media/rep/

Repositorium für die Medienwissenschaft

Giulia Taurino

Dialogues with Audiovisual Archives

2024

https://doi.org/10.25969/mediarep/24077

Veröffentlichungsversion / published version
Zeitschriftenartikel / journal article

Empfohlene Zitierung / Suggested Citation:
Taurino, Giulia: Dialogues with Audiovisual Archives. In: *VIEW. Journal of European Television History and Culture*. Archive Television: Storing, Structuring and Accessing Audiovisual Content in the Time of Algorithmic Curation, Jg. 13 (2024), Nr. 26, S. 87–124. DOI: https://doi.org/10.25969/mediarep/24077.

Erstmalig hier erschienen / Initial publication here:
https://viewjournal.eu/articles/10.18146/view.358

Nutzungsbedingungen:
Dieser Text wird unter einer Creative Commons - Namensnennung - Weitergabe unter gleichen Bedingungen 4.0 Lizenz zur Verfügung gestellt. Nähere Auskünfte zu dieser Lizenz finden Sie hier:
https://creativecommons.org/licenses/by-sa/4.0

Terms of use:
This document is made available under a creative commons - Attribution - Share Alike 4.0 License. For more information see:
https://creativecommons.org/licenses/by-sa/4.0

DFG Deutsche Forschungsgemeinschaft

Philipps Universität Marburg

volume 13 issue 26/2024
VIEW
JOURNAL OF EUROPEAN TELEVISION HISTORY & CULTURE

DIALOGUES WITH AUDIOVISUAL ARCHIVES

Giulia Taurino
Northeastern University
g.taurino@northeastern.edu

Abstract: This curated section[1] explores the state of audiovisual archiving in Europe and North America, emphasizing the dynamic interplay between digitization efforts and the integration of artificial intelligence (AI) technologies within traditional archival workflows. Through interviews with archival professionals from key institutions, including the Latvian State Archive of Audiovisual Documents, Institut National de l'Audiovisuel, National Library of Sweden, Austrian Broadcasting Corporation, GBH, and the American Archive of Public Broadcasting, this partial ethnography of audiovisual archives provides insights into the histories, management structures, and mission statements that shape these heritage institutions. The findings reveal that while all institutions share the common aim to preserve audiovisual heritage, differences in their cultural contexts and operational frameworks influence their approaches to digital management and AI implementation. Key challenges identified include funding constraints and the necessity of maintaining archival integrity amidst technological transitions. Each archive's strategy for integrating AI tools for metadata extraction and content analysis is discussed, highlighting both opportunities and challenges. The study ultimately underscores the importance of balancing traditional archival practices with innovative technological advancements to enhance the accessibility and discoverability of audiovisual collections, ensuring their preservation for future generations while addressing evolving digital landscapes.

Keywords: algorithmic curation; archival management; artificial intelligence; audiovisual heritage; copyright; digital preservation; media history; television archives

Introduction

The state of television archives has become an essential focus of contemporary audiovisual archiving, particularly in light of increasing digitization efforts in broadcast archives and the integration of artificial intelligence (AI) technologies for the operationalization of tasks pertaining to digital preservation, cataloguing and information retrieval. These 'dialogues with audiovisual archives' present a comparative analysis that revolves around several key interviews with archival professionals and specialists from various heritage institutions in Europe and North America. Drawing a partial ethnography of audiovisual archives, this section provides insights into the histories, management structures, and mission statements that shape these heritage institutions. For the purpose of a comparative study, we have interviewed European archives - namely, the Latvian State Archive of Audiovisual Documents at the National Archives of Latvia (LSAAD, Latvia), the Institut National de l'Audiovisuel (INA, France), the National Library of Sweden (NLS, Sweden), the Austrian Broadcasting Corporation (ORF, Austria) -, as well as US archives selected for their significance - the GBH

(the first public US broadcasting station to create an archive[2]), and the American Archive of Public Broadcasting (AAPB, a long-term initiative to preserve and make openly accessible significant historical content broadcasted by local public television stations[3]).

Ordered by longevity and geographic location from the introductory interview to the closing piece, these institutions differ in terms of their governance structures, mission focus, funding sources, and the specific roles they play within their respective national or local media landscapes. While they all share the common goal of preserving audiovisual heritage, their histories reflect various cultural, organizational and political contexts. For instance, the LSAAD, founded in 1963[4], is a state archive, and as such it is primarily funded by the government, with budget constraints impacting its operations. Given its focus on preserving and providing access to audiovisual documents that reflect Latvia's cultural and historical heritage, it acts as a state institution responsible for managing the local audiovisual heritage under specific regulations for document submission. Similarly, INA and NLS, respectively created in 1975[5] and 1978[6], are national libraries tasked with preserving national audiovisual heritage, including audiovisual materials (mainly sound and moving images, but also printed documents or photographs) collected under the mandate of local legal deposit laws. INA was originally created from the remnants of France's public broadcasting system. It receives funding from the French government and maintains commercial operations for the reuse and sale of its materials, thus functioning as both an archive and a commercial entity, supporting cultural institutions and media industry professionals alike. NLS is also supported by government funding, with an emphasis on collecting national cultural heritage. It serves both as a preservation institution and an access point for scholarly use, emphasizing a broad approach to media preservation that involves research and education. Archives like the ORF Archives (operating since the 1920s as a radio archive and since the 1950s as a film archive and restyled as a service-oriented archive in 1986[7]) and GBH Archives (operating since the 1950s and officially founded in 1979[8]) have a different institutional identity than the LSAAD, INA and NLS. They both are public broadcasting media corporations, and as such they focus on the preservation, access, and distribution of their own content. For example, the ORF Archives highlight Austrian audiovisual culture and programming broadcasted by ORF channels, which are funded through a mix of broadcasting fees, advertisements, and government support. ORF covers two institutional functions as both a content creator and archiver, prioritizing its own productions and other relevant external material, and so does GBH, a Public Broadcasting Service (PBS) member television station. Other than holding its own archival *fonds*, GBH is part of a collaboration with the Library of Congress aimed at preserving public media across the US and facilitating access to historical broadcasting material under the name of American Archive of Public Broadcasting (AAPB). AAPB is funded through a combination of government grants, public funding, and private donations. It acts as both a repository for digital audiovisual materials and a partner in archiving, community engagement and outreach initiatives for public media. While we tried to cover a diverse range of institutions, we know that this selected group of audiovisual archives is not necessarily comprehensive of all archival experiences and many more can be added to this list. Nevertheless, the ensemble of these perspectives provides yet a closer look into present-day television archives, in the form of a preliminary state-of-the-art survey on the hybrid nature - both analog and digital - of audiovisual archiving methods in Europe and the US.

The archives interviewed in this section operate under distinct frameworks and historical contexts leading to different technology adoptions. However, they do share common features in their approach to digital archiving methods. The LSAAD has faced significant challenges concerning the description and digitization of its vast collection, which includes hundreds of thousands of audiovisual documents. Despite modifications initiated in 2016, limitations remain in adequately documenting these materials, which affects public access and research opportunities. As a solution, specialists at LSAAD have explored multimodal similarity search that relies on visual and sound data, rather than on textual description only. INA, contrastingly, has established a more robust digitization program and B2B access platform, enabling professional users to engage with their extensive audiovisual media holdings and their metadata bases. At first glance, this highlights a distinction in how archives align their operational structures with digital management systems and AI solutions, depending on the state of the *fonds* acquired, archival priorities in handling record-keeping and knowledge organization, as well as

available assets - including, among others, long-term financing plans for digitization and metadata extraction. However, both LSAAD and INA share the commitment to integrate AI in their workflows in ways that align to their missions and facilitate the ultimate scope of providing broader access to audiovisual collections. Meanwhile, in a similar fashion, the NLS has focused on building a comprehensive framework that emphasizes user access through a hierarchical storage management system, enhancing the searchability and discoverability of its content while addressing the complexities of metadata management. The ORF Archives reflect a comparable trajectory, having digitized 95% of their collection while also navigating the intricate relationship between physical and digital preservation. In the United States, the AAPB and GBH archival initiatives exemplify all these experiences with digital archiving combined. On the one hand, they leverage adaptive computational methods and collaborative partnerships with local stations to enhance the archival landscape for public broadcasting media. On the other hand, they acknowledge existing issues for audiovisual archiving in the digital age.

Through this comparative lens, the integration of technologies for records' digitization, along with the introduction of AI toolkits in archival workflows and automated algorithmic curation pipelines, emerges as a widespread and urgent topic across the archival community. In light of the archival turn in AI,[9] audiovisual archives are facing their own field-specific questions on how to ensure a responsible use of AI for recordkeeping and mitigate possible risks of a parallel AI turn in digital archiving. Despite recent advancements in archival sciences and computation, all institutions encountered in these interviews share concerns and difficulties, primarily around funding, resource allocation, and maintaining the integrity of archival practices amidst technological transition. The adoption of AI tools is thus welcomed with some amount of internal discrepancy across different audiovisual archives: all institutions interviewed are torn between efforts to integrate AI in their workflows for metadata extraction and content analysis, on the one hand, and more cautious and strategic approaches, on the other hand. These conversations with archival specialists ultimately speak for an ongoing quest of our times, which witnesses attempts to balance algorithmic innovation - being it in the form of a speech-to-text automated transcription model, a face recognition software or a multimodal similarity search tool -, with traditional methodologies from the archival paradigm - such as documentation of records-provenance, *respect des fonds* and other well-established approaches to information management and knowledge extraction.

As physical storage systems evolved into digital repositories, these dialogues with audiovisual archives emphasize the importance of maintaining archival authenticity and contextual integrity while adjusting to rapidly changing digital landscapes. Moreover, the ongoing transformation of these archival repositories illustrates a broader trend to professionalize a digital archival discipline where traditional methods are harmoniously blended with modern technological practices, enhancing the preservation, accessibility and discoverability of audiovisual heritage for future generations. With this in mind, we have attempted to bridge these conversations with use cases that appear in other contributions to this editorial issue, stressing on the ultimate scope of audiovisual archiving: to guarantee long-term conservation and access to audiovisual records.

For this purpose, the interview section concludes with an 'in-concert interview', a conversation with an ensemble of archival professionals from Centrum Cyfrowe Foundation (Poland), EuroClio (Netherlands), Europeana Foundation (Netherlands), Netherlands Institute for Sound & Vision (Netherlands), TV3 Corporació Catalana de Mitjans Audiovisuals (Spain). This last interview about the EU project *Watching Videos like a Historian* echoes Sauer et al.'s reflexion on audiovisual archives as learning environments, as well as Blume et al.'s paper on unlocking treasures from television archives for research. In a way, it also reconnects the archival discussion to Treleani et al. work - developed in collaboration with six national channels in France and Italy, namely France 2, Arte, TF1, Rai Uno, Rai Due and Canale 5 - on the traces of audiovisual archives in television and on the web, and Re's analysis of curatorial practices left over outside of the archival limits, in the tension between a finite 'archive model' and the open-ended 'network model.' Finally, the scope of this interview section is to guide the reader behind the curtains of audiovisual archives, so as to understand the archival logics that regulate the transition of audiovisual heritage into the age of digitization, computation and AI, and how this affects the circulation of audiovisual records within and out of the original archive.

LSAAD

Interview with Natālija Lāce, Rūdolfs Henčels, and Svyat Elkind (Latvian State Archive of Audiovisual Documents, 1963-present) [10]

This interview with archive professionals Natālija Lāce[11], Rūdolfs Henčels, and Svyat Elkind walks the reader through the history of the Latvian State Archive of Audiovisual Documents (**LSAAD**), from its origins up until its current outlook. Officially established in 1963 after evolving from earlier archives, it has played a key role in preserving Latvia's audiovisual heritage through various political climates. As of 2024, the LSAAD houses over 2 million records among photographic, video, and sound material. Although more than half of the documents are described, only 20% of the audiovisual material is currently accurately catalogued. The majority of descriptions are not reliable, and many lack digital copies, affecting accessibility. Starting from a detailed presentation of the archive, the speakers bring our attention to a sometimes overlooked issue in audiovisual archival management: the need to update existing labels and metadata in order to correct former classification practices relying on ideological beliefs, outdated language or poor documentation. This conversation with archive and digitization experts sheds light on the role of digital record management and computational processing in favoring positive improvements in database functionality and document searches in multimedia collections. While the speakers acknowledge that archivists face various challenges in integrating IT skills and ensuring that the contextual historical significance of archival collections remains intact, they share an important perspective on the future of AI in audiovisual archives. Aiming to balance traditional methods with modern technology to enhance digital accessibility through digital literacy programs and CS training might just be the solution for the long-term sustainability of large-scale digital archiving of audiovisual heritage.

Figure 1. Image of miscellaneous records from the Latvian State Archive of Audiovisual Documents. © Photo by Ieva Ozolina.

1. Can you present the LSAAD collection at its present stage?

   As of 2023, the archive holds 1,705,704 photographic documents that have been described, with 300,559 remaining undescribed. For what concerns film documents, 111,310 are described, while 12,835 are undescribed. The archive holds a total of 54,000 video documents, of which 7,000 have been cataloged. Additionally, 55,479 sound documents have been described, leaving 8,794 undescribed. Most documents were described during the Soviet era, corresponding to the ideology and censorship of that time. And despite the large number of described documents, only 20% are accurately described, while the remaining descriptions require revision. It is also important to note that, although these documents have been described, they don't have digital copies, which significantly complicates their accessibility and use by the public. The archive's database contains only 385,631 documents, without distinguishing between document types. In 2024, nearly 1 million photographic documents and several dozen film documents were submitted to the archive. These documents are not yet included in the overall statistics, as the evaluation of their content and their physical condition are still ongoing. The process is very slow due to limited human resources, as the archive currently consists of four departments and employs only 38 people (including administrative and technical staff).

2. What is the history of this archive in terms of its institutional and curatorial lineage (provenance, acquisition, preservation)?

   The Latvian State Archive of Audiovisual Documents traces its origins back to the early 20th century, when new types of records—such as audiovisual, photographic, and sound documents—began to emerge in archival practices. Initially, these records were deposited in the National or Historical Archives of Latvia, alongside other documents which were submitted to the archive by private individuals or institutions, following the principle of the indivisibility of fonds. This principle required that all documents had to be kept together to preserve their historical context. However, since the storage conditions required for photo, film, and sound documents differ significantly from those of paper records, challenges arose. The archive initially operated under constrained conditions, including inadequate storage facilities. By the late 1960s, collections were spread across multiple locations, including the historic Bulduri Church, where film documents were stored. The church was repurposed for archival use despite being unsuitable for long-term preservation. The archive's holdings steadily grew, incorporating films, photographs, and sound recordings from various public and private sources, including the Rīga Film Studio and other state institutions. This growth led to the decision to officially establish the archive in 1963 as the Central State Archive for Film, Photo, and Sound Documents of the Latvian SSR, making it the first institution in Latvia dedicated to the collection and preservation of these types of documents. Throughout its history, the Archive has navigated various political regimes, playing a vital role in preserving a significant portion of Latvia's material cultural heritage. Today, it safeguards six distinct types of documents across more than fifty media formats, including films, photographs, video documents, sound recordings, posters, and electronic files. Each type serves as a valuable historical medium, providing insights into Latvia's culture, history, and societal developments from multiple perspectives. The archive has received documents from private individuals, state, and municipal institutions, as well as private companies. Record creators deposit their documents into state custody without compensation, but they have the option to receive digital copies of their submitted records. Since the restoration of Latvia's independence in 1991, the submission of documents by state and municipal institutions has been regulated by Cabinet of Ministers regulations. These rules specify the timeframes and the required level of description and organization for the documents to be handed over. Film studios that have produced films using archival materials and received state co-financing are legally required to submit a copy of their film to the archive. All other submissions are made voluntarily. Upon receiving documents, the archive signs an agreement with the record creator, outlining the public accessibility and usage terms of the materials. Copyright issues are managed by the Copyright and Communication Consulting Agency/Latvian Authors Association. The archive does not handle copyright matters, but if documents are subject to copyright, their availability in the public database is restricted. However, access within the internal database is not limited, and any client can view these documents in person at the archive.

