https://doi.org/10.2352/issn.2168-3204.2025.22.1.21

AI-Driven Metadata Extraction and Semantic Search for Audiovisual Archives

André Rattinger, Giacomo Alliata, Kirell Benzi and Sarah Kenderdine; Laboratory of Experimental Museology – EPFL; Lausanne, Switzerland

Abstract

ArchiveVault is a next-generation digital archiving system designed to enhance access to audiovisual collections through automated metadata extraction and advanced retrieval mechanisms. Traditional archiving methods are labor-intensive, requiring extensive manual annotation that often leads to incomplete and inconsistent metadata. ArchiveVault addresses this challenge by employing AI-based transcription, named entity recognition (NER), speaker diarization, and pose detection to extract structured metadata from audiovisual archives [1]. This allows for rich, searchable metadata that improves retrieval precision beyond traditional keyword-based approaches.

By leveraging state-of-the-art AI techniques, ArchiveVault enables researchers, archivists, and content creators to perform semantic searches across large collections, discovering moments of interest more effectively. Our deployments in a national broadcast archive (RTS) and the Olympic Games media collection demonstrate how AI-driven processing unlocks previously inaccessible content, from spoken-word analysis to pose-based retrieval for sports footage [1, 4].

Introduction and Context

Over the last decade many institutions and media companies have digitised hundreds of thousands of hours of legacy film and tape. Heritage and media archives continue to expand rapidly, encompassing decades of digitized text, images, audio, and video content. Ensuring accessibility and effective retrieval in these vast collections presents a growing challenge. Traditional archival systems rely on manual cataloging and keyword-based media management, requiring significant time and expertise. A one-hour broadcast, for instance, can demand up to multiple hours of manual annotation, making large-scale indexing impractical [8]. Existing solutions often fail to adequately address the scalability, accuracy, and flexibility required by rapidly growing archives, highlighting the necessity for innovative approaches integrating automated metadata extraction with advanced retrieval techniques. The archives we are working with are well-annotated, but certain aspects are never feasible to annotate by hand. Examples of those are transcribing every word ever spoken in the archive or even describing the objects in every frame of every video.

In contrast, modern AI-driven solutions provide scalable and automated metadata extraction, enabling deeper and more structured content indexing [2, 3]. Progress in deep learning has made automatic speech recognition (ASR), language understanding, and computer vision accurate enough for production systems. By analyzing speech, text, and visual elements, AI models can produce detailed, machine-readable descriptors that facilitate more flexible and precise retrieval. Speech-to-text transcription converts spoken words into searchable text, while NER identifies people, places, and organizations, allowing for enriched entity-based retrieval. Additionally, pose detection in video content enables similarity searches across sports and performance archives, further expanding the capabilities of content discovery [1]. Transformer-based ASR (e.g. WhisperX) delivers <10 % word error rate on in-studio audio. Modern entity linking bridges textual mentions to graph IDs such as Wikidata, providing disambiguated, language-agnostic descriptors. Visual models can now recover human skeletal keypoints in unconstrained broadcast footage, unlocking body-pose queries invaluable for sports historians.

ArchiveVault harnesses these advances to automate metadata extraction at scale. The system produces a multi-modal index that supports natural-language, faceted, and similarity search in a single SQL corpus. Our contributions are threefold:

- **Scalable AI pipeline.** A local or Kubernetes-orchestrated set of workers processes audio, text, and vision tasks in parallel, writing structured outputs to a unified object store.
- **Hybrid retrieval engine.** PostgreSQL/pgvector fuses inverted indices with Approximate-Nearest-Neighbour (ANN) search, enabling Boolean filters over semantic embeddings at interactive latency.
- **Empirical validation.** We report the first large-scale evaluation of end-to-end AI indexing on 60 000 h of French-language television and millions of Olympic keyframes.

Core Components

ArchiveVault introduces a scalable AI-driven metadata extraction and search framework with the following key components: AI-Based metadata extraction: speech-to-text and speaker diarization, Named Entity Recognition and Linking to external knowledge sources and different feature extraction models depending on the needs of the archive. This is combined with scalable infrastructure for fast retrieval. The pipeline always goes through the following steps:

1. Indexing the source files:

Source files are added with the metadata to the database. They will be later used to reference back to with the metadata and subsequent files that have been generated from them. The source files will not be added to the system by default as the system needs to know what will be served as media files later on. This could be full videos, but also derived clips.

