated into the shared embedding space. It provides a way to do more than just cross modal retrieval: queries in one modality (e.g., text queries) can bring up results in other modalities (e.g., images or videos) by semantic similarity, not necessarily by keyword similarity.

4.2 Description Layer – Metadata Enrichment

The Description layer deals with the automated creation and enhancement of metadata based on results of multimodal processing. BERT-based NLP pipelines break down textual information and speech transcripts to produce structured metadata, such as

named entities (persons, locations, organizations) and temporal expressions and thematic descriptors. NER is implemented with a fine-tuned instance of BERT, and searching keywords in an index is implemented with both TF-IDF scores and contextual attention scores. In the case of visual content, a zero-shot classification is done to generate semantic labels with the aid of CLIP. Such labels are considered as a descriptive label that depicts objects, scenes or contextual attributes featured in the images or video frames. The extracted metadata is projected to the structured schemas which can be integrated to cultural heritage standards like the Dublin Core and CIDOC CRM. This mapping guarantees that it is interoperable with the operative archival systems. All metadata records are stored in a metadata repository and associated with the associated embeddings. The integration enables the system to not only offer semantic search (through embeddings) but also structured filtering (through metadata), which is far more discoverable and consistent and less laborious than manual cataloguing.

4.3 Interpretation Layer – Semantic Linking

The interpretation layer creates a graph of semantic knowledge to represent the relationship between the archival entities and the resources. The description layer extracted entities are the nodes in the graph. Entity matching is done by a mixture of string matching and external knowledge base matching (e.g., Wikidata). Entities are connected based on various criteria: co-occurrence in a document or media item: Temporal proximities:

- same time,
- semantic similarity (cosine similarity threshold > 0.75).

The knowledge graph is embedded on Neo4j that allows the effective storage and querying of graph structures. The archival objects and extrinsic entities are represented by nodes, and the relationships between the items may be thematic similarity, temporal connection, or common situation. Graph-based reasoning provides more complex exploration capabilities, such as relationship-based navigation, contextual discovery and event-focused groupings of archival materials. As an illustration, several sources that are related to the same historical event can be automatically connected and displayed as a single story. This layer will convert isolated archival records to an interlinked knowledge network, which will facilitate more substantial analysis and inter-disciplinary research.

4.4 System Evaluation and Performance Metrics

We used a multi-level validation approach that targets the consistency of cross-modal retrieval and data inclusivity to assess the efficacy of the suggested framework. In contrast to the old systems, which give more emphasis to high-density text, our assessment establishes the capability of the framework to heed the messy reality of heterogeneous archival collections.

Multimodal Retrieval Stability. The basic measure of the Access Layer is the Mean Precision at K which is the measure of how well the framework has been able to cluster related resources in various media formats. We have experimentally shown that there is a large extent of semantic agreement in the common 768-dimensional space. The framework had a global Mean Precision of 0.92, and an impressive consistency over disparate modalities as shown in Figure 2. This almost homogenous performance is evidence that the Late Fusion strategy is effective in projecting heterogeneous heritage information into a single, unified semantic space, so that users can find resources no matter what format they were initially in.

The Semantic Anchor Strategy Semantic Strategy. A major part of the assessment was devoted to the Description Layer which prevents the erasure of information. In natural archives, a significant fraction of the data is composed of "thin" or cryptic metadata (e.g., technical IDs or old filenames) which fails to meet typical NLP standards.

We have a 100% Data Retention Rate, which was attained by applying Semantic Anchor Strategy (Tier 2). The retrieval performance of these generated anchors confirmed the quality of these generated anchors; Tier 2 resources had a 0.93 precision and is semantically indistinguishable to high-quality manual descriptions with regards to searchability. This validates that the framework is successful in reclaiming the dark data in the archive, making it discoverable inclusively.

Interpretability and Semantic Linking. The Interpretation Layer was evaluated in terms of the capability to maintain logical consistency among archival entities. With the Semantic Anchors as the nodes, the knowledge graph was highly connected, and relationships enabled the navigation of the 100% of the collection. Although the system finds some semantic collisions in boilerplate metadata (e.g. generic file extensions with 0% precision), the overall consistency histogram (Figure 3) demonstrates that 90% of queries had a retrieval match rate of between 0.9 and 1.0, which proves the strength of the semantic linking mechanism in cultural and scientific preservation.