Regarding the types of documents accepted by the archive, they can be divided into two distinct periods: during Latvia's occupation by the USSR and after the restoration of independence. During the Soviet era, documents that aligned with the ruling regime's ideology and propaganda were prioritized. These documents portrayed Soviet Latvia as an economically and socially successful, and continuously developing nation, shaped by the influence of the USSR. Many documents related to Latvia's history before World War I were not submitted to the archive during this period due to fear of repression. People often hid these records in attics, under floorboards and other places. These documents were submitted to the archive after Latvia regained its independence in 1991. Since the early 2000s, a significant number of documents have been donated by Latvians who lived in exile after World War II. These contributions have included numerous radio recordings, such as broadcasts from 'Radio Free Europe' and 'Voice of America', which aimed to inform the Latvian population about events beyond the Iron Curtain. Although Soviet authorities tried to block these broadcasts, some Latvians managed to intercept them, providing a crucial link to the outside world during the Soviet occupation. Starting from 1991, the archive has accepted documents based on their content and significance within the broader context of Latvia's historical, social, political, or economic events, regardless of the historical time period. Upon acquisition, these documents undergo a meticulous process that includes cleaning, cataloguing, detailed description, and digitization, before being integrated into the archive's database.

3. When was the first digitization, data migration and metadata integration effort initiated at the LSAAD?

The first data migration took place in 2016, when an archival database was created, and video and film document records were migrated from the previous Microsoft Access database. In 2018, photo and audio document records described in the Access database were also migrated. As a result of this migration, a separate entry was created in the database for each described document, containing the main descriptive metadata but without the associated digital files. The digital files for photo documents were bulk uploaded to the records in 2018 using a specially developed script, which matched each file to its corresponding record in the database based on the document number assigned to each file. This was possible because photo document digitization in the archive involves scanning positives with a standard document scanner, while negatives are digitized using a film negative scanner. For larger-scale negative digitization or photo albums, these items are sent to a professional photo lab. As a result, the archive had the necessary digital files available to attach to these records. Video, film, and audio document records initially had no associated files, as the archive lacked the necessary equipment for digitization at that time. The first digitization equipment 'Telecine' was donated to the archive in 2008, allowing for the transfer of documents to DVCAM format. However, the first large-scale digitization of video, film, and audio documents only began in 2017 as part of an EU-funded project, with the film studio 'Lokomotive' providing the service. A total of 9,872 archival units of film, video, and audio documents were digitized. The Archive's Film and Video Digitization Center, established in 2021, has significantly enhanced its capability to digitize and preserve video, film, and audio documents in-house at a high resolution. The digitization process utilizes cutting-edge equipment, including the Lipsner Smith CP9400 HFS ultrasonic film cleaning machine, a Blackmagic Cintel Scan with Davinci Resolve 19.2 software for 35mm film image and optical sound scanning, and a Filmfabriek HDS with Cine2digits software for 16mm and 8mm films. Two Dell scanning stations equipped with Windows 10 Enterprise, Intel(R) Xeon(R) Gold 6242 CPUs, 128 GB RAM, Nvidia Quadro RTX 6000 GPUs, and dual Philips UHD reference monitors provide robust hardware support. The stations also feature ARECA RAID storage with a 40TB capacity and an Isilon storage server offering 400TB, connected via 10 Gbps optical fiber. Tools like Md5checker for checksum generation and Terracopy for file migration ensure data integrity during the digitization process. For postproduction and digital restoration, the center employs two additional Dell workstations, featuring similar high-performance specifications, including TVlogic 4K color-precise monitors and 4TB SSDs for swift data processing. These workstations leverage software such as Davinci Resolve 19.2 for color correction, Adobe Premiere 2024, and Adobe Encoder 2024 for editing and deliverables, as well as Digital Vision Phoenix 2024.2 for visual restoration and Acoustica 7 for sound processing. This comprehensive setup allows the archive to deliver high-quality restoration and preservation services, ensuring that its growing

collection of audiovisual documents remains accessible and preserved for future generations. By 2023, more than 4,300 film, video, and audio document units had been digitized, with this number growing annually. Digital files are added to both the records migrated in 2016 and newly described documents in the archival database. Since 2023, the archive has also been preparing data packages that are sent to Preservica.

4. How did the digital archival management of the LSAAD collection change over time?

   Since both this practice and the archive's database are relatively recent, it is difficult to speak of significant changes in digital archival management. Over time, we have improved the database's functionality by adding new features that facilitate document search, synchronizing author and person classifiers, and introducing bulk record creation through CSV file imports. Our current focus is on making the database as user-friendly as possible for archive staff, enabling more bulk operations, and ensuring that the system is compatible with those of other institutions, thus enhancing interoperability.

5. Could you provide a brief overview of how the digital repository system works in synergy with the material archive for tasks like information retrieval and querying of the records?

   Electronic records form items described in inventory lists, just as paper records form items described in separate inventory lists. At this level, it is possible to link items across different inventory lists, which aids in retrieving information from the archive. Currently, searches are primarily a manual process, as our information system, digital repository, and material archive are not yet integrated.

6. What are the main challenges and opportunities you encountered during the development of digital and/or computational projects at the LSAAD?

   Since we don't have our own IT department, all our needs are managed by the Culture Information Systems Centre, which also supports the information systems of libraries and museums. Unfortunately, archives are not their primary focus. As a result, any project we undertake is a challenge, as we don't know if the project will start as planned or if we'll have financial support through to its completion. Nevertheless, we remain optimistic.

7. Does the LSAAD plan on implementing AI methods for the management of its collections? If it did already, which AI projects are currently ongoing? And if not, do you foresee integrating any algorithmic processing to the archival work?

   The Latvian State Archive of Audiovisual Documents has begun integrating artificial intelligence (AI) into its archival processes, although its efforts are still in the experimental phase. In 2020, the archive partnered with NOSCOS, a Latvian IT company, to implement a facial recognition tool for identifying individuals in photo and audiovisual documents. This tool was integrated into the archive's database via an API, allowing for the automated identification of individuals based on confidence scores. However, the system required significant manual verification and training to improve its accuracy. Despite its benefits, such as compatibility with older systems and customization to local needs, the process was resource-intensive and time-consuming. In early 2024, the archive tested Archipanion's AI tool for image and video search, as well as sound transcription in Latvian. While the results were impressive, the service posed budgetary challenges due to its subscription model, where fees increase with the growing number of searchable documents. Therefore the archive is exploring more sustainable alternatives and has created the Multimodal Similarity Search Prototype. This internal AI-based tool, utilizing open-source models such as Deep Face and Meta deep learning frameworks, enables vector-based searches across approximately 276,000 TIFF files. It can perform face and image similarity searches, though there are ongoing challenges with false positives. The prototype represents a significant step forward in enhancing search capabilities without reliance on external providers. Given that the entire process of document

acceptance and description is very slow, as previously mentioned, the archive plans to address this by securing project-based funding. Through outsourcing, the archive aims to mass-digitize the undescribed photographic documents and store them in a separate digital repository. This approach would allow both archive staff and clients to search for and access relevant images using the AI tool, even before the documents have been described and added to the archives database. In the long term, the archive is considering developing custom AI tools for description, transcription, and image search. This would allow the archive to maintain full control over its systems and ensure they are specifically tailored to Latvian audiovisual collections. Given Latvia's linguistic diversity, which includes documents in Latvian, German, and Russian, these AI tools must accurately handle complex multilingual recognition and transcription. However, financial and human resource constraints limit the archive's ability to develop and maintain such advanced systems. To overcome these challenges, the archive is actively working on expanding international collaborations and securing additional funding. Currently, it is involved in several global organizations, such as the International Council on Archives, and has joined the ARCAM-net initiative ('International Network for Advanced Research on Computational Audiovisual Media'). These partnerships help enhance the archive's competencies and foster international cooperation, paving the way for the successful implementation of advanced AI solutions.

8. In general, how does the LSAAD as an institutional entity position itself in the tension between algorithmic and non-algorithmic archival practices? And how do you, as archival experts, foresee future evolutions of AI applications for the management or use of audiovisual archives?

   The Latvian State Archive of Audiovisual Documents positions itself as a forward-thinking institution, balancing traditional archival methodologies with the integration of emerging algorithmic practices. Historically, the archive has relied on manual, non-algorithmic processes for cataloguing and managing its collections. These practices emphasize human oversight, ensuring that the historical, social, and cultural contexts of documents are meticulously preserved. However, given the archive's growing collection and limited resources, it has increasingly recognized the potential of algorithmic tools to enhance efficiency and accessibility. As a result, the archive has taken the previously mentioned steps to implement AI tools, particularly to facilitate the search and retrieval of documents within its digitized but undescribed collections. These advancements aim to bridge the gap between traditional practices and the modern demands of large-scale digital archiving, ensuring that the archive remains both efficient and user-focused. While we are actively working on implementing AI tools to support the archival workflow, it is important to state that streaming platforms have not yet affected the work of the Latvian State Archive of Audiovisual Documents. The archive continues to rely on traditional methods for critical tasks, such as evaluating the historical and cultural significance of documents and ensuring contextual integrity in their descriptions. At the same time, it actively seeks ways to leverage AI for processing larger volumes of documents. However, the archive avoids integrating tools into its database that perform on-site subtitle generation for video, film, and audio documents, or those that automatically generate annotations and automated content indexing or filtering. This decision is based on testing, which has shown that such tools achieve a maximum accuracy of around 80%. Implementing these tools would require significant human resources to manually identify and correct errors, a capacity the archive does not currently have. Given the archive's commitment to publishing only high-quality documents and metadata in its public database, the use of imprecise tools would compromise its stringent quality standards. While it is acknowledged that no existing AI tool can deliver 100% accuracy, and users would be informed of the potential for errors, there is a significant risk that some users might uncritically accept AI-generated inaccuracies as factual. If these errors are not identified and corrected by archivists, it could lead to the falsification of historical records, a risk the archive is unwilling to take. This underscores the institution's cautious approach to integrating AI in ways that do not compromise the integrity and reliability of its collections. We could say that the LSAAD is navigating the tension between algorithmic and non-algorithmic archival practices by adopting a hybrid approach. We recognize the value of AI as a tool to complement traditional methodologies, ultimately aiming to enhance the accessibility and management of its collections. As AI technology continues to evolve, the archive is willing to adapt and implement innovative solutions that respect both the integrity of its collections and the growing demands of its users. However, in our view, it is crucial to approach AI tools with caution and critically evaluate their results.

Rather than embracing them uncritically, we should only adopt those tools that are genuinely beneficial and help streamline daily operations, rather than burdening staff with additional tasks of verifying results and correcting errors. It is equally important to consider the specific nature of each archival document and the significance of its contextual background. While AI can identify visual elements in an image, it cannot determine the context in which a document was created—a task that requires explanation and interpretation by the archive for the benefit of the public. Therefore, the implementation of AI tools in archival practices should not be an ultimate goal. Each tool must be carefully assessed for its utility and relevance, ensuring it supports the archive's mission to preserve and communicate the contextual and historical significance of its collections without compromising on accuracy or quality.

9. Is there anything you want to add based on your work with audiovisual archives, digitization and AI ?

   An important aspect—and arguably a challenge—in Latvia is that the majority of archive staff hold degrees in history, philology, political science, film directing, or as cinematographers. These qualifications are invaluable for tasks such as the acquisition, evaluation, description, and presentation of historical documents to the broader public. However, with the rapid development of AI tools in recent years, it has become clear that this expertise alone is no longer sufficient. As I mentioned, the archive does not have its own IT department and Culture Information Systems Centre does not specialize in developing or implementing AI tools for audiovisual documents. For example, while a text transcription tool called 'HUGO' has been developed, it is unsuitable for creating subtitles as it only recognizes clearly enunciated Latvian in standard language, which is rarely found in historical documents. Thus, it does not meet the archive's needs. As a result, archive staff must continuously acquire knowledge in IT, programming, and AI solutions on their own. This is particularly challenging, given their prior education and the lack of educational institutions in Latvia that provide training specifically related to the management of audiovisual documents. While programming and IT skills can be learned, and archivists can theoretically outline the functionalities needed for AI tools to meet their specific requirements, there is a critical gap in bridging these areas.

   Addressing these challenges, as well as improving competencies, relies heavily on the initiative and enthusiasm of archive staff, who must piece together missing elements from various global sources. Changing the mindset of archive employees, many of whom come from humanities backgrounds, is also a gradual process. However, I believe this process will ultimately be successful. We must keep pace with the times, and this is a shared understanding within the archive.

INA

Interview with Eléonore Alquier (Institut National de l'Audiovisuel, 1975-present) [12]

This interview with Eléonore Alquier discusses the history and current status of the French Institut National de l'Audiovisuel (**INA**), detailing its evolution since the establishment of the archive 50 years ago. The interviewee guides the reader through INA's archiving journey, from the initial preservation of public radio and television content from the Office de Radiodiffusion Télévision Française (ORTF) to present-day documentation efforts in web archiving. Today, the archive houses approximately 30 million hours of content, comprising both professional use and legal deposit collections. The former, accessible via the INA Mediapro platform, contains materials available for commercial use, while the latter is a national collection primarily reserved for academic research. Among the topics touched upon is the digitization process, started in 1999, an initiative that marked a shift from physical to digital formats, resulting in enhanced metadata integration and improved accessibility to audiovisual records. In this migration to digital management, the use of AI emerges as a solution to long-standing archival needs particularly for

the automation of metadata extraction, content analysis, and film restoration. The institution aims to balance algorithmic methods with traditional archival practices, ensuring transparency in AI applications and prioritizing digital literacy among the staff to support responsible use of AI. Overall, the interview reveals INA's commitment to preserving audiovisual heritage while adapting to technological advancements and evolving collection management strategies. Future initiatives will likely focus on continuing to document computational processes and impacts of AI on archival methods - keeping the focus on the traditional archival imperative: tracking provenance and documentation.

Figure 2. Screenshot of INA website featuring INA *éclaire* l'actu, a retrospective section curated daily by the editorial team with a selection of archival material based on current events, and INA's streaming platform *madelen*.

1. What is the history of INA in terms of its institutional and curatorial lineage (provenance, acquisi- tion, preservation)? Can you present the INA collection at its present stage?