2. Indexing derivative audiovisual files:

The user specifies the ways they want to ingest the media. Either the full files or defining a way to split the video into multiple clips. A split can come in many different ways and can already be annotated beforehand. A split can also be automatically chosen by the system based on transcripts and speaker diarization.

This step adds the derivative files to the s3 storage, even the full source files if so wanted.

3. Metadata generation:

After the system has ingested the necessary media files, derivative features can be generated from the source files. This can be either based on the transcripts or the split (or left intact) media files at this point.

4. Serving the data:

After all the metadata files have been generated, the user can serve the archive now over an API to make it accessible.

Figure 1 shows the high-level overview of the processing steps as well as how the typical processing pipeline is used in practice. The following steps are at the core of the system and describe the options the user has for the generation process:

1. AI-Based Metadata Extraction

To ensure comprehensive metadata coverage for audiovisual archives, ArchiveVault employs a modular and expandable pipeline that integrates cutting-edge AI techniques for speech, text, and visual analysis. This pipeline enables the automatic extraction of structured descriptors that enhance retrieval, linking, and exploration capabilities. Below, we outline the key components of this system and their role in improving archival accessibility.

Speech-to-Text and Speaker Diarization:

ArchiveVault leverages WhisperX [10], a state-of-the-art ASR system that transcribes spoken content with high accuracy. With the help of the output of WhisperX we also perform speaker diarization, distinguishing different speakers in an audio stream. The diarization pipeline aligns speech segments with speaker identities using pre-trained neural embeddings, enabling structured segmentation of dialogues and discussions. Speaker diarization, cuts between the visual content and the extracted paragraphs of the text are used to split long videos into smaller easier to digest clips of content [1, 5].

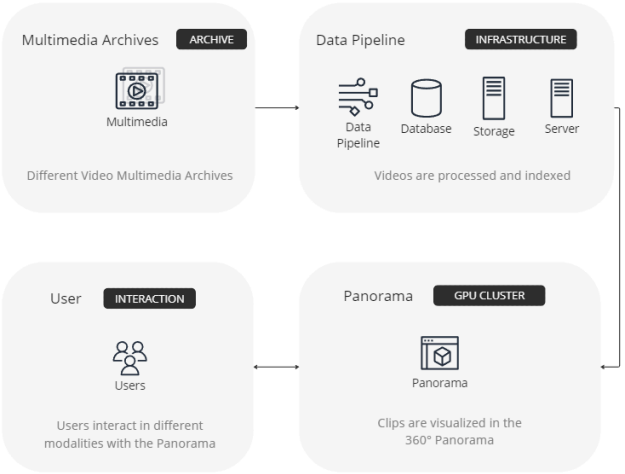

Figure 1. *Overview of the end-to-end steps in the pipeline processing. ArchiveVault starts at the archive level, indexes it and makes the information available for consumption by visualization systems or manual search to find pertinent video clips and content.*

Named Entity Recognition (NER) and External Linking:

ArchiveVault employs Natural Language Processing (NLP) through spaCy's Named Entity Recognition (NER) [7] to identify and categorize key information within the transcribed text of audiovisual content. These entities include names of people, geographical locations, and organizations. Currently, identified locations are automatically linked to corresponding entries in Wikidata, a large, open knowledge graph, which enriches the metadata with additional information and context. The system's design allows for the future integration of links to other types of entities, thereby further enhancing the depth and interconnectedness of the metadata. Additionally, to handle diverse data types or to clarify the meaning of specific terms, users have the option to upload custom thesauri, providing a mechanism for disambiguation and tailored entity recognition.

Pose-Based Search for Sports Archives:

ArchiveVault utilizes OpenPifPaf [9] to perform human pose estimation on video content, allowing the extraction of skeletal data that represents body positions. This capability enables content-based searches based on the similarity of these positions, which is particularly valuable for applications such as comparing athletic techniques across different performances or facilitating visual storytelling through the identification of similar actions. Models like OpenPifPaf analyze video frames to detect and track key points on the human body, generating a skeletal representation that can then be compared across different video segments. In the context of our work, this serves as a powerful tool for indexing and retrieving specific moments based on the observed human pose.