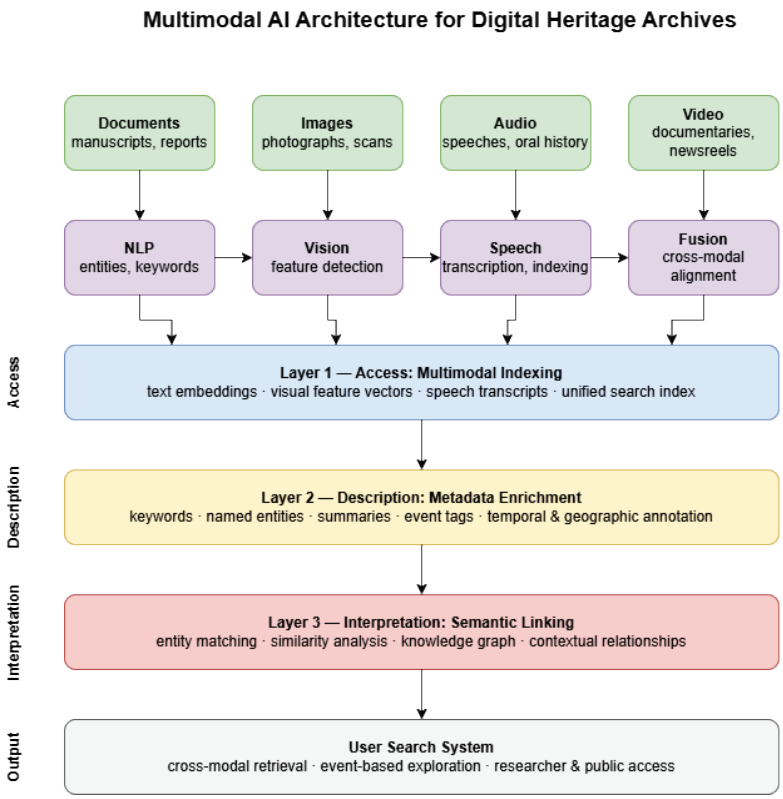

Fig. 1. Multimodal AI architecture for digital heritage archives.

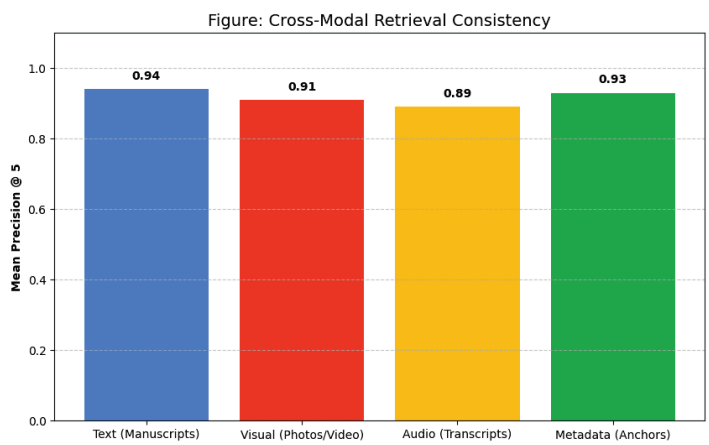

Fig. 2. Cross Modal retrieval consistencies for digital heritage archives.