   The institution by itself, Institut National de l'Audiovisuel (INA), was created 50 years ago from the disappearance of the original public radio and television broadcaster in France, which was called Office de Radiodiffusion Télévision Française (ORTF). When INA was created, its first mission was to guarantee and ensure the good archiving of public productions, both radio and television programs, so that they could be reused mainly for professional uses by producers, directors and news broadcasters. This still represents the core of the collection at INA, which holds the rights on reuse and sale of audiovisual and sound material from ORTF. The historical archives are available on a B2B website for media professionals, which is called **INA Mediapro**, along with more recent acquisitions, mainly public productions from local television and public radio - namely, France Télévision (FTV) and Radio France. Additionally, INA manages negotiations with private owners of audiovisual content, so that the collection proposed on INA Mediapro is not only

focused or based on public productions, but it also includes private productions - for example, from private broadcasters like TF1 and La Chaîne Info (LCI). As for the news content, we also establish partnerships with sport federations or with production companies in order to propose a larger catalogue with a high diversity of news, sport, cultural content and so on. What I just described is the part of the archive dedicated to professional and private use. INA holds commercial use rights on this entire collection, which today contains 2.5 million hours, either because it directly owns the copyrights or because it has a mandate to represent the original owner of the content. Starting from the 1990s, a second collection was created called 'legal deposit.'[13] It was inspired from the legal requirement to submit books and printed documents at a national library. In the name of legal deposit, today INA manages the continuous archiving of 100 TV channels and 80 radio channels, which represent more than 1.5 million hours each year. As of today, this legal deposit collection counts around 27 million hours. The main difference between these two collections is that INA holds the rights on the archival collection for professional use, whereas it doesn't own any right for the exploitation (including reuse and sale) of the legal deposit collection. This second collection has the status of a national collection. It was founded with the main purpose of managing and preserving audiovisual heritage and it is available exclusively for academic use, which means that only researchers, scholars, students or users with a specific research project have the right to access the collection. In order to consult the archival material for research purposes, users have to come into specific on-site locations. For example, we have a reading room within the French National Library, and computational devices that serve as access points across 50 partner institutions, such as libraries, university libraries, and film institutes. And from these access points, users can consult a total of 27 million hours of radio and television content.

Since 2006, we also created a legal deposit online, because INA is archiving part of the French web dedicated to audiovisual content. As part of this internet archiving project, INA collects not only websites, but also YouTube accounts, Twitter accounts, Tiktok accounts, Twitch accounts, which make the whole collection of INA, including the entirety of these different uses and purposes altogether, around 30 million hours at its present stage. Other collections at INA, less important in volume and less known but still relevant, comprise written documents and photographs. The photographic records mostly come from public television broadcasts, and include photographs taken during the recording of an emission or television program, the so-called 'studio photographs', while the written records come from several sources. The written collection includes documents like reference books regarding the audiovisual issues, technical documentation with broadcaster's comments, descriptions and explanations of how a television program was developed. Furthermore, over the years, INA has collected archives from producers and directors, and, more recently, it has acquired several collections from set and costume designers who worked in broadcast theatre, including figurative records (e.g., costume design sketches) among others.

2. For the archiving of audiovisual material from the web, are you partnering with any organization, such as the Internet Archive or others, that can support this web archiving project at scale?

While not in the context of a formal partnership, we did initially work with the Internet Archive. When INA began archiving audiovisual material from the French web in 2006, the Internet Archive[14] already existed and had collected a number of records of interest for our project. At that time, INA established a collaboration with the Internet Archive to acquire everything that was in our perimeter of archiving for the timeframe between the late 1990s and 2006. At the moment, we don't have many partnerships with other institutions because we decided to develop our own tools for collecting, preserving and archiving audiovisual material - even the format is specific to INA. In this sense, the archival team at INA works independently from outside partners. However, we do value and seek professional exchange with other institutions in the field of audiovisual archiving. For example, we are part of the International Internet Preservation Consortium (IIPC), through which we are connected to other institutions in charge of archiving their national web records, even though currently we have no operational partnerships with them.

3. When was the first digitization, data migration and metadata integration effort initiated at INA? How did INA's digital archival management change over time?

To be accessible online, the records in the INA collections had to be, of course, digitized. Starting from 1999, INA began a huge plan of digitization, known in French as the Plan de Sauvegarde et de Numérisation (PSN), which was committed to safeguarding, preserving and digitizing audiovisual material of national interest. The initial digitization plan focused on the part of the INA archive for professional use, because, at the time, that was the main section of the collection. Before this moment, when we started working in a digitized framework, the act of archiving documents was based on physical carriers for the audiovisual component, and on paper for the metadata part, because this was the type of content that broadcast offices would send to INA. Due to analog television carriers still being the standard in the industry, at the beginning of INA's archival operations, from the 1970s to the 1990s and even into the early 2000s in some cases, the practice of collecting audiovisual material was very much physical. For instance, INA would receive either original documents or printed copies including detailed descriptions of daily broadcasts, which were used to reconstitute and contextualize audiovisual records with information on which show was broadcasted at which moment of the day. At the end of the 1990s and beginning of the 2000s, INA initiated a large-scale retro-engineering effort to migrate all the analog information and documentation acquired up until then into digital formats and build the metadata bases on which we still work today. So the heart of INA's archival documentation and its historical origin comes from analog information - that is to say, analog audiovisual and sound recordings, as well as handwritten or typewritten paper documents.

Today the process is very different because we are dealing with very different volumes of content. The current process of collecting metadata has very much evolved, since we had to industrialize and operationalize metadata extraction and integration. For instance, instead of collecting specific shows selected during the day, within the legal deposit activity we now have a large-scale approach to systematic automated recording, because we don't have the means to watch 180 channels, one by one, 24 hours a day. We also rely on several sources of information to create a general frame of metadata for each record: we have partnerships with both public and private broadcasters, as well as with private actors and players in the audiovisual ecosystem. This is the first step for the organization of records in the collection. Afterwards, according to what information is easily available on the web or in the audiovisual ecosystem, we add more detailed metadata. In addition to this process, we have developed custom digital tools, designed specifically for the news channels, to automatically identify the exact time of broadcasting of different shows from each channel. Finally, all this information is gathered together. So, to some extent, we don't generate the metadata associated with each record, but we rather work as an aggregation platform by merging all these different information sources into a unique model that gives them consistency and ultimately constitute INA's own catalogue. We collect, combine and publish this information in the name of the legal deposit, to which we have the obligation of publishing anything that represents the national bibliography of French television and radio. Beside merging metadata from a variety of external sources, INA also produces its own metadata documentation. Important teams of documentation experts at INA autonomously produce information about a selected part of the collection dedicated to records that are most interesting for professional uses or rich in content, such as news programs. To this regard, from the beginning of the legal deposit in 1995, INA chose to maintain a specific and systematic documentation of the news from general broadcasts, namely the six historical French channels and the evening news. For the purpose of a systematic description of these records, the team of documentation experts at INA visualizes the records, adds a short text describing the content of each part of the news, and indexes the material for each topic, thus generating a continuous documentation of all the news records from historical French channels, a documentation process that has been going on for almost 30 years. For what concerns the collections for professional use, there were two main changes in the last 30 years. One is the systematic digitization of the historical collections, started in the late 1990s and still ongoing. It was thanks to this massive digitization initiative that we could open the B2B website, INA Mediapro, and the B2C website. So when I mentioned that we have 2.5 million hours available as part of this professional collection, they are all

digitized. The other change concerns INA's presence on social media platforms, which we use mainly to share archival records to the public and to show how old archives can reflect on what is happening today in our current society, in addition to other important editorialization activities and publication initiatives.

4. Could you provide a brief overview of how the digital repository system at INA works in synergy with the material archive for tasks like information retrieval and querying of the records?

The large-scale digitization of archival material and the creation of a digital repository have had a huge influence on the development of archival uses, but today I would say that the vast majority of records we collect, at least for professional use, is native digital. However, we sometimes have contact with people who still have analog collections, and this is why it's also important to still have the means to organize the digitization of documents that were natively analog. Even for analog records that were digitized 20 or 30 years ago, it is in some cases necessary to have resources for digitization, especially in the area of film recordings. For example, when it comes to audiovisual content recorded on 35 mm film, the quality of a digitization made 20 years ago is not as good as the quality we can achieve today thanks to more recent technologies. This is why we embrace a politics of remastering to produce, among others, 4K format movies and television programmes, based on the same records originally digitized 20 years ago but relying on different scanning techniques. The difference is that 20 years ago INA underwent a huge volume of digitalization, while today we are taking a more selective approach. This means that we opt for remastering a document only when it belongs to a specific editorial project: for example, when we want to curate a publication on INA's SVOD platform, ***madelen***, for a special program, or a retrospective on a specific producer or director. We can say that the digital processing is based on editorial priorities, especially when it comes to the part of the collection dedicated to professional use.

As to what concerns the legal deposit, there was a major transition in 2001 from a physical collection to an entirely digital repository. Between 1995 and 2001, it truly was a material deposit in the sense that the broadcasters had to make a copy of the documents that belonged to the status of legal deposit. During these 5-6 years, we received physical recordings from which we made copies and DVDs that we proposed in our reading rooms. At the beginning of 2001 we stopped collecting physical content, and opted for digital recordings, initially focusing only on the six historical channels and then developing the digital collection to include 180 channels. It was at this moment that we began not only recording digital copies of audiovisual material but also collecting metadata linked to each record. The integration of metadata in the digital repository had to be operationalized and industrialized. Before moving to digital recordings, INA used to receive the physical material from the broadcasters, along with relative information and documentation. Once we embraced a new system of continuous digital recording of broadcasting material, we had to develop our own tools for collecting metadata. For this reason, we established partnerships with professional firms in the audiovisual ecosystem that manipulate large volumes of metadata. This change from physical to digital in the management of the legal deposit was also the preliminary condition to develop new uses and guarantee broader access to the collection. This plan for the digital recording of audiovisual content at scale led to establishing 50 partnerships across libraries and academic institutions in France, and gave INA the possibility to offer public access to its content outside of its headquarters and on-site reading-room locations in Paris. Once the documents in the collection have been made available online in protected virtual spaces, we were able to develop a wider network of academic partnerships by providing access to INA's archives outside of their physical holdings. The ensemble of these archival decisions, digitization efforts, creation of born-digital recordings and establishment of new virtual access points inevitably has consequences on the management of the digital repository.

As of today, INA's archives have a huge capacity for digital storage. We opted for a hierarchical storage management (HSM) system[15], which is a very interesting option for managing huge volumes because it gives the possibility of scaling up operations, ingesting different types of data and transferring between different generations of data storage technologies. For example, thanks to the HSM system we are able to

automatically change LTO-7 formats to LTO-9 generation, which upgrades the physical storage capacity by three times more: if today we have the ability of storing 60 petabytes of content, in one year, with the same infrastructure, we will have the capacity of 180 petabytes thanks to this change of format. With its ability to record and copy content from older generations to the new one, the HSM system allows us to operationalize large volumes of data and optimize storage space in order to merge all the archiving on the same storage system. So it's an optimization of the way we organize the digital repository, but this isn't the only interesting outcome. The HSM system also allows us to have a continuous process for checking that information is readable, and that there was no destroying effect on data digitized from a tape, for example.

As for the digital interface, we don't have such large volumes of content available on the website open to the broader public, so the traditional navigation for the general audience via a digital catalog is sufficient for information retrieval tasks.

5. How do you deal with having to scale up metadata, along with the audiovisual content? What are the main challenges, solutions and opportunities you encountered during the development of digital and computational projects at INA, with respect to the media archive and the metadata base associated with it?

   Regarding the link between the media and the metadata, they belong to two different managing systems. The large automated storage system that I described before regards the media part, which comprises the largest volumes of content. For context, when we remaster the content in 4K format, one hour corresponds to one terabyte. On the contrary, even if we describe millions or even billions of objects, metadata represents a very small volume from a computational point of view. So the storage systems are very different. Not only are there discrete storage plans for archiving media versus metadata, but we also make independent choices in their management. Much like in other digital archives, the two management systems are then linked via unique identifiers, which connect each media object with its ID, title, description and other information. This allows us to access the information we need by using the identifier to retrieve metadata on a given media object on the technical interface and metadata base, without having to retrieve the media by itself. And similarly, we just need to know the place in the digital repository where the media is located in order to retrieve its information: for example, I can use a web service to open the media object in a player, without needing a direct access to the media itself. This is also a matter of security. As to the challenge and solutions that we found, one issue is that, since media and metadata are stored in different locations, we had to coordinate the archive to connect the media, the data models and the structure of the metadata. To approach this, INA always made the choice of inventing its own models. The issue we have today is that all models are getting quite old. That's why we have been working on a new modular model, which has been conceived to be as flexible and as scalable as possible. We were inspired by other metadata structural models that exist, for example, in the frame of libraries and museums. There is a standard called Functional Requirements for Bibliographic Records (FRBL), which inspired us to imagine a system based on a graph model, meaning a flexible model where we can link different kind of objects together without being limited in the number of characters or the number of fields that we describe. This allows us to add as many descriptions and as many titles as we want. For example, if we want to describe a content in French, English and German, we just have to add a new link to a new language for the same kind of metadata information and it will, altogether, constitute the description of the object. This is what we have been building this year in the frame of a new structure of metadata storing which we have called 'data lake.' This infrastructure will not only be the heart of all the historical metadata that we have created in these past 50 years, but it will also be the starting system for all the metadata created by AI. We hope that the data lake initiative will build consensus between all kinds of information that represent keys to access content and research tools. The idea is to have a unique database including all this information from different sources merged together.

6. Apart from data management systems in libraries, did you take inspiration from any other information system or model to design INA's 'data lake'? For instance, did you adapt any content management indexing or filtering system found on streaming platforms to INA's archival management system?

I wouldn't say that the streaming platforms have influenced us for data storing and preservation, since storage and preservation is everything that you don't see, everything that is maintained in the back-end before access is being granted. If we have been influenced by streaming platforms, it would be in the way we publish the media content. For the front-end design, we replicated some of the editorial choices found in commercial streaming platforms, notably in the way of organizing the screens on the interfaces. For instance, INA's SVOD platform, *madelen*, relies on the archival content management systems for content description, but when it comes to the user experience and navigation, as well as content organizing and selection, it is similar to other streaming services on commercial online platforms. For the interface, we took inspiration from existing interface design and web practices at large, with the aim to make INA's SVOD platform more attractive for the audience and simplify access to content for all users. In this same scope, we even integrated a specific training and documentation program within the archive for the description and publication of audiovisual records for external uses. It goes without saying that describing a document or compiling a database for internal or professional publication requires a distinct set of priorities compared to record description done for circulation among public audiences. It's a similar challenge when we have to publish a Tweet or share a Facebook script. For these public use purposes, we designed a set of practices both for describing records and for linking them to up-to-the-minute news coverage and other cultural productions. In general, I would say that we drew from best practices in online editing and editorialization, which are of course influenced by the 'giants of the web' and main streaming services, among others.

7. Does INA plan on implementing AI methods for the management of its audiovisual collections? If it did already, which AI projects are currently ongoing? And if not, do you foresee integrating any algorithmic processing to the archival work?

   As far as INA's plan for integrating AI applications, we have deployed AI tools mainly for industrializing, operationalizing and automating the process of collecting metadata from audiovisual resources. Especially when it comes to archival material made available for professional use, where we propose 2.5 million records, we have been using AI solutions based on semantic annotations that combine AI-generated descriptions to traditional descriptions made by a document specialist. The very first tests we made with existing AI tools had precisely the goal of making the discovery of news content from the whole collection easier and to enrich the description by performing frame segmentation and object detection on news programs. This helped us produce more precise and exact metadata compared to the ones that we had before, thus improving information retrieval.