Extensible Feature Extraction Pipeline:

ArchiveVault's metadata extraction is designed to be modular and expandable. Any feature that can be represented as a vector embedding—including deep audio embeddings, image features (e.g., ResNet), and environmental sound classification—can be indexed and searched. The system allows users to define custom AI pipelines, seamlessly integrating new embeddings as archival needs evolve.

2. Scalable Archival Infrastructure

To support the efficient storage, indexing, and retrieval of audiovisual data at scale, ArchiveVault employs a robust and scalable infrastructure orchestrated by Kubernetes, or it can be run locally. While a local setup changes the ingestion time, smaller archives can still be scaled.

The system allows for seamless expansion as new AI models and feature types are introduced, efficiently processing audio, text, and vision tasks in parallel using a unified object store. ArchiveVault can also be started from multiple machines in parallel, provided the database and S3 storage are accessible by the other computers, as well as the archive itself. This scalable architecture enables efficient archival management and high-performance search capabilities across vast collections. Below, we outline the key infrastructure components that enable efficient archival management and high-performance search capabilities.

Hybrid Indexing for Fast Retrieval:

Our methodology combines PostgreSQL's inverted indexing for textual metadata with vector-based semantic search (pgvector), enabling rapid and semantically nuanced retrieval. The combination of these approaches enables users to execute queries based on both literal textual content and underlying semantic meaning. This dual-faceted indexing architecture allows users to locate relevant information quickly and accurately by searching for specific keywords or exploring conceptually related ideas. The inverted index facilitates efficient keyword-based searches by identifying documents containing the specified terms. Concurrently, pgvector generates vector embeddings of the content, capturing semantic relationships between information. This allows the system to retrieve conceptually similar results, even without identical keywords, enhancing information retrieval scope and precision.

Separation of Raw Content & Metadata Storage:

ArchiveVault employs a dual-storage architecture, separating raw audiovisual data from extracted metadata. This distinction facilitates efficient metadata updates without processing large files and improves query performance by enabling rapid searching and filtering based on metadata attributes. This approach aims to streamline content management and enhance data accessibility. The strategy utilizes Minio for raw audiovisual data storage and Postgres for metadata storage to achieve this separation.

Indexing for Real-Time Queries:

The system uses a dual-indexing strategy for real-time queries, combining inverted indexes for keyword searches with vector-based search for semantic similarity. Inverted indexes quickly map terms to documents, while vector embeddings capture semantic meaning, allowing the system to find conceptually similar documents. This integrated approach delivers fast, scalable, and comprehensive search results for growing data and user demand without impacting performance.

Expandable Search Capabilities:

The architecture allows for seamless integration of new vector-based retrieval approaches. Any new feature descriptor, such as face embeddings, environmental sound classifications, or video action recognition embeddings, can be incorporated into the search framework without rearchitecting the system. The architecture is designed to readily accommodate the integration of novel vector-based retrieval methodologies.

This inherent flexibility ensures that diverse feature descriptors. This adaptability eliminates the need for substantial system re-engineering or redesign when new feature extraction techniques or data modalities become available, providing a future-proof and extensible search capability.

3. Advanced Retrieval Mechanisms

Retrieving relevant content from vast audiovisual archives requires more than just keyword searches. ArchiveVault incorporates advanced retrieval mechanisms that allow users to search content based on semantics, visual similarity, and structured metadata. By leveraging AI-driven techniques, the system ensures high-precision results across multiple modalities. Below, we describe the key retrieval features that enhance the user experience and improve content discoverability.

Semantic Search Across Modalities:

ArchiveVault allows users to perform content searches using natural language queries, retrieving results based on speech transcripts, named entities, and visual elements. This uses the same mechanism as the transcript generation and vector embeddings.

Pose-Based Retrieval for Sports Media:

By indexing body postures and movements, ArchiveVault enables retrieval based on athlete pose similarity, providing novel ways to explore archival content.

Context-Aware Search Filtering:

Users can refine search results based on modality (e.g., spoken content vs. visual elements), leveraging intelligent ranking algorithms to improve content relevance.

Vector-Based Retrieval Expansion:

The system is not limited to specific search paradigms but is inherently designed to support additional AI-driven search modalities, making it adaptable to future research needs and emerging retrieval techniques. This enables multiple possibilities, such as easily setting up a Retrieval Augmented Generation (RAG) system on top of the existing architecture, using the same database structure.