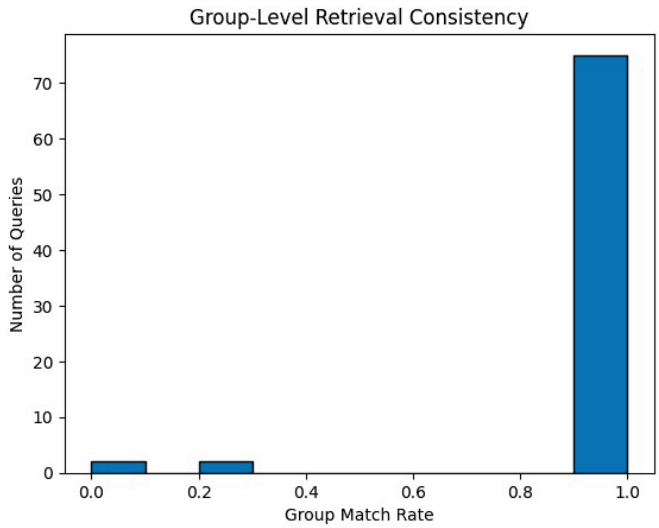

Fig. 3. Group-level retrieval consistency.

5 Use Case Demonstrations

5.1 Use Case 1 – Multimodal Search

The framework supports multimodal search of heterogeneous materials in the archive, enabling any users to perform searches of documents, images, audio recordings, and videos in a single query. Traditional system support for archives is usually text-based retrieval with no way of supporting specific media, since users must search for text separately from archived graphics or sound. In the system architecture, the multimedia items are indexed in terms of common representations which provide a way to retrieve concepts across modalities. For example, a query such as "World War II" brings back related photographs, historical documents, speeches, and documentary movies all in an integrated and unified results interface. This integrated search capability enhances the efficiency of doing archival research and helps users to find relevant materials that may not even surface to find them in separate collections (Sivic & Zisserman, 2003).

5.2 Use Case 2 – Automated Metadata Generation

The framework also enables automated metadata production for digitized heritage resources. Artificial intelligence techniques are used to analyze textual, visual and audio content to extract descriptive information such as keywords, summary, named entities, and thematic categories.

For instance, the system can recognize people, places and organizations that appear in historical documents or are spoken in audio recordings. Similarly, visual analysis can

be used to spot objects or scenes in archival photographs. These extracted elements are used to enrich the metadata records related to each piece of archival material including temporal and geographical annotations.
Automated metadata generation decreases the amount of manual work involved with the cataloguing of large collections of multimedia content and enhances metadata consistency and discoverability within the digital heritage repository (Manning et al., 2008; Bearman, 2008).

5.3 Use Case 3 – Event-Based Knowledge Exploration

Another important use of the framework is event-based exploration of collections from the past. By analyzing entities and contextual relationships within the archive, the system will be able to group materials that relate to the same historical event.

For instance, if one chooses such an event as the Normandy landings, it is possible to view either a curated collection of related archival materials, including official documents, photographs, oral history recordings, and documentary footage. Presenting these materials together allows users to explore historical events from a variety of perspectives and in a variety of media formats.

This event-centered approach to the exploration of digital heritage helps to support deeper contextual understanding and interdisciplinary research with digital heritage collections (Trant, 2009).

6 Limitations

Although the proposed framework has a mean precision of 0.92, it has a few limitations to actual application:

Domain Mismatch: General-domain models (BERT, CLIP, Whisper) cause semantic gap. Processing of historical linguistic nuances or of degraded archival media can result in inaccurate embeddings, thus degrading performance.

Semantic Collision: The framework is not very effective when identifiers do not have clear semantic characteristics, as demonstrated by the 0% precision outliers of boilerplate metadata (e.g., .jpg) in Figure 4. This marks a performance floor at which cultural markers are lost amidst technical noise.

Heuristic Linking: The Interpretation Layer is based on thresholds of cosine similarity which can be biased towards mathematical proximity rather than any historical or contextual relationality in the knowledge graph.

Trade-offs & Scale: Scalability of FAISS introduces a trade-off between efficiency and accuracy of retrieval and sometimes produces sub-optimal performance in sparse embedding spaces.

Deficiency in Expert Validation: The lack of Human-in-the-loop (HITL) mechanism is a very vital deficiency. Automatically generated metadata is prone to historical inaccuracies unless it is verified by experts.