   Over the past few years, we have developed and implemented our own AI solutions, which we are quite proud of because they are already in use in our platform and were published on our website a little over a month ago. For instance, we designed an AI tool for data appraisal that produces information about how broadcast media like TV and radio speak about current society. The tool generates descriptors with data about topics, subjects, notable individuals, location and places that are being mentioned in a given news coverage. This allowed us to create a huge legal deposit collection that relies on systematic speech-to-text transcriptions of information from the last five years of news coverage in France and proposes social and cultural indicators. By doing so, we don't need to re-publish the audiovisual content online, but we can directly publish a synthesis of statistics based on what was said on which topic, and by whom. This system based on speech-to-text models is not as precise as we would hope, but at least it allows us to produce data visualizations for a greater audience and provide insight into how television is changing the way it speaks about things. The ultimate goal of this system is to equip users with a very factual approach to interpreting news, a tool that allows the audience to gain objectivity from data and quantitative information - such as, how many times a given news is spoken about, how often TV programs talk about politics, ecology or something else. This was made possible thanks to AI and automated metadata extraction from five years of content. Without AI we wouldn't have been able to transcribe speech to text manually in such a short time and we would not have been able to propose such a systematic statistical approach on so much radio and television material. This is maybe not a new way of using AI, but it does help us produce new

kinds of content, new keys of understanding how radio and television reflect current society. By relying on AI to industrialize part of the process, we have produced not only comprehensive descriptions of INA's archival collections, but also a new use of the archive for the analysis of journalism discourse. For what concerns audiovisual content, we can also deploy AI on the images themselves. For example, we have been approached several times by AI specialists to upscale or restore early film material at a volume that wouldn't be possible through manual restoration. This is still an area of testing: we are exploring AI-based image processing techniques to evaluate whether it can make early television content more easily understandable by new generations through colorization of black and white material, for example. Of course, this raises ethical questions about the authenticity of AI-processed material: we have to be transparent and very clear about how the original material was modified. These AI approaches to metadata extraction, content analysis and film restoration are important for both the back-end archival management as well as for the editorial content organization at the front-end and interface design level. We have been working on AI projects to generate descriptions of news programs for two years now, and the quality of the outputs is quite interesting. We are testing machine learning (ML) models on only two channels out of the 100 TV channels that we collaborate with. So far, we have found that AI tools can be particularly useful to extract metadata, but the images, dictionaries or references from ML training sets often need to be adapted to better fit our focus on audiovisual archives.

8. You mentioned speech- to- text and AI-based image processing for automated restoration or colorization. Do you also use computer vision for things like object detection? Are you doing any close reading analysis of each frame?

   Yes, we have been using automatic segmentation for news channels for two years now. This includes different technologies. For instance, we are implementing computer vision techniques, such as face recognition or logo recognition, to generate metadata on audiovisual records. These applications are useful, for example, to extract information on recurring hosts in programs like weather announcements or breaking news. In these cases, we show to the ML model different images of the same face so that it's able to recognize it and then to make the analogy between the identity of the host and a given show. In this sense, it's similar to a human learning practice, but based on AI for image recognition. These are examples of very specific uses, but is only the beginning of this experience in using ML for archival management. The images, dictionaries or references on which the tool is based to identify faces or logos require constant updates, so it's quite an investment in terms of time allocation and resources, a valuable one but not negligible. Larger-scale AI applications on massive datasets will be the next step. For instance, we plan on implementing a multi-modal approach to complement the semantic annotation, which will be based not only on speech-to-text, but also on some vision recognitions and other computer vision systems. However, I would say that broader uses of AI for archival management at scale are still experimental at this stage.

9. In general, how does INA as an institutional entity position itself in the tension between algorithmic and non-algorithmic archival practices? And how do you, as an archival expert, foresee future evolutions of AI applications for the management or use of audiovisual archives?

   As a non-profit organization and national archive, we don't have a lot of choice in terms of advancing many in-house AI applications. It is not always possible to hire AI specialists as employees, so INA works with a lot of external partners for computational solutions, but we did hire three lead data scientists and AI experts to take upon algorithmic-driven archival work in the frame of computer vision and NLP. The goal of INA is to create a team able to develop internal AI solutions, but also to organize digital literacy and pedagogy initiatives with the rest of archivists working on more traditional, 'non-algorithmic' archival tasks. The role and duty of these new strategic hires in AI is to work in synergy with the rest of the team and explain how AI technologies apply to the archival work, what they involve and what changes they bring. I think it's very important that we have these technical experts internally to make us aware of the consequences, both technical and ethical ones, as well as the legal ones. This is why at INA we try to find a balance between

limited capacity in hiring, and the need to have internal AI experts that can walk the team through this new way of working and archiving.

10. Is there anything you want to add based on your work with audiovisual collections/databases and digitization/AI?

    I don't think I would have been able to predict what happened two years ago, when ChatGPT exploded as a social phenomenon. However, I would say it's important to always remember the core principles of the archival work we do: we are here to keep the traces, we are here to document the change of the records. Digitization was only a first step in the change from the physical carrier to a file, a digital record. In the transition to AI-assisted forms of archiving, archival professionals need to keep in mind that it is maybe not so important to preserve everything, as much as it is to keep the documentation that explains how different archival processes have evolved, how records have become more accessible, and which tools were used. Similarly, AI is important to improve records accessibility and discoverability, but it's also important to document which AI models have been used, which kind of technologies and for which corpora. Which archival processes have been automated? What were the human influences? And so on. As advice to archivists in this transition to algorithmic practices, I would say always document which AI tools you use and how, in order to guarantee transparency and to be able to track down the original document. And even so, what is an original document? Has the very definition of authenticity changed? Even before AI, INA had made some digital restoration. For example, ten years ago we used 100% manual processes, but did not document it. At the time, the audiovisual content was restored due to the physical carriers being chemically corrupted (colors had faded to green or yellow or red and we had to rebalance the colors). In a way, AI doesn't change these fundamental questions about restoration practices, but it does increase their impact at a massive scale, because it gives us the possibility to undergo 100 times more restoration initiatives than we were able to achieve years ago. The archival concerns and priorities remain the same - meaning, always keep track of the provenance of a record and its documentation history. We need to go back to this pedagogical imperative when we publish AI-generated content or metadata to a broader audience, making it aware of the preservation practices and issues, archival devices, and technologies that existed now and then.

NLS

Interview with Mikael Johansson, Anna-Karin Lundgren, F a ton Rekathati, Emma Rende, Chris Haffenden[16] (Swedish National Arc hive of Recorded Sound and Moving Images - National Library of Sweden, 1978-present) [17]

This interview with Mikael Johansson, Anna-Karin Lundgren, Faton Rekathati, Emma Rende and Chris Haffenden outlines the historical development and current status of the National Library of Sweden's (**NLS**) audiovisual collection. Initially governed by legal deposit laws established in 1979, the Swedish National Archive of Recorded Sound and Moving Images (SLBA) organically made its transition to digital in line with technological changes in the mediascape, evolving legal frameworks and market conditions. Over time, it increased its focus on digital formats following a major shift in 2005 when broadcasters began delivering content digitally. By 2015, new regulations included e-legal deposits for web-published materials, further expanding the collection to encompass television and radio programs, published recordings, and donated materials. This insightful conversation with AV-technicians and data specialists speaks for NLS continuous adaptation to new technologies for digitization and preservation,

particularly with regard to curatorial approaches that improve archival accessibility and discoverability of records in the collection via cloud-based solutions and guarantee an efficient storage management system for multimedia records. Similarly to other audiovisual archives in Europe and abroad, NLS is deploying state-of-the-art AI applications and it leverages machine learning models for metadata creation - such as speech-to-text automated transcription models fine-tuned on Swedish audio files. In the conclusion, the interviewees highlight these initiatives as part of a long-term effort to bridge the divide between algorithmic and traditional archival practices, enhancing discoverability and enriching the repository through collaborative initiatives between data scientists and archivists. Challenges remain, such as funding and maintaining high-quality archival standards amid the vast amount of content requiring preservation.

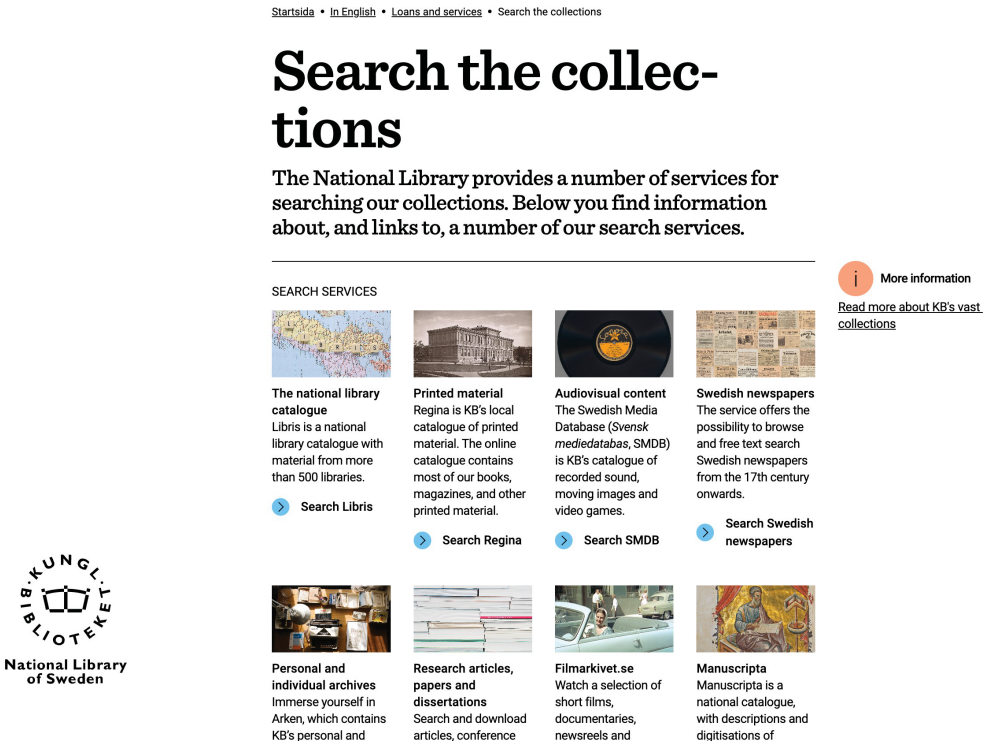

Figure 3. Screenshot of the 'search the collection' feature on the NLS website.

1. Can you present the NLS audiovisual archival collection at its present stage?

   The original NLS archival collection was based primarily on physical carriers, according to the legal deposit law for audiovisual material, introduced in July 1978. From 2005 to 2011, most of the broadcasters switched to delivering digital files instead of physical carriers. From 2015, a law for legal deposit of material only published on the web or electronic resources, called e-legal deposit, supplemented the earlier law. The collection now contains television and radio programs, published sound recordings (such as music), video for rental and selling, theatrical film copied to video. In addition, donated material that enriches the collections to complete the audiovisual media history. The compliance recordings we receive from the broadcasters contain the whole day of broadcasting for each broadcaster and their channels, and we digitize and preserve all the content recorded on a carrier. So it is a broad approach without any single focus. In 2005 we were given extra funding from the government to prioritize the public service broadcasting companies' compliance recordings. Apart from that, for other broadcasters,

we base our decision on whether the content is requested often by the researchers, together with the risk of degradation of the carrier and losing information. In addition to the carrier we also include the capability to perform playback, do we have machines, status of them, enough spare parts and so on and so forth. We have had some specific projects where we have prioritized the digitization of material relating to a certain culture area, like moving images of the Sápmi culture in Sweden, to name one of them.

2. What is the history of NLS audiovisual archives in terms of its institutional and curatorial lineage (provenance, acquisition, preservation)?

   The NLS audiovisual archive collection is based on legal deposit laws that have changed over the years due to changes of the media market; to simplify it is all television and radio broadcasted from Sweden, published sound recordings, video for sale and rental and so on. The provenance is broadcasters, distributors and record companies because it is mandatory for them to provide a copy (for broadcasters a compliance recording) to our institution. We preserve the physical carrier or file and have the right to make copies for research, due to an exception in the Swedish copyright law that allows our institution to do so. For the most part, the researchers are external people such as scholars, students, journalists or genealogists etc. Our collections have around 50 traditional audiovisual formats plus interactive media such as computer games. We do some acquisition for objects that are missing in our collections or broadcast material created before legal deposit law introduction. Since 1979 a governmental body called the Swedish National Archive of Recorded Sound and Moving Images started to receive the material, in Sweden with acronyms ALB and later on SLBA, which also took over the collection of published sound recordings from the National Library of Sweden. In 1994, the institution started to migrate some of the earliest television material to Digital Betacam and later DVCPRO and some published sound recordings to DAT (Digital Audio Tape) some years earlier. In 2009, this institution merged with the National Library of Sweden.

3. When was the first digitization, data migration and metadata integration effort initiated at the NLS audiovisual archives?

   The metadata integration in data form started in the late 80s, where both metadata description for content and the physical carrier are present. Revised around 2005 when a modern GUI for the database was developed. In addition, the database records were revised to conform to the cataloguing rules. Some of the first digitization projects began by 1997 and 2000, digitizing older music and newsreels. From 2006, regular mass digitization and migration to file based storage started. It was the starting point for preserving the content. It also made it more accessible for our researchers by making a web GUI for the database and a playback facility for the files. The files had already been accessible for research by SLBA before as a collaboration from 2006. The first storage migration took place around 2009 for the media files from DLT to LTO tapes. In 2011, an institution handling nonfiction films merged with the National Library of Sweden. It is Swedish private, educational or corporate material on film, donated and made available by the donors for research. It started as a part of the Swedish Film Institute. In 2011, it became part of the National Library of Sweden's organization. We have no close partnerships besides our community relations with other similar organizations in the Nordic countries, as well in the EU and the world.

4. How did the digital archival management of the NLS audiovisual archives change over time?

   From 2006, the archive file (preservation) has been on data tape and access file on disc solution. The latter is what the researchers can access. We use a hierarchical storage management system that has changed from one manufacturer to another. The disc solution has also changed over time. The tape format for the data has been DLT and now we use LTO. We have disc storage as a private cloud for the latest received or digitized files. However, the primary storage for archive files, and backup of access files, are on data tape, and access

files are on a disc solution. Today we have a checksum for every file in our repository. The National Library of Sweden has two internally developed digital repository systems, one of which is OAIS compliant. The majority of the audiovisual files are in the digital repository originally developed at SLBA (which is not OAIS compliant). Within the next few years, the National Library is aiming to develop one new digital repository, replacing the two current ones.

5. Could you provide a brief overview of how the digital repository system at NLS works in synergy with the material archive for tasks like information retrieval and querying of the records?

   Via a web GUI of our database, the Swedish Media Database (**SMDB**), our researchers can access by logging on at the facility in Stockholm. In the database, the archive file is described and linked to the content, and a backbone look up system recognizes the name difference (the access file has another prefix than the archive file), and finds the correct file for playback from the disc solution. We also provide access by sending physical media, DVD, to the libraries we collaborate with in Sweden, Norway and Denmark. During the pandemic, we tried distant access through a cloud solution for researchers, still in use with several of the collaborating libraries. We are investigating solutions to securely stream the material from the NLS archives, through restricted researcher access and prohibiting download or other unauthorized use of the material. Our intention is to adopt a streaming solution that complies with current laws, regulations, policies and that takes into account matters like security, copyright and so on. As a governmental body, the solution must be in-house and closely connected to our current systems for retrieving material. Researchers visiting our premises in Stockholm can access their ordered digital material right away through our web catalogue, and our aim is that researchers in other parts of Sweden could use this flexible manner as well. We do also get requests for archived materials from broadcasters or production companies that have cleared the copyrights, which is necessary when it is not for research. We have system solutions to read the archive files from storage to disc and secure transfer to those who requested the material.