Case Studies

ArchiveVault's AI-driven approach was rigorously tested through the processing of diverse audiovisual archives. These tests were specifically designed to highlight the system's inherent functionality and overall effectiveness in managing and extracting value from varied content. Our advanced metadata extraction and intelligent retrieval techniques were applied across a spectrum of distinct use cases, each presenting unique challenges and requirements.

The objective was to demonstrably showcase how the implementation of automated workflows can substantially enhance both the accessibility and the discovery potential within extensive, large-scale digital collections. By addressing different types of audiovisual material and user needs, we aimed to provide compelling evidence of ArchiveVault's broad applicability and significant advantages. The subsequent case studies offer detailed illustrations of the system's tangible impact and practical benefits within authentic real-world scenarios and institutional contexts. These examples will further elucidate the power and versatility of ArchiveVault in transforming how audiovisual archives are managed and utilized.

1. RTS Swiss Television Archive

ArchiveVault was validated in collaboration with Radio Télévision Suisse (RTS), the French-speaking national broadcaster of Switzerland. RTS possesses a vast archive of audiovisual content accumulated over decades, significant portions of which lack comprehensive metadata. By leveraging AI-powered speech transcription, named entity recognition (NER), and speaker segmentation, ArchiveVault successfully indexed thousands of hours of this footage. This indexing enables users, including journalists and archivists, to conduct entity-based searches, specifically by Locations, Persons, or other Entities, retrieving precise clips where these entities are mentioned.

Additionally, the system supports natural language queries, providing users with the most relevant clips in response. The RTS archive encompasses 212,338 videos, totaling over 60,000 hours of

content from 1950 to the present, including 1,999,672 editorial clips. While some structural catalog data exists, it only partially assists with detailed discovery. During ingestion, challenges included variability in audio quality and inconsistencies in entity tagging. Adaptive preprocessing and iterative refinement of entity linking significantly improved the metadata quality and retrieval accuracy.

Processing:

We ingested the whole archive available to us through ArchiveVault. WhisperX-large generated word-level transcripts with diarization; spaCy tagged named entities, which we reconciled against Wikidata IDs; speaker and sentence embeddings were stored in pgvector. Overall the slice produced 138.3 M words, 1.1 MIL unique entities and 52 ± 2 time-coded clips per source video. Table 1. gives an overview over the entities processed in the archive. With millions of entities, disambiguation becomes a challenge on its own with many entities to process. With this, this demo only focused on the entity types Person and Location. Figure 2. shows an example of the processed archive with its entities in the form of the Map of Switzerland. The entities and locations have been matched against wikidata to receive the locations on the map.

Table 1: Entities processed in the RTS archive. Most of the entities only appear a single time, with the average mention being 4.7 times.

Entity	No. of Results
Person	511,124
Location	2,487,030
Organisation	657,205
Miscellaneous	1,460,374

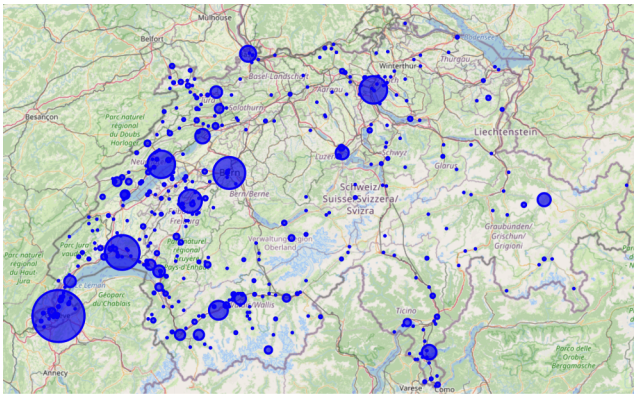

Figure 2. *Example annotations performed by ArchiveVault. Most of the french speaking swiss content is centered around the western part of Switzerland.*

2. Olympic Games Pose Matching

We applied ArchiveVault to the extensive Olympic sports footage archive, targeting a specialized retrieval task: pose-based searches for athletic movements. Since the application supports embedding of various types, similar poses can be efficiently retrieved. Using pose extraction algorithms, we indexed images and video segments based on skeletal motion data, enabling comparison of poses across different events. For example, a curator looking for iconic Olympic movements, such as high jump techniques over several decades, could utilize ArchiveVault to find comparable postures. This capability facilitates comparative athletic analyses and thematic storytelling within sports media. Additionally, the system revealed unexpected similarities, such as pose alignments across different sports, demonstrating how AI-driven analysis can uncover new insights within archival collections.