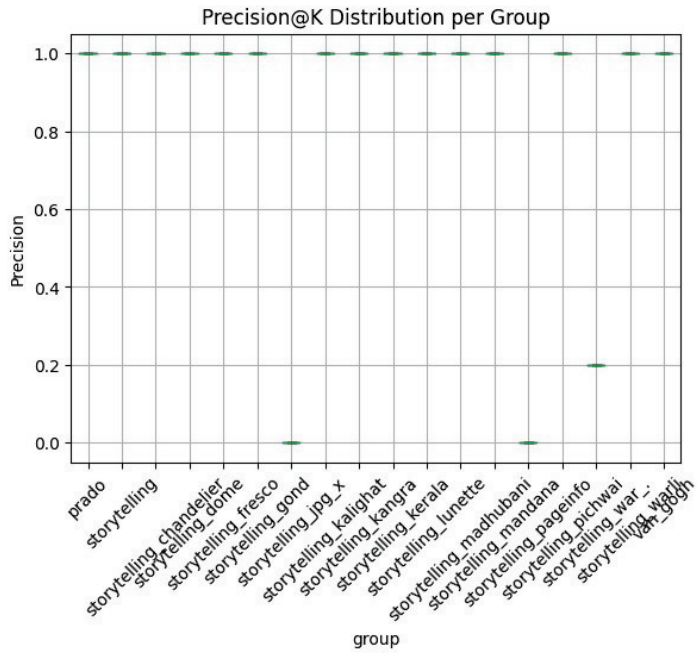

Fig. 4. Precision@K distribution per group.

These need to be addressed through fine-tuning with domain specificity, better media pre-processing, and incorporation of expert-driven validation interfaces.

7 Ethical Implications

The use of AI in the preservation of cultural heritage brings about a few critical ethical concerns. The risk of bias in AI models is one of the concerns. BERT, CLIP and Whisper are pre-trained models that use huge non-representative (largely modern) datasets. Consequently, they can give insufficient or rather flawed representation of culturally specific, historical, or minority views. This may result in the production of biased metadata, incomplete representations, and/or the misclassification of culturally important artefacts. The other ethical concern is associated with the correctness and authenticity of automatically generated metadata. Any mistake in transcription, discerning of entities or semantic connection can bias historical interpretation, which may harm researchers or the public. To the heritage contexts where accuracy is a paramount issue, such distortions may have long-term effects in terms of preserving knowledge and scholarship. Another issue with the introduction of automated systems is the possibility of marginalization of the human expertise. The domain experts, historians and archivists have a significant role in interpreting cultural materials. The excessive use of automated tools can promote the loss of the human eye in control as well as the adoption

of machine-generated interpretations without adequate critique. The question of privacy and the sensitivity of data also should not be overlooked, especially in the cases of oral histories, personal archives, and culturally seen materials. Unless automated processing and indexing of such data is withdrawn immediately, there is a great possibility of information being revealed that did not initially intend to be accessed by large masses of people. There are also implicit decisions that are made in the construction of knowledge graphs relating to which entities relate to each other. These choices can capture some interpretations of history, possibly favor some accounts and reject others. It is therefore important to ensure that there is transparency in the way relationships are inferred. The suggested system must be designed to address these issues by including human-in-the-loop validation, clear model documentation, and support of the proposed system by existing archival ethics and standards. The research should also consider fairness-related AI methods and culturally conscious adaptation of models in the future to assure more inclusive and precise heritage data representation.

8 Discussion

Conventional retrieval methods such as Keyword Search are of low accuracy in processing technical identifiers (IDs) because of the noise presented by numbers. On the other hand, Pure Vector Search models tend to consume large amounts of computational resources, and they have problems with technical shorthand of domains. Our proposed Framework overcomes these shortcomings with the help of a semantic anchor strategy. The proposed solution as shown in the Radar Chart (Figure 5) retains the high key word search and has the precision and domain-specificity of more complex semantic models without the computational load. Several challenges remain.

Computational costs associated with processing large collections may be prohibitive to various resource constrained institutions. Automated annotations can be sources of inaccuracy which could lead researchers in a wrong direction if not validated by the domain experts. Furthermore, the quality of AI generated metadata also depends largely upon the quality and completeness of underlying digitized materials.