6. What are the main challenges and opportunities you encountered during the development of digital and/or computational projects at the NLS audiovisual archives?

   The main opportunities are better overview and being able to make a management system for digital preservation. Also, getting a system for researcher access to the content, faster than in the analog era. As in many projects the challenge is to get a common language between different groups like audiovisual technicians, metadata personnel, developers and IT management. It takes time but is also an opportunity. Funding is also a challenge – finding resources for things like staff, digitization projects, large data storage, and IT development to make changes when there are missing functions for accessing the content or for preservation. The legal deposit law for audiovisual material does not give us the right to dictate delivery format or quality parameters, for neither analog nor digital materials. That and the amount of millions of hours of audiovisual content to preserve is challenging.

7. Does the NLS plan on implementing AI methods for the management of its audiovisual collection? If it did already, which AI projects are currently ongoing? And if not, do you foresee integrating any algorithmic processing to the archival work?

   KBLab, the NLS's datalab, is currently working on fine-tuning a Whisper model for the Swedish language using the library's audio-visual collections as training data, among other sources of high quality, carefully curated sound data for Swedish. This model will then complement our existing work with speech-to-text models that can be used to improve the accessibility and discoverability of the NLS's audio-visual archival holdings. We have produced a demo of a search application for the library's radio archive, which shows how the creation of transcriptions using AI models makes the radio material amenable to both free text search and metadata

enrichment for named entities that then also become searchable (i.e. names, places, organizations etc). The next step will be producing transcriptions for the radio archive in its entirety in order to radically improve the discoverability of the material. Due to copyright restrictions we will not be able to make this publicly available, but the plan is to have a dedicated zone at the library's premises to allow researchers to use this. This example illustrates a positive symbiosis between i) using the library's collections to improve the AI infrastructure for a smaller language like Swedish and ii) using these improved AI models to enhance access to the library's collections[18].

8. In general, how does NLS as an institutional entity position itself in the tension between algorithmic and non-algorithmic archival practices? And how do you, as an archival experts, foresee future evolutions of AI applications for the management or use of audiovisual archives?

   In terms of the relationship between algorithmic and non-algorithmic archiving practices, we are interested in exploring areas where machine learning can help us better understand collections that currently lack metadata – or require extensive human efforts to produce descriptions. An example of the latter is a pilot project at KBLab that looks at using LLMs to produce summaries of various TV broadcasts where we lack incoming metadata, requiring librarians to watch the broadcasts in real time to produce descriptions – a time-consuming and somewhat thankless task. A similar instance has involved making digitalized postcards lacking metadata amenable to the sort of online search techniques we're used to from our smartphones[19].

   Both of these projects have grown out of a new initiative at the NLS called KBx, which seeks to combine the insights about large-scale discoverability from the data scientists at KBLab with the domain expertise of library specialists about specific archival holdings, in identifying where AI might be useful for the library's working practices. In this sense, we are interested in exploring synergy, overlap and collaboration between AI/ data science experts and library/archival experts, rather than cementing the potential divide between these perspectives.

9. Is there anything you want to add based on your work with audiovisual collections/databases and digitization/AI ?

   We try to incorporate the metadata that is available for objects in different collections as search filters. However, for audiovisual collections this metadata can be scarce, and typically only includes the media channel, the date and time of broadcast, as well as possibly some additional listings information about the item, such as brief description of topic, names of presenters etc. As mentioned above, it is this fundamental scarcity of metadata that has prompted us to explore AI-based methods as an alternative to enable content-based search.

ORF

Interview with Christoph Bauer and Mag. Herbert Hayduck (Austrian Broadcasting Corporation Archives, 1986-present) [20]

This interview with Christoph Bauer and Mag. Herbert Hayduck provides an in-depth look into the history, current development, and future of the **Austrian Broadcasting Corporation**'s audiovisual collection (**ORF Archives**). The conversation focuses on digitization efforts, the role of AI in archival management, and the evolving relationship

between technology and archival practices. A detailed description of ORF plans for metadata extraction and integration, from pre-digital to digital operations, allows us to gain insight into the valuable, and yet often invisible, archival work that hides behind archival interfaces, whether in their digital repository or in other online platforms. This interview also demonstrates that, internal control over the management of IP rights can improve assistance to archival users in the form of broader access to media content and relative metadata, resulting in an overall more effective information retrieval experience that prioritizes users' needs. Finally, from this interview ORF emerges as a leading player in balancing traditional archival practices with AI methods. Their approach to emerging technologies relies on internal initiatives for improving digital pedagogy and computational expertise across the entire archival team, which strengthen a responsible and informed digital record management workflow. As the interviewed speakers outline, this internal pipeline has evident positive outcomes on the ultimate goal of enhancing access, discoverability and reuse of audiovisual records for both educational and commercial purposes.

Figure 4. View through multiple sections of ORF empty vaults, showing some of the last cassettes to be migrated at the end of the shelves. © Photo by Christoph Bauer / ORF.

1. What is the history of the ORF Archives in terms of their institutional and curatorial lineage (provenance, acquisition, preservation)? Can you present the ORF collection at its present stage?

   Overall, the ORF audiovisual collection consists largely of material that was created directly as part of in-house productions. In addition to this, we also have content that was purchased for use in ORF programs or that was supplied to us by agencies - e.g., EBU, Reuters, etc. In order to make the work of our colleagues in production easier, large parts of this external content were retained and digitized or migrated. As of 2024, our archive contains content from the beginnings of television broadcasting in Austria to the present days. It includes a very broad range of audiovisual content, from news-programs to TV-shows, from political magazines to science documentaries, from international edutainment-coproductions to national and regional raw-footage, from sport-events to recordings from cultural highlights and rare gems. The focus is on Austrian- and 'ORF'-content – therefore 'Hollywood' and alike play a minor role in our collection, since those productions are not in our scope of 'preservation-duty'. In addition to the final productions, programs and contributions, we also focus on preserving raw physical materials. This gives our customers (mainly in-house editorial teams) direct access to the originals. At its present stage, our collection is already 95% digitized and it is therefore almost completely available to our customers in digital production file formats. This enables us to supply our production units extremely fast and conveniently.

2. When was the first digitization, data migration and metadata integration effort initiated at the ORF Archives?

   With regard to data migration, a distinction must be made between media content and metadata. Metadata was 'migrated' as early as the mid-1960s. At that time, the switch was made from 'stock lists' to index cards. The first major switch to digital databases took place in the mid-1980s. Following this first major shift, approximately 15 years ago these digital databases were transferred to a system co-developed by ORF together with the broadcasting companies of the ARD (full name: Arbeitsgemeinschaft der öffentlich-rechtlichen Rundfunkanstalten der Bundesrepublik Deutschland), the consortium of regional public-service broadcasters in Germany (ORF has a good experience in cooperation with other broadcasters, not only in (TV and radio) production, but also in research & development of AV-related soft- and hardware (e.g. our main TV-Archive-database, data-mining-tools, preservation-related tools, etc.).

   As to what concerns media content, the audiovisual material was repeatedly transferred to the latest carriers and formats that were used in the company's operations/production at each given time. Some content that was produced on film has thus been through a true odyssey across the history of AV recording: from film to 2-inch VTR, to 1-inch VTR to MII, IMX and so on and so forth. Around 15 years ago, the ORF began to digitize content in specific areas. In 2016, our largest migration project (DiMi) finally started, in which 300,000 hours were migrated from IMX and Digi-Beta to files (MXF). This last project will be completed in about 5 months by the end of 2025. In addition, various special parts of our collection (e.g. film; about 20,000 hours) were directly digitized.

3. How did the digital archival management of the ORF archival collection change over time?

   One of the decisive steps here was the already mentioned migration of our old database (FARAO and predecessors) to the current FESAD system (in collaboration with ARD). For the first time, preview media (preview and keyframes) could be offered across the board, allowing our customers to access the content directly during research. Another related in-house development is our 'customer access system', known as mARCo, which not only makes research much easier for our customers via a simplified and cross-data source research system, but also enables direct access (including downloading the content in production quality). However, this is only possible because we also have a very advanced rights management system, which is indispensable for this purpose.

4. Could you provide a brief overview of how the digital repository system works in synergy with the material archive for tasks like information retrieval and querying of the records?

   Because over 95% of the content and materials are already available in file formats, access is quite simple for users: via the mARCo system (see above), the customer can conveniently research the content needed on their device (PC, mobile device) and also usually initiate the transfer of the parts they have selected to specific destinations. If it is necessary for legal reasons, they can also retrieve rights clarification or other technical information they may need directly from the system. The selected content is then made accessible on their preferred production storage within a few minutes. In special cases, experts in the archive are also available to support special research and even provide a complete service - namely, research support, rights clarification and editing. If the material is only accessible on an analog physical medium, the archive will arrange for it to be digitized upon request.

5. What are the main challenges and opportunities you encountered during the development of digital and/or computational projects at the ORF Archives?

   One of the biggest challenges in the beginning was finding a common language between experts on the computational side and experts on the archival side. By introducing 'archivists with computational

skills' in the team and thanks to the dedication of everyone involved, with a lot of shared enthusiasm and willingness to learn the needs of both sides involved, we managed to ace this quite fast. This synergy between professionals in computer science and archival sciences provided an excellent basis for all projects and developments ahead. Along with this collaboration, the greatest opportunity has been the extremely easy access to information and content made possible by digitization. Thanks to these efforts, 'our' archival collection represents a large, important and therefore indispensable part of ORF current television production.

6. Does ORF plan on implementing AI methods for the management of its archival collections? If it did already, which AI projects are currently ongoing? And if not, do you foresee integrating any algorithmic processing to the archival work?

   AI has been at least 'under observation' for over 25 years. We have been increasingly involved with the topic for around 10 years now. Speech2Text is currently being used for large parts of our collection and many new acquisitions, favoring automated textual transcriptions from sound material. Thanks to this technology, very old parts of the ORF collection have been provided with descriptive metadata for the first time. Text analysis methods (NER, summarization, etc.), research-supporting analysis methods (query by image, etc.) and image content analysis technologies are currently being evaluated or are in a test phase, since the use of LLMs is changing the game once more.

7. Other than implementing automated speech-to-text transcription, as well as testing NLP techniques and LLMs to facilitate the archival work in the back-end (e.g., metadata creation), did you adapt any content indexing and filtering system commonly found on streaming platforms to ORF archival management system?

   Streaming platforms played a minor role in our decisions, since for a long time the legal situation had prevented ORF to run a dedicated streaming platform and only allowed us to focus on traditional broadcasting (the usual 7-day-catch-up-TV, etc. …), while the 'internet-channel' remained forbidden up until recent years. For this reason, so far we have been delivering final products (mainly fiction) to other related platforms (e.g. 'Fidelio'). Nevertheless, we always hoped that the legal background would change. In the meantime, we introduced and opened our content to academic and school platforms, in order to gain insight and experience in online content distribution. This forward-thinking approach, along with our general digitization efforts, proved to be right when the new 'ORF-law', adopted in summer 2023 and in place since the beginning of 2024, allowed us to run dedicated streaming services. However, since we believe in providing carefully curated services, a general and open access platform to our complete collection is not foreseen in the near future. For now, we will offer user-oriented curation via geo-references, and other similar approaches. But who knows what the future will bring…

8. In general, how does ORF as an institutional entity position itself in the tension between algorithmic and non-algorithmic archival practices? And how do you, as an archival experts, foresee future evolutions of AI applications for the management or use of audiovisual archives?

   We try to approach the matter as unprejudiced as possible. We rely on an approach based on evidence, meaning that if the quality of the algorithmic processes is suitable for the respective purpose, there is nothing to speak against using them. However, it turns out that many of the technologies currently available also involve a great deal of effort in the area of operation planning and sometimes even specific computational model training or fine-tuning. These efforts are not always in proportion to the results achieved. It is therefore necessary to refrain from making blanket judgments and decisions based on them and instead proceed detailedly. Experience has shown us that there are already algorithmic tools that can be used profitably and in a resource-saving manner.

Another greatly debated topic in AI in application to archives is the automated labelling of the metadata generated by algorithmic processes. On the one hand, this can give the customers (e.g., researchers or students) the opportunity to decide for themselves whether they want to use this data as a basis for their work; on the other hand, this promotes a certain 'ghettoization' in the data area and forces one to always offer at least a certain basic set of non-algorithmically generated data. Another important aspect or argument to keep in mind with respect for using AI-generated metadata to label content is the fact that, according to some experts in the field of AI, deploying these metadata for further training or development of AI-models might lead to major problems regarding the quality of the final ML models. At the same time, most of the data emerging from the last decade is already influenced by AI: for example, this very text has been checked and corrected by an AI-based spell-checker – and some parts have been translated from German to English by an AI-based tool. So how to keep track of all these AI-influence on digital content and metadata creation? Certainly, this topic won't be off the table quickly and it might take years before we answer this question.

9. Is there anything you want to add based on your work with audiovisual archives and digitization/AI?

When it comes to preserving the audiovisual heritage and providing access, digitization is a big step forward in achieving these goals. Similarly, AI as well can be really helpful here when used wisely. To this point, we are currently running a national project with the University of Vienna and try to deal with the many and currently heavily changing legal aspects regarding the reuse of our data (both content and metadata) for training AI-models. The EU has an important role here in providing a legal framework, but this won't be enough. Our main customers are (still) the ORF journalists. Although our content is used very often and widely outside the company as well, no large-scale (re-)use of (parts of) our collection outside of ORF-based initiatives is currently ongoing or planned. Nevertheless our content can be found in important collections like the Europeana-collection or in/at other local and international exhibitions - but here the content is hand-picked and small-scaled. In the long term, we hope that digitization and AI initiatives will allow us to make large amounts of hidden, valuable audiovisual content discoverable and usable as a result of more efficient digital and computational operations. The ORF Multimedia Archives are on top of all these developments, providing excellent service to our customers and will continue to have a pro-active role in the future of archiving.

GHB | AAPB

Interview with Karen Cariani (GBH Archives, 1979-present and American Archive of Public Broadcasting, 2013-present) [21]

This interview with Karen Cariani discusses how US public broadcasting archives are dealing with archival management, preservation and access when it comes to handling digital records. The conversation primarily focuses on two archival initiatives: the American Archive of Public Broadcasting (**AAPB**[22]) and the **GBH**[23] Archives. On the one hand, the AAPB is a collaboration between GBH and the Library of Congress originally launched in 2013 by the Corporation for Public Broadcasting (CPB). It aims at preserving and making accessible historical records from US public media collections via digital formats. Over the years, the AAPB has grown from 40,000 to over 200,000 items, thanks to widespread digitization efforts across the country supported by a series of grants. On the other hand, the GBH Archives dates back to the 1950s and has also undergone significant digitization. Guiding us through these two archival projects, Karen Cariani highlights challenges and opportunities

in archiving both analog and digital records in a rich conversation on IP rights, funding constraints, metadata creation, and the use of AI tools for improving content organization and accessibility. The interview also touches on the importance of preserving physical materials, the potential of sharing datasets for research, and the need for sustainable digital archiving models.