Why a Separate Pipeline?

RTS archives primarily require retrieval based on spoken-word content. In contrast, the IOC's historical sports footage emphasizes visual motion similarity, necessitating a dedicated pipeline.

Pose Extraction

For an initial set of Olympic Games coverage, we employed OpenPifPaf to extract 17-point skeletal vectors from several million keyframes (approximately 10 frames per second). Each pose vector was stored alongside its corresponding frame time-code and broadcast metadata. These vectors were indexed using HNSW in pgvector, enabling nearest-neighbor lookups within 200 milliseconds, even on commodity GPUs. Subsequently, OpenPifPaf was executed on every single video frame, providing precise pose data and enabling parallel annotation across a vast quantity of videos.

Conclusion

ArchiveVault is presented as a system designed to enhance the management and accessibility of audiovisual archives by integrating Artificial Intelligence-driven metadata extraction with archival workflows. It addresses the increasing complexities of managing large volumes of media through scalable indexing and intelligent search capabilities that aim to improve content discoverability. The system utilizes a range of AI technologies, including speech-to-text transcription for spoken content analysis, named entity recognition for identifying entities such as individuals, organizations, locations, and events, and pose-based retrieval for the analysis of visual elements like body language and actions.

This approach facilitates a detailed analysis of archived material beyond traditional keyword-based methods. The automated generation of detailed metadata aims to reduce the resources required for manual cataloging. The enhanced accessibility provided by ArchiveVault intends to enable researchers, historians, journalists, and the public to explore audiovisual resources, potentially facilitating broader utilization of archival content. The platform's scalability is designed to accommodate archives of different sizes and expanding collections. Future work includes integrating advanced multimodal

embeddings and extending the pipeline to support real-time analytics and retrieval in live archival scenarios.

References

[1] Alliata, G., Rattinger, A., Benzi, K., & Kenderdine, S. (2024). AI-Driven Workflows for Unlocking Switzerland's Collective Memory: Distant Listening of the RTS Archive. In DARIAH Annual Event 2024 Submission.

[2] Alliata, G., Yang, Y., & Kenderdine, S. (2023). Augmenting the metadata of audiovisual archives with NLP techniques: Challenges and solutions. Digital Humanities 2023

[3] Colavizza, G., Blanke, T., Jeurgens, C., & Noordegraaf, J. (2021). Archives and AI: An overview of current debates and future perspectives. ACM Journal on Computing and Cultural Heritage (JOCCH), 15(1), 1–15.

[4] Lewis, P., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459–9474.

[5] Radford, A., et al. (2023). Robust speech recognition via large-scale weak supervision. In International Conference on Machine Learning.

[6] Rudnik, C., et al. (2019). Searching news articles using an event knowledge graph leveraged by Wikidata. In Proceedings of the 2019 World Wide Web Conference, 1232–1239.

[7] Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). spaCy: Industrial-strength Natural Language Processing in Python. doi:10.5281/zenodo.1212303

[8] Jaillant, L. (2022). Archives, access and artificial intelligence: Working with born-digital and digitized archival collections. Bielefeld University Press.

[9] Kreiss, S. et al. (2021) .OpenPifPaf: Composite Fields for Semantic Keypoint Detection and Spatio-Temporal Association

[10] Bain, M., Huh, J., Han, T., & Zisserman, A. (2023). Whisperx: Time-accurate speech transcription of long-form audio. arXiv preprint arXiv:2303.00747.

Author Biography

André Rattinger is a Senior Machine Learning Engineer at the Laboratory for Experimental Museology (eM+), EPFL, Lausanne, Switzerland. His current work focuses on designing and implementing distributed AI pipelines for large-scale multimedia archives, including extensive projects with Swiss Television (RTS) and the International Olympic Committee (IOC). André specializes in multimodal feature extraction, scalable data infrastructures, and immersive visualization systems, often leveraging Retrieval-Augmented Generation (RAG) techniques for advanced multimedia analytics. Previously, he developed machine learning-driven anti-abuse systems at ProtonMail and contributed to semantic visualization tools at CERN. André holds an M.Sc. in Computer Science from Graz University of Technology, Austria, and has authored several publications on information retrieval, semantic embeddings, and network visualization.