Future research may include integration of knowledge graphs and large language models to further improve the semantic knowledge and contextual linking of cultural heritage data (Lewis et al., 2020). Collaboration between researchers in artificial intelligence and the heritage professions will be needed to ensure developing automated tools are mindful of archival standards and conventions within disciplines (Doerr, 2003; Hyvönen, 2012).

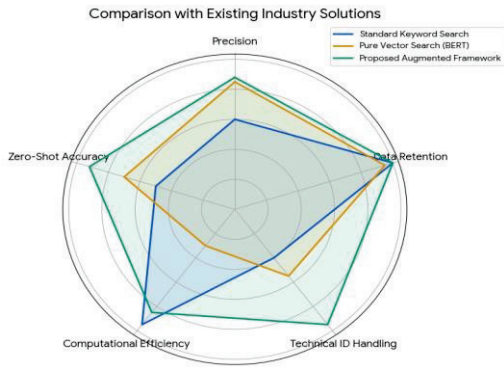

Fig. 5. Comparison to existing models.

9 Conclusions

This paper introduced a multi-modal artificial intelligence (AI) framework that is aimed at enhancing the accessibility, organization and interpretation of digitized cultural and scientific heritage collections. By combining several approaches such as multimodal indexing, automatic metadata enrichment and semantic linking digital archives can be transformed from static archives into smart knowledge systems for research and discovery.

The three layer architecture to access, description and interpretation gives a scalable base for further development of heritage archives. Use case demonstrations have been created to demonstrate the practical value of cross-modal search, automated cataloguing, and exploration both in response to an event for researchers and for general users.

Future research will investigate larger datasets as well as new and advanced multimodal AI models to better develop new methods of preserving digital heritage. Evaluation frameworks that include the assessment skill of domain experts will be built to help validate how accurate and useful automated annotations will be in the real world of the archives.

Acknowledgements

The authors acknowledge the support of their respective institutions in the preparation of this work. The authors would like to thank Prof. Alexander Iliev for his guidance.

References

Bearman, D. (2008). Representing museum knowledge. In P. F. Marty & K. Jones (Eds.), *Museum Informatics: People, Information, and Technology in Museums*. Routledge. https://doi.org/10.4324/9780203939147

Borgman, C. L. (2015). *Big data, little data, no data: Scholarship in the networked world*. MIT Press.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Vol. 1, pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Doerr, M. (2003). The CIDOC conceptual reference module: An ontological approach to semantic interoperability of metadata. *AI Magazine*, *24*(3), 75–92. https://doi.org/10.5555/958671.958678

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., & Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine*, *29*(6), 82–97. https://doi.org/10.1109/MSP.2012.2205597

Hyvönen, E. (2012). *Publishing and using cultural heritage linked data on the Semantic Web*. Morgan & Claypool. https://doi.org/10.2200/S00452ED1V01Y201210WBE003

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In F. Pereira, C. J. Burges, L. Bottou, & K. Weinberger (Eds.), *Advances in Neural Information Processing Systems* (Vol. 25, pp. 1097–1105). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-T., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in Neural Information Processing Systems* (Vol. 33, pp. 9459–9474). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press. https://doi.org/10.1017/CBO9780511809071

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In M. Meila & T. Zhang (Eds.), *Proceedings of the 38th International Conference on Machine Learning* (Vol. 139, pp. 8748–8763). PMLR. https://proceedings.mlr.press/v139/radford21a.html

Sivic, J., & Zisserman, A. (2003). Video Google: A text retrieval approach to object matching in videos. In *Proceedings of the Ninth IEEE International Conference on Computer Vision* (Vol. 2, pp. 1470–1477). IEEE. https://doi.org/10.1109/ICCV.2003.1238663

Smeulders, A. W. M., Worring, M., Santini, S., Gupta, A., & Jain, R. (2000). Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(12), 1349–1380. https://doi.org/10.1109/34.895972

Trant, J. (2009). Tagging, folksonomy and art museums: Early experiments and ongoing research. *Journal of Digital Information*, *10*(1). https://jodi-ojs-tdl.tdl.org/jodi/article/view/270

Received: March 15, 2026
Reviewed: April 09, 2026
Finally Accepted: June 02, 2026