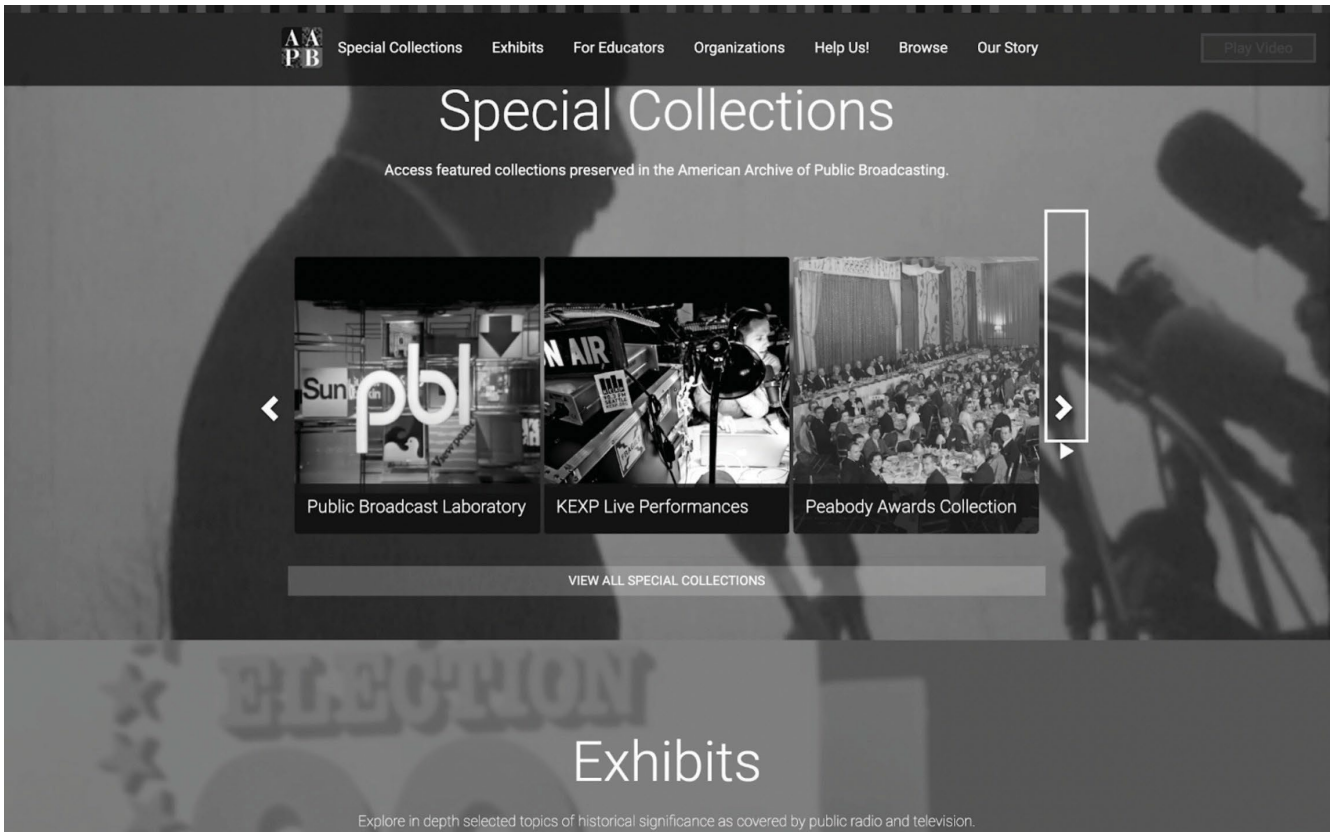

Figure 5. Screenshot of the AAPB website showing a snippet of their special collections and curated exhibits.

1. Can you provide a brief presentation of the AAPB, its history and core mission for archival management, preservation and access?

   The American Archive of Public Broadcasting (AAPB) is a collaboration between GBH and the Library of Congress in Boston. Its mission is to preserve public media, both television and radio, across the country, and make it as accessible as we can on an online platform - namely, the **AAPB website**. As to the division of responsibility between the two organizations, the Library of Congress is responsible for the long term preservation of the digital files, and the GBH is responsible for community engagement and outreach to public media. Together, we manage the website, which is the access point to the digital files, curation decisions, and legal review. Currently, we have almost 200,000 items in the collection including both digitized audiovisual files and sound recordings.

2. What is the history of the AAPB in terms of its institutional and curatorial lineage (provenance, acquisition, preservation)? When were the first digitization, data migration and/or metadata integration efforts initiated at the AAPB and how did the digital archival management of its AAPB audiovisual collection change over time?

Originally, the project was part of an initiative launched by the Corporation for Public Broadcasting (CPB). CPB started a series of pilot programs for the initial digitization of audiovisual and sound material from local television and radio stations, and about 11 years ago (2013), then they turned over the initiative to GBH and the Library of Congress. The initial pilot digitization project was a 2-to-3 year project. Roughly speaking, it managed to digitize about 40,000 items from audiovisual archives across the US. Since then, the digitized collection increased to 200,000 records coming from stations who had received various grants to digitize their materials and already had digital files. Alongside the acquisition of digitized records, the AAPB also helped a number of organizations write grants to manage the digitization process. Among other funding resources, 2 years ago, we received a large Mellon Foundation grant to support more digitization of broadcasting materials across the country, which we are managing as part of an ongoing project.

3. Could you provide a brief overview of how the digital repository system at the AAPB works in synergy (or not) with the physical archives for tasks like information retrieval and querying of the records? For instance, how do you handle long-term sustainability and access to digital records? What is the role of analog records in enhancing long-term preservation?

   The AAPB is purely a digital collection. We're not directly in charge of nor do we own any of the physical materials. Those all go back to the stations, who manage their physical collections in line with their own corporate business and preservation strategies. The entire AAPB collection, along with the metadata, is available on site at GBH and the Library of Congress. Users can view over 120,000 items in the Online Reading Room, streaming only on the website from anywhere in the US. Of the 200,000 items in the digital repository, only 120,000 are available for online streaming through the AAPB website due to IP rights issues. For example, we have the entire *Sesame Street* collection, but due to the various IP rights around puppets and actors, we can't make it available online. Apart from copyright issues that affect access to digital archives, long-term sustainability inevitably relies on available funding. The Mellon Foundation has been extraordinarily generous and provided a unique opportunity to up our staff, expand the digitization project and improve the website. There is always going to be a sustainability issue for both digital and analog collections alike long term. When we talk to local stations and they ask us 'what should I do with these analog tapes now that it is digitized?', we advise that they keep the analog tape for as long as they can, because you never know what will happen to the digital files. We are privileged to have the Library of Congress as a collaborator who will manage the digital preservation and migration into the future. But good archival practice says that we should keep the analog formats. A number of the stations just don't have space anymore or they don't have the equipment to play the tape, so they are trashing them. This is happening and it's a real concern, but given the constrained resources, understandable.

4. This is indeed an urgent issue in archival management across news archives: due to lack of space and high storage costs, many television, radio and journalism archives are being 'dismissed'. To limit this practice, some collections have been acquired by university libraries in rescuing operations. In your experience with broadcast archives, are media companies and television networks entrusting their collections to academic libraries and archives for long-term storage, digitization and access?

   Some audiovisual archives like the Northeast Historic Film are maintaining the WCVB collection. They do keep and preserve the physical materials, and they have received some grants to digitize part of the collection. The Boston Public Library (BPL) also took on the WHDH collection - including, film and broadcasting material - a long time ago, and they still maintain it today. A few years ago, the Boston TV News Digital Library, an initiative funded by the Institute of Museum and Library Services (IMLS) and the Council on Library and Information Resources (CLIR), created an online catalog of broadcasting material from the WHDH collection (1960-mid-1970s), Cambridge Community Television (1988-1999), Northeast Historic Film's WCVB collection (1970-1979), WGBH-produced *The Reporters* (1970-1973), *Evening Compass* (1973-1975) and the *Ten O'Clock News*. The digital repository counts almost 50,000 records, brought together thanks to a collaboration between the BPL, Cambridge Community television, Northeast Historic Film and WGBH Educational Foundation. Other institutions, such as the West Virginia State Archive, do take on analog materials via local initiatives and commit

to collecting resources for long-term physical and digital preservation. However, I think it's becoming less and less likely. Some universities around the US are openly taking older formats, but I think they're becoming more wary about it, mostly in terms of copyrights. Most academic institutions are open to rescue physical audiovisual materials, but there are still challenges: what can they do with it? Can they make it available to the university community? Can researchers, professors and students access it? Can they put it online? Universities in California, like UC Santa Barbara and UCLA, have preserved and curated audiovisual collections from early television. Indiana University Libraries holds a large audiovisual collection. Washington University Libraries in St. Louis took on the Henry Hampton collection, including all film materials for *Eyes on the prize*. They've been collecting various other things too. To answer your questions, these 'rescuing operations' are mostly in the hands of universities around the US that are willing to take upon large physical collections in their storage.

5. You are also the Executive Director at the GBH Archives, could you give us an overview on the GBH archival collection and its digital management, access and long-term sustainability plan?

   GBH has been on the air since 1951 for radio, 1955 for television, and we are committed to managing the digital preservation of and the access to all files housed at its archive. We have a physical collection of material dating as far back as the 1950s. Since the goal is to preserve anything GBH creates or distributes, the GBH Archives includes all types of radio and television formats, genres and broadcast material. Some titles in the archive are *Masterpiece Theater*, *Nova* and *Frontline* and *American Experience*. A couple of years ago, we received a generous NEH Challenge Grant to digitize about a third of the collection, 85,000 items, a project that we are about to finalize. Once the digitization is complete, we will make available some of the digitized material online through our open vault website, and contribute it to the AAPB at some point. GBH is also starting to utilize old television programs for redistribution. For example, some programs, like *Victory Garden*, are now available on YouTube and we are looking into redistributing other titles from GBH digitized collection. In this sense, the GBH Archives is an active archive that is constantly being tapped and used. As far as the analog and digital long-term preservation plans, we are keeping the physical materials for as long as we can. We might move some of them out of the temperature and environmentally controlled vault into a warehouse storage, just because we've got numerous digital copies, and we feel like the digital files are pretty safe at this point. The sustainability constraints for physical tapes really do come down to space. There is a cost involved with maintaining a digital collection, and there is a cost involved with maintaining a physical collection. And when you've got a resource-tight company like public media, something's got to give, right? People are going to start saying, 'Why are we keeping both?'. Both digital and analog records are both costly to keep. In some cases, I would argue that the digital is actually more expensive to keep than analog material because of long term migration needs and software upkeep to manage the digital files, but upkeep of obsolete machines is also expensive. As of now, for the AAPB are using two external vendors, one for the digitization, and one for temporary digital storage. The digitization vendor is George Blood LP. He has the capacity to process large amounts of material simultaneously, as well as the ability to repair machines, because he has such a big bank of them and has backups. For us, it is much more cost and time effective than having to process the digitization internally. In house digitization would mean having to hire one engineer, one deck, and it would take years and a lot of resources to get through 1000s of items. For the AAPB, we're using Digital Bedrock for temporary digital storage and preservation for the LTO tapes, because the Library of Congress can't ingest as fast as we are creating the digital files. This is a stop gap temporary solution. Long term, the Library of Congress is responsible for making sure they have resources to continue to migrate and store the digital files in the foreseeable future.

6. What are the main challenges and opportunities you encountered during the development of archival projects at the AAPB or GBH? Not only in terms of analog versus digital content, but also in terms of cataloguing different types of records for public access - like distinct broadcasting formats, forms or genres.

When it comes to the physical collection, one of the most obvious stumbling blocks is material preservation. For instance, in the early 1960s a fire destroyed a lot of the tapes stored in the GBH Archives. Moreover, at the time of early television broadcasting tape was so expensive that older tapes were often used to record over stories. Despite these challenging circumstances, the GBH Archives were able to preserve programs like the original recordings from the *Julia Child* show, which presents a unique opportunity to provide access to early television programming. In the case of radio programs, cataloging problems arise from the lack of visual material and contextual information associated with audio files. It's just radio and sound recordings that need to be transcribed, without much other information, which is indeed challenging. As far as archiving what we currently produce, it is partly dependent upon the production units delivering the material to the archive. The material doesn't always get to us, but we still try to archive as much television content as we can - from *Antiques Roadshow*, to GBH children's programming, to local shows. The AAPB is pretty similar. Its collection ranges from music, some jazz and radio shows, news programming, local cooking shows, children's programming. For the AAPB as much as for GBH, we are prioritizing the archive of local content, news, and of other content created by local stations. We are particularly invested in preserving news programs and local news for public access for two reasons. On the one hand, local stations retain the rights on local television programming, which makes it easier to acquire the rights to make local news more accessible to the broader public. On the other hand, we focus on local news records because they are more at risk, as opposed to national news records. Nation-wide broadcasters and networks probably already have copies of anything that they aired and they have access to well-funded archives. We try to focus on local audiovisual collections that might disappear without external help and provide them with the opportunity to digitally preserve records that would otherwise be lost. When it comes to the digital collection, metadata is the real challenge, because even archival professionals and metadata specialists often do not have much access to information about older radio and television programs. To make up for this, we are using speech-to-text transcripts generated by ML models that provide us with at least some sense of what's in the program. Of course, these transcripts are not entirely accurate, there can be mistakes in the model outputs, but we are still able to extract some keywords that can be considered as cataloging labels and added as metadata in the digital repository. This does help improve the keyword search, if nothing else, and it increases discoverability of content. Otherwise, besides the transcripts and whatever the station gives us, the metadata is pretty thin in a lot of places. As I mentioned before, radio shows often do not have very good metadata, because of the lack of visual material. We are creating transcripts of radio programs too, but to achieve a good metadata integration process for the entire archive is going to be a long project.

7. Speaking of the opportunities opened up by computational tools (and their shortcomings), do AAPB and GBH plan on implementing additional AI methods for the management of their audiovisual collections? If they did already, which AI projects are currently ongoing? And if not, do you foresee integrating any algorithmic processing to the archival work?

   We're working on using some AI tools and natural language processing (NLP) tools to capture single frames in audiovisual records that might contain metadata, like the slate at the beginning of a show. Thanks to tools for optical character recognition (OCR), speech-to-text transcription, speech identification, we are working on being able to extract the actual title, the air date, the producer, the copyright holder for many of the records we have in the digital archive both at the AAPB and GBH. We are attempting to implement ML to capture para-textual data and other information contained in the credits, which is then added to our digital records as metadata, but it is still a work in progress. We are trying to extract all the data that appears within the program as text off of the videos. The main problem with the AAPB and the GBH collection is that they are so diverse, so varied and come from so many different sources that there isn't a standard textual format in use across records, or a standard font or even a consistent page layout, meaning a place on the screen where text appear. And, as you know, AI tools need consistency among records in order to be trained and fine-tuned to produce more accurate outputs. So we've been struggling with this issue quite a bit. For this AI initiative, we have partnered with Brandeis University's Research Lab for Linguistics and Computation on a set of open-source tools and workflows known as CLAMS (Computational Linguistics Applications for Multimedia Services). Professor James Pustejovsky and his team are helping us develop and train tools for metadata extraction. So far, we've been using Whisper for speech to text

transcripts. We know that the bigger Whisper models are probably more accurate than the smaller models, so we've been testing those. Unfortunately, we don't have the funds for large compute power as of now.

8. In general, how does the AAPB/GBH as an institutional entity position itself in the tension between algorithmic and non-algorithmic archival practices? And how do you, as an archival expert, foresee future evolutions of AI applications for the management or use of audiovisual archives?

   I think the AAPB and GBH are probably more open to implementing AI applications compared to other archives, because we are collaborating with experts in the field of AI that understand its mechanisms, applications and potential shortcomings. A lot of the smaller archives don't have the funding and the resources to figure out how to implement AI in their workflows or how to manage AI tools in the long run. That said, there's a lot of commercial systems out there that utilize the tools and promise to be able to extract it. Being commercial systems, however, they are costly and have some constraints. Personally, I am reluctant to pay a vendor for a tool that is likely to change in three years and does not necessarily do exactly what we need at the archival level. Commercial AI companies make a lot of promises and ask a lot of money, but they do not respond to archival needs, so they don't seem like the best option forward for archival management. In addition, they are often using your content to train tools that they will then sell. For these reasons, at the moment, we are refraining from using commercial AI tools and we are not yet using automated content filtering as seen on commercial streaming platforms. We have had internal conversations among the archival team about ranking items in search results based on 'things like this' metadata weighing in order to facilitate content discoverability, but this has not been implemented so far.

9. Is there anything you want to add based on your work with audiovisual collections/datasets and digitization?

   One of our desires is to make the material available for public use and this implies a struggle with improving metadata integration via AI as much as with addressing IP rights for open access. Everybody wants to have the records on their site, and they want a dataset that they can take away. This is something that would be wonderful to try to break through in terms of either figuring out how we can enhance that (better metadata) and support that (long-term open access). In future evolutions of current projects at the AAPB and GBH, creating datasets is a really good way not only to give back to the community, but also to improve AI tools and their applications in archival settings. There is a lot of analysis and research that could happen across the AAPB collection as a whole, as a dataset. For this to happen, the IP rights issues really need to be figured out. There is 60 years worth of audiovisual material that's just not available because of IP rights, and it's going to just turn to dust on a shelf. And it's our history, right? It's our visual history, it's our sound history, and it's important to make that as available as any other document captured in live events on contemporary media over the internet.

WVLH Project

Interview with Montse Bailac, Isabel Beirigo, Natalia Cetera, Maria Drabczyk, Georgia Evans, Christine Hveding, Ivan Milovanov, Helena Vicente Galán, Dorota Wiśniewska (Watching Videos Like a Historian, 2022-present) [24]

The project ***Watching Videos Like a Historian*** (WVLH), funded by ERASMUS+, aims to enhance media literacy among young audiences by integrating European audiovisual heritage into local history and citizenship education. Launched in April 2022 and concluding in March 2025, the project has produced key resources, such as a field report and a comprehensive toolkit for educators that includes teaching strategies, e-learning activities, audiovisual galleries,

and classroom guides based on extensive feedback from teachers. In collaboration with various partners - namely EuroClio and Europeana, Centrum Cyfrowe, Webtic, EUscreen/Netherlands Institute for Sound and Vision, and 3Cat -, the project curates thematic audiovisual galleries from European archives. While not employing AI tools directly, the initiative emphasizes critical engagement with media and audiovisual data, addressing the role of AI in education and media literacy.

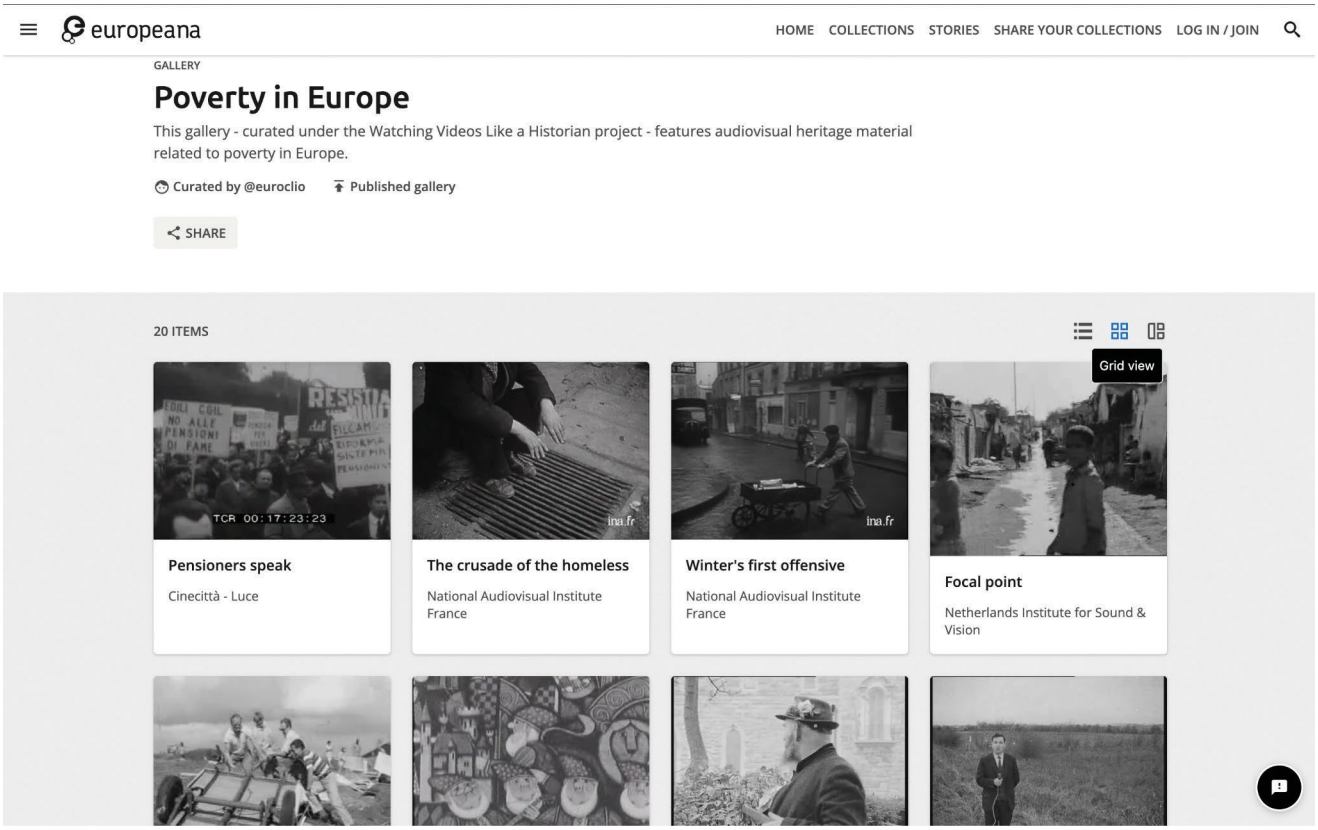

Figure 6. Screenshot of audiovisual galleries on the topic of 'Poverty in Europe', curated by EuroClio as part of the project Watching Videos Like a Historian on Europeana.

1. Can you present the project *Watching Videos Like a Historian*, by providing an outline of its history from its initial design and development to its launch, goals and layout at the current stage?

   Watching Videos Like a Historian is an ERASMUS+ funded project that, at its heart, seeks to increase the number of well-informed, media-literate young citizens across Europe. To do this, the project is developing materials that support the teaching of history and citizenship through audiovisual heritage. The project was born out of a consciousness that media literacy is now considered a crucial skill for citizens to critically navigate information sources (online, on social media or through other channels). The classroom is a key place to start supporting young people (as current and future citizens) to be critical and conscious consumers of media, but this can only be achieved if educators themselves have access to relevant sources, tools and information to help students. *Watching Videos Like a Historian* seeks to develop resources which draw on the wealth of audiovisual heritage materials available in archives across Europe to meet this need. The project kicked off in April 2022, and ends in March 2025, so we are now entering the final sprint! But we are already very pleased by what the project has been able to deliver. Among many activities, the project has published a **State of the Field Report on Media Literacy supported by AV collections** (an analysis of how audiovisual materials are used for education, the challenges teachers and students face in using them, and how this can be addressed),

and updated the **WATCH Media Literacy Competence Framework**. We have developed and published (in Dutch and English) a **Practitioner's Guide for AV Collection Holders**, a practical tool to help archives and broadcasters with audiovisual collections turn their cultural heritage assets into open educational resources. We have run a series of capacity-building events to support educators and get feedback on the resources the project is developing. And, most notably, we have published the first version of the **Watching Videos Like a Historian Toolkit**, which aims to provide all the resources that an educator would need to incorporate audiovisual materials into their classroom to teach history and media literacy - and motivation to do so. This has been accompanied by a **webinar series** of three episodes, aimed at sharing the Toolkit with educators and providing them with hands-on activities to effectively integrate audiovisual materials into the teaching process. In the final three months of the project, we look forward to receiving feedback from teachers on the Toolkit to refine and further develop it, and wrapping up all our project activities. **Find out how you can give feedback on the Toolkit**.

2. Who are the core team members and outside collaborators of the initiative *Watching Videos Like a Historian*? Which archives are part of the project, and how do you work in synergy with them?

   The project is led by **EuroClio (the European Association of History Educators)**, an organization committed to supporting the development of responsible and innovative history, citizenship, and heritage education. Five other partners complete the project consortium: **Europeana Foundation**, which deploys the **common European data space for cultural heritage** and provides access to millions of cultural heritage objects through **Europeana.eu**; **Centrum Cyfrowe**, a locally and internationally renowned Polish NGO specialised in diagnosing social and cultural changes taking place in our society with the influence of digital technologies; the multidisciplinary design and development agency **Webtic**; **EUscreen**, a Foundation and network that brings together audiovisual and broadcast archives across Europe, administered by the **Netherlands Institute for Sound and Vision**; and **3Cat**, the primary television channel of Catalan public broadcaster Televisió de Catalunya.

   Project partners are based across different countries and sectors. We see this as a strength of *Watching Videos like a Historian*, as it has ensured access to diverse expertise to develop the project's tools and high-quality and relevant materials from archival collections. Project partners collaborate with each other through regular meetings (online and in person) and editorial sprints to develop resources, but we also take care that the project represents and reaches relevant audiences through capacity-building activities and outreach with teachers across the countries partners are based in.

3. Can you provide an overview of the audiovisual material you are making available as part of the *Watching Videos Like a Historian* Project? Which type of formats or genres are available through the toolkit you created?

   Research conducted by the Europeana Foundation between 2018 to 2021, involving more than 700 respondents, showed that teachers favour using videos and images as media content in the classroom, but that technical barriers prevent their consistent use. The project has tried to support teachers to overcome these barriers not by making new content available, but by making the content that exists - for example, on Europeana.eu and EUScreen - more accessible to teachers, and providing them with the resources to use them with confidence in the **Toolkit**. The Toolkit includes four main formats for teachers: Teaching Strategies (elaborate classroom activities that enable educators to insert any topic or content they require); e-Learning Activities (interactive tasks and assignments made available in digital formats suitable for remote and blended learning); audiovisual galleries (thematically curated collections of historical footage providing visual insights into the past); and Classroom Guides ('crash course' type content that helps educators find the most useful path for their work with audiovisual materials). As we improve the Toolkit, we will also include a fifth type of resource, namely the Learning Scenarios, which are specific, shorter activities that are 'ready to go' with minimum preparation necessary. The audiovisual materials that these formats make available and highlight include many focusing on historical events. For example, one **e-learning activity** focuses on German TV advertisements from the late 1960s, and embeds videos from Insel-Film GmbH, available through Europeana.eu. Similarly, the galleries curated as part of the project highlight audiovisual media, including **news clips**, **commercials**, **documentaries**

and **interviews**, bringing together materials from archives across Europe. While the Toolkit aims to support teachers to use audiovisual materials, the project's **Practitioner's Guide for AV Collection Holders** also provides help and guidance to archives with audiovisual collections guidance by advising them how and why to open them up for educational reuse. Through this, we approach supporting the reuse of audiovisual materials for education from both sides.

4. Could you talk about how the toolkit was designed to implement information about the videos in the collections for educational purposes?

   The Toolkit was designed and developed co-creatively with project partners and teachers based in three different countries: the Netherlands, Spain and Poland. Structured as described above, the Toolkit was designed to support teachers in using videos in the classroom, not only as a tool for teaching history but also as a means to create discussions where students are invited to critically analyze who created the videos, from what perspective, and for what purpose. Through capacity-building activities, the *Watching Videos Like a Historian* project tested the Toolkit's methodology with small groups of history teachers. Workshops and webinars were organized where teachers had the opportunity to design and implement their own teaching strategies. Additionally, they were encouraged to use Europeana.eu to find audiovisual materials on a variety of topics that could be incorporated into their lessons. Ultimately, collaboration between professionals from different countries contributed to the creation of the Toolkit, which reflects the diversity of education in various European contexts. In addition, there are plans to publish a second iteration of the Toolkit based on feedback from teachers who have tested it in their classrooms.

5. The project includes the creation of audiovisual galleries containing curated collections of historical footage. How were these galleries curated? Did the curator take into account the available metadata, or was it based on personal expertise?

   Ten audiovisual galleries were curated for this project on Europeana.eu, bringing together audiovisual materials centered on specific themes or topics. These galleries feature clips ranging from advertisements to news reports, available in multiple languages. The curation process typically involves selecting a focused topic, evaluating the depth and quality of the metadata, and checking whether the content is visually appealing. The clips were chosen for the galleries created in this project to enrich historical narratives and make them more engaging. This required selecting clips based on their thematic relevance, educational value, and historical accuracy. Additionally, essential considerations in the selection process included choosing clips with sufficient metadata to contextualize content and make it traceable. It was also crucial to ensure that all clips had either cleared copyright for educational use, were labelled as 'In Copyright – Educational Use Permitted', or licensed under Creative Commons terms, such as the Creative Commons Non-Commercial ShareAlike (BY-NC-SA) or the Creative Commons Universal License (CC0).

6. Which data and metadata were you given access through Europeana.eu or other platforms, and were they sufficient for this project? Can you describe how the information retrieval system was developed so as to highlight existing video categories, labels or metadata via keyword searches or queries?

   Existing metadata available through Europeana.eu and EUScreen was used and worked well for the project, in terms of being easy to locate, research and provide context for audiovisual material. For example, in the e-activity devoted to old TV advertisements, we used existing metadata to locate relevant AV sources on Europeana.eu. We began by entering broad search terms: 'advertisement' OR 'ad' OR 'commercial' into the platform's search bar. This initial query returned nearly 9 million results, which we then refined using filters to streamline our search. Since we were specifically looking for video materials, the 'Video' filter under 'Type of Media' was particularly valuable, as it excluded files in text, audio, and still image formats. This adjustment reduced the results to just over 9,000. From this refined list, we focused on items suitable for educational use, prioritizing those marked as being in the public domain or available with permissible usage rights ('Yes,

with conditions' or 'Yes' under 'Can I use this'). Additionally, we conducted a search using the 'Providing Institutions' filter to identify materials contributed by our project partners. In this particular e-activity, language was not a critical selection criterion, as the advertisements are straightforward and short, thus usually can be analyzed without speaking the language. However, during webinars with teachers, the language filter proved useful. The diverse linguistic backgrounds of participants highlighted the importance of tailoring resources to meet their specific needs.

7. Do you plan on implementing AI methods in this project at any point in the workflow? If you have already, which AI tools have been implemented? And if not, do you foresee integrating any form of automated algorithmic processing into this project, for metadata aggregation, enrichment or else?

   We have not implemented AI tools in the project workflow, and we do not foresee that we will use automated algorithmic processing. However, this does not mean that the project is not considering AI. We are very conscious that many of the students we are addressing through the project use AI in their everyday lives, and that there is a crucial need to support them in critically engaging with this technology. Alongside the Toolkit's educational activity offerings - all of which aim to support students to think critically about the media they are consuming - it also includes a dedicated Classroom Guide for teachers on using AI in history education, focusing particularly on how GPTs can be used in the classroom, as well as the considerations and challenges to make students aware of.

   The project has not focused on additional media content aggregation or metadata enrichment as part of its activities. In the future, it may be interesting to think about ways we could classify the audiovisual material used in *Watching Videos Like a Historian* by educational stage and age groups. Although educational stages vary by country, age-related needs are quite similar, and this metadata would be helpful for educators.

8. How does this project position itself in the tension between algorithmic practices as seen on streaming platforms and hybrid archival practices (both algorithmic and non-algorithmic)? And how do you, as the curator for this project, foresee future evolutions of AI applications for the management or open access and reuse of audiovisual resources for media literacy?

   Although we are not developing AI tools in the project (which was never the plan), many archives are already testing and implementing AI-based and algorithmic solutions. What perhaps indeed differentiates the archival approach from the algorithmic approach in streaming services is its attachment to quality control and standards concerning privacy, copyright and ethical questions linked to the use of AI. Many archives do use algorithmic processing more and more - to improve the searchability of its collections, support archival processes and explore the user's experience and interaction with the collection. And yes, AI brings many questions about ownership, including open access, lawful practices and competences needed to critically and consciously manage it. And those soft skills, critical thinking and media literacy will become even more important in the near future, allowing us to make wise analyses on the use of our data and interpretation of AI-developed data.

Notes

1. All interviews were designed and conducted by Giulia Taurino, Ph.D., who curated this section on audiovisual archives.
2. **https://openvault.wgbh.org/about**.
3. **https://americanarchive.org/about-the-american-archive**.
4. **https://portal.ehri-project.eu/institutions/lv-002874**.
5. **https://www.inamediapro.com/eng/Editorial/The-National-Audiovisual-Institute-Ina**.
6. **https://www.kb.se/hitta-och-bestall/om-samlingar-och-material/tv.html**.

7. As per Christoph Bauer (ORF), while no official 'year-of-foundation' was established for the ORF Archives, archiving operations at ORF date back to the foundation of the ORF Radio-service in 1924 (for audio) and of ORF TV-service in 1955 (for film and later video and television programming). The efforts for developing a 'service-oriented version' of the ORF Archives began in 1986, and eventually led to the fusion of the television archive with the radio archive in 2016.
8. **https://openvault.wgbh.org/about**.
9. Taurino, Giulia, and David Smith. "Machine Learning as an Archival Science: Narratives behind Artificial Intelligence, Cultural Data, and Archival Remediation." *NeurIPS Workshop on AI Cultures.* (2022). **https://ai-cultures.github.io/papers/machine_learning_as_an_archiva.pdf**.
10. This interview was conducted remotely in written form.
11. For additional information on AI initiatives at the Latvian State Archive of Audiovisual Documents refer to this ICA/PAAG webinar conducted by Natālija Lāce during the 2024 World Day for Audiovisual Heritage: **https://www.youtube.com/watch?v=enjfN3YcV7U**.
12. This interview was conducted remotely in oral form and transcribed.
13. **https://www.ina.fr/institut-national-audiovisuel/collection-preservation-and-documentation-of-audiovisual-heritage**
14. The Internet Archive was established in 1996.
15. The hierarchical storage management (HSM) system is a data storage and management technique that organizes data into tiers based on how often they are used and their importance, thus favoring cost-effective management of archival collections at scale through the automatic assignment of data to high-cost or low-cost storage media.
16. Questions 1-6 are answered by Mikael Johansson; for question number 5 regarding researcher access, Mikael has consulted Anna-Karin Lundgren. Questions 7-9 are answered by staff at the NLS data lab (KBLab): Faton Rekathati, Emma Rende, Chris Haffenden.
17. This interview was conducted remotely in written form.
18. Read more about the making of KBLab, the data lab at the National Library of Sweden (NLS). in the following OA article: **https://crl.acrl.org/index.php/crl/article/view/26325/34254**.
19. Read more about NLS' work on how AI methods can be used to improve the searchability and accessibility of visual heritage collections at: **https://kb-labb.github.io/posts/2023-10-20-unearthing-forgotten-images-with-the-help-of-ai/**.
20. This interview was conducted remotely in written form.
21. This interview was conducted remotely in oral form and transcribed.
22. AAPB is a collaboration between the WGBH Educational Foundation and the Library of Congress aimed at digitally preserving and making accessible historically significant public radio and television programs from local stations across the US.
23. WGBH-TV, known as GBH, is the main PBS member television station in Boston, Massachusetts (US).
24. This interview was conducted remotely in written form.

Biographies

Eléonore Alquier is Assistant Director and Head of the Data Department, Direction Data & Technologies at the Institut National de l'Audiovisuel, where she works as project manager and development manager in the collections division. She's in charge of the refoundation of the collections management IT system and of the collections development strategy, coordinating transversal projects connected to audiovisual description and conservation. She began working within the French National Archives (2010-2013) first as Archives Department manager, and then as Coordinator of the relocation and redeployment of the Archives between three cities. She was then Head of the Records and Archives Office at the French Social Ministries (Labor, Health, Youth and Sports (2013-2015), where she was responsible for the collecting policy applied to political records and archives produced by the ministers and their advisers. She is also an active member of the International Council on Archives, representing France in front of the Photographic and Audiovisual archives working group (PAAG). She has been involved in training sessions at the international level around records and archives management.

Montse Bailac is a Researcher at the archive of TV3 (CCMA), the Catalan Audiovisual and Media Platform. For the WVLH project, she works on capacity building, impact assessment (evaluation and analysis) and dissemination of the project, overseeing the inclusion of CCMA's audiovisual content in the Toolkit and helping the use and development of educational tools based on audiovisual content in classrooms.

Christoph Bauer, AV-Archivist, is the System administrator for the Archive-Systems and R&D Senior Expert at ORF. Over the years, he covered several roles at ORF, including Project Officer for several EC/IST/ICT/H2020/ FAA-Projects (PRESTO, PRIMAVERA, FIRST, NODAL, PRESTOSPACE, eCHASE, PRESTOPRIME, DAVID, EUNOMIA, TailoredMedia, FAIRmedia, etc.) and Senior Specialist for preservation, digitization and restoration in the ORF Archives department. In addition, he acts as system-administrator for Archive-Systems and AV-Digitization, workflow-developer and AI-mining-specialist (audio&video). He is member of the ARD Media Expert group Mining (AI) and member of the Digitization & Migration Commission of FIAT/IFTA (International Federation of TV-Archives).

Isabel Beirigo is Research Communication Specialist and Outreach Coordinator at the Netherlands Institute for Sound & Vision, where she coordinates Inward Outward communication and dissemination activities alongside other international projects. Historian by education, Isabel holds master's degrees in both History and Applied Museum and Heritage Studies. She has experience with audiovisual collection management and currently is focusing on studies about diversity and inclusion. As part of the WVLH project, she coordinates capacity building activities with her colleagues, including training sessions, workshops, and webinars.

Karen Cariani is the David O. Ives Executive Director of the GBH Archives and GBH Project Director for the American Archive of Public Broadcasting (AAPB), a collaboration with the Library of Congress to preserve and provide on-line access to content created by public media. Karen has 40+ years of television production, project and archival management experience for projects focused on providing access to audiovisual collections online and long-term digital preservation. She served on the AMIA board, Digital Commonwealth board, and the National Stewardship Digital Alliance (NDSA) Coordinating Committee. She chaired the NDSA Levels of Preservation revision WG, co-chaired the NDSA Infrastructure WG, and AMIA committees including Open Source, Local TV and Doc, and Copyright. Other projects include the GBH Open Vault, and the Boston Local TV News Digital Library.

Natalia Cetera, Head of Open Culture Studio at the Centrum Cyfrowe Foundation (CFF), has a background as a project manager in the field of digital transformation of cultural institutions, trainer in the field of new technologies, and art historian. For the past 11 years, she has been supporting the process of digitizing museum collections, creating digital collections, engaging online exhibitions and technological tools for creative sharing of cultural heritage. She co-created the digital collections of the National Museum in Warsaw, the National Museum in Krakow, the National Museum in Slovakia and the National Gallery in Bulgaria on the international Google Arts & Culture platform. One of the coordinators of the 'Open Zachęta' project - one of the largest digitization projects in Poland. Co-author of the publication 'Digital heritage of culture. How to prepare for digitization and making your collection available online'.

Maria Drabczyk, Chair of the Board and Head of Policy and Advocacy at the Centrum Cyfrowe Foundation (CCF), is a sociologist, researcher, project manager interested in the impact of technology and copyright regulations on culture, education and access to knowledge. She is a board member of the EUscreen Foundation, member of the Europeana Association Members Council and Chair of the FIAT/IFTA Value, Use and Copyright Commission. In the past, she acted as a project manager at the National Film Archive –Audiovisual Institute (FINA), in charge of international cooperation, mostly focused on access and creative re-use of the Institute's digital collection for research, educational or artistic purposes. She co-created a crowdfunding site for culture Wspieramkulture.pl and worked as an international relations expert at the Polish Ministry of Culture and National Heritage.

Svyat Elkind is Head of the Electronic Documents Department at the National Archive of Latvia, where he has been working for more than 10 years. For 5 years, he served as an expert on electronic records in the Institution Records and Archives Management Department. Currently, he leads the Department of Electronic Records, where he and his team implement best practices for the archiving of electronic data and records, while also providing support to institutions on archiving processes. He regularly delivers lectures on electronic record archiving at the State Administration School and other venues.

Georgia Evans is Senior Editorial Officer at the Europeana Foundation, where she is in charge of writing and developing communications aimed at professionals working in and across the cultural heritage sector. Before working

at Europeana, Georgia worked in the higher education sector in the UK. She has a BA degree in English Language and Literature, and an MA in Early Modern Studies.

Chris Haffenden is a research coordinator at the National Library of Sweden, where he works at KBLab with initiating and facilitating research projects that explore novel approaches to the library's digital collections. He holds a Master's degree in Political Thought and Intellectual History from Cambridge University and a PhD in the History of Science and Ideas from Uppsala University.

Herbert Hayduck has been Deputy Head of Documentation and Archives at ORF since 1989 and Head of the ORF Archives since 2008. Since 1995, he has been in charge of the ongoing ORF coordination of EU research projects in the field of 'Audiovisual Archiving and Preservation'. Outside of ORF, he collaborates with FIAT/IFTA (International Federation of Television Archives) as a member of the 'Documentation Commission' and later of the 'Media Management Commission'. He participates in ongoing lecturing activities at international conferences on all topics relating to media archiving. and he is lecturer at the University of Vienna.

Rūdolfs Henčels is Head of the Document Digitization Department at the Latvian State Archive of Audiovisual Documents. Rūdolfs has over a decade of experience in film and video digitization, holding various positions within the archive. Throughout his career, he has demonstrated exceptional expertise and enthusiasm in the digitization and restoration of image and sound, consistently devising and implementing optimal solutions using available technical resources. Rūdolfs' extensive skills and knowledge have enabled him to provide expert consultations on selecting technical equipment for establishing high-quality, sustainable film digitization infrastructures. Additionally, he has designed efficient workflows for film digitization and restoration, tailored specifically for long-term preservation purposes. His contributions have significantly advanced the archive's mission to preserve and make accessible Latvia's audiovisual heritage.

Christine Hveding, Business Development Coordinator for Education at Europeana Foundation, is responsible for programming, partnerships and projects in the field of formal, non-formal and informal education. As a Project and Business Development Coordinator, Christine works to support the seamless delivery of EU-funded projects with Europeana Foundation Participation and exploring new funding and partnership opportunities. She has a lifelong love for the arts and heritage field and a passion for its intersection with technology. She holds a BA in Arts Management and a Master's in Arts and Heritage: Policy, Management and Education. As part of the WVLH project, she ensures the seamless delivery of Europeana's contributions for this and other EU-funded projects.

Mikael Johansson, AV-technician and member of the digital preservation team, has been working with the audiovisual collections of the Swedish National Archive of Recorded Sound and Moving Images and the National Library of Sweden since 1995. He is also part of the NLS's Digital Preservation Team.

Natālija Lāce is Archive Expert at the Access and Inquiry Department, Latvian State Archive of Audiovisual Documents, where she specializes in the description and cataloguing of audiovisual collections. Natālija has played an important role in developing strategies for integrating AI tools into archival processes. Her responsibilities include compiling archive fonds' histories, maintaining databases, and improving archival systems. Additionally, she has experience working in museums and as a teacher, enriching her expertise in both educational and cultural sectors. Natālija holds a master's degree in history from the University of Latvia, complementing her professional skills in digital and AI-driven solutions. As a member of the International Council on Archives, Natālija contributes to the global archival community. Her current research focuses on promoting sustainable archiving practices and enhancing the use of AI in small archive systems.

Anna-Karin Lundgren, Librarian, coordinates the Research Service of Audiovisual Media at the National Library of Sweden, and has worked with the audiovisual collections of the Swedish National Archive of Recorded Sound and Moving Images and the National Library of Sweden since 1997.

Ivan Milovanov is Professional Development Coordinator, Project Manager and Educational Journey Designer at EuroClio, where he leads L&D projects, offline and online events including EuroClio's Annual Conference, and the Historiana platform. Previously, Ivan's notable contributions as an entrepreneur include the 4M Storytelling method, which integrates visible storytelling with collaborative design, educational city walking tours focused on history and remembrance, and Stories that Move, a blended-learning platform. His approach to creating learning experiences blends innovation, psychology, and playful storytelling, with a focus on active citizenship, democratic values, critical thinking, history, emergent technology, and environmental and economic sustainability. He co-manages the WVLH project project and co-authored some of the texts in this Toolkit.

Faton Rekathati is a data scientist at KBLab (National Library of Sweden), where he works on improving the searchability and accessibility of the library's collections. He holds a Master's degree in Statistics from Linköping University.

Emma Rende is a product manager at the National Library of Sweden, where she works at KBLab with launching the initiative KBx, which combines AI insights and library expertise to enhance the library's working practices. She holds a Master's degree in Computer and Systems Science from Stockholm University.

Helena Vicente Galán is a Psychology graduate and a trainee at EuroClio. She holds a Bachelor in Psychology from the University of Santiago de Compostela and a Masters in Multilingualism and Interculturality from the University of Strasbourg. Her research interests include youth participation and impact and traumatic historical narratives. She has collaborated in the drafting and editing of the WVLH project Toolkit.

Dorota Wiśniewska is Assistant Professor in the Department of History and Civics Didactics, Institute of History, University of Wrocław, Poland. She has been involved in several projects led by EuroClio, including **Online Teaching in the Visegrad Region**, and more recently **School EducatioN for Sustainable and Equal Inclusion**. For the WVLH project, she developed eLearning activities that will help teachers put into practice the theoretical approaches outlined in the Toolkit.

VIEW Journal of European Television History and Culture Vol. 13, 26, 2024
URL: **https://doi.org/10.18146/view.358**
Publisher: Netherlands Institute for Sound and Vision in collaboration with Utrecht University, University of Luxembourg and Royal Holloway University of London.
Copyright: The text of this article has been published under a **Creative Commons Attribution-ShareAlike 4.0 International License.**
This license does not apply to the media referenced in the article, which is subject to the individual rights owner's terms.