volume 13 issue 25/2024
VIEW
JOURNAL OF EUROPEAN TELEVISION HISTORY & CULTURE

THE AUDIOVISUAL ARCHIVE IN AN ERA OF DISINFORMATION AND MISINFORMATION

Jacqueline Pietsch
Agence France-Presse
jacqueline.pietsch@afp.com
in collaboration with the MediaNumeric Team

Abstract: Audiovisual archives should reflect on their mission and goals in an era of overwhelming computer power. Will they be able to make good use of Large Language Models (LLMs) to unlock archived materials? Should they and can they be an ally in combating misinformation and disinformation? As part of a larger project on data literacy for journalists and other media and creative industries professionals, archivists were questioned about the challenges facing audiovisual archives today. Rather than focus on the specific missions of either national or broadcaster's archives, they focus on how the archive has an important role when it comes to the politics of representation in public debate and civil life. In convivial conversation they speak from their experience at the French National Audiovisual Institute INA, The Netherlands Institute for Sound & Vision, the EBU Academy, the Spanish RTVE archive and WITNESS, a human rights non-profit organisation based the United States that supports activists in archiving and preserving their video.

Keywords: audiovisual archives, digital media, AI, fact-checking, politics of representation, misinformation

1 Introduction

Audiovisual content creation has exploded. This includes the journalism and (public) broadcasting landscape which is changing at breakneck speed. We've collectively experienced a move away from the daily newspaper and the News at Six/Eight/Nine just a few decades ago, to 24-hour rolling news channels, and now information on demand at any time of the day or night, from any location in the world via the internet and mobile phone networks. Audiovisual archives that focus on broadcasters' journalistic and other news and documentary content are faced with a challenge. As repositories of both the past and the present, they take in new content daily. Will they be able to make ethical and fair use of the new Large Language Models to make content identifiable? Will such computer programmes also be of help in making the archive accessible where material may have been digitised but not been given comprehensive metadata to make it findable? Large Language Models (or LLMs) are a type of artificial intelligence (AI) that can recognize and generate texts. Such models use enormous data sets, which archives have no trouble supplying. Once the programmes have been trained, they are able to recognize and interpret human language and similar complex data: facial recognition is an obvious example. Also obvious is the fact that the quality of the AI and its ability to generate short summary descriptions, for example, depends on the quality of the data that it was fed and whether the programme was 'taught' well by its often forgotten human coders.

As part of MediaNumeric, an extensive European-funded project on education in data literacy, journalism, and multimedia storytelling, representatives of national audiovisual and broadcast archives were interviewed.[1] These interviews revealed interesting observations about how LLMs might be enormously useful and also how the archive could be of service in making high quality data sets available. The interviews also hinted at older rather than new questions that come to the fore in this era of enormous content production by trustworthy and untrustworthy professional organisations and amateurs alike. Archives are of evident importance in providing reliable reference to historical data for journalists, researchers and activists. By maintaining accurate records and providing robust metadata to make archived materials available, archives can help combat misinformation and support the integrity of factual storytelling. The question however arises as to what exactly integrity signifies in this context? Are any and all images as they were broadcast, or gathered for the purpose of public broadcast sufficient in themselves to give a sense of the past? And, was the content that was archived created with a critical, fair and balanced eye? In additional conversations, three key figures in archiving who are close to the MediaNumeric team elucidated what they saw as the real challenges in the period ahead.

The three extra conversations were with Antoine Bayet of France's National Audiovisual Institute (INA), Willemien Sanders of Utrecht University who works closely with The Netherlands Institute for Sound & Vision, and Johan Oomen, Head of Research and Heritage Services at Sound & Vision. Both INA and Sound & Vision are national archives that specialise in public broadcasting. Of the earlier MediaNumeric interviews, those with Alexandre Rouxel, Hélène Rauby Matta, Virginia Bazán-Gil and Yvonne Ng were used. Rouxel and Rauby Matta are affiliated with the European Broadcasting Union's Data and AI section and its Academy, Bazán-Gil with the Spanish public broadcaster archive RTVE.[2] Here, the focus is on how the challenge to combat misinformation and disinformation which spurred the MediaNumeric project, via the lens of the possibly new role of the archive, calls forth older questions regarding the politics of representation. For what is archived was broadcast, and hence journalists, other makers and editors decided on what (and who) to include and who not. For this reason, and to include a different perspective, we also draw on the MediaNumeric interview with Yvonne Ng of WITNESS, a non-profit organisation focused on citizen journalism and supporting "activists to archive and preserve their video so that human rights abuses cannot be denied or forgotten over time."[3]

Misinformation and disinformation refer to actual falsehoods and lies. There is however a grey area where misinformation rests on suggestion. It comes perilously close to how dominant codes of representation in a given time space might have also held suggestions, stereotypes and absences. The real mission then, for audiovisual archives in this day and age, is to make archived materials not only easily available and discoverable, but also to provide guidance and reflection on how we might understand the fragments and snapshots captured in those archived materials. They do not offer 'truth', they offer a historically specific and always partial piece of information that the audiovisual industry as well as citizens will have to learn to view with a critical thinking lens, and learn to understand its value.

Open, convivial conversations felt like the best course to bring journalism, data science, research and archives further into dialogue with one another. We need the perspective of the practitioners in these fields to bridge the gap between these spaces, and to understand what is needed of these spheres. This paper is purposefully positioned as practice-oriented work. The text ruminates on, rather than interrogates, the 'state of affairs' of audiovisual archives, their collaboration with journalists, and the potential role national audiovisual archives could play in combating dis- and misinformation. In doing so, we believe we have been able to bring to the fore how practitioners, like academics, reflect on their field but in more hands-on ways that helps inform the mission of the audiovisual archive of the near future.

2 Seeing is Believing: Images, Misinformation and the Need for Fact-Checking

The MediaNumeric work started from the shared straightforward notion that the spread of misinformation needs to be contained. We welcome the power of the media ordinarily and how it is a force of connection in society.[4]

False information and information that intends to cause harm is, however, a problem. In a report for the European Commission, Wardle and Derakhshan distinguish three broad categories of problematic content: misinformation - information that is false but not intended to cause harm; disinformation - information that is false and deliberately created to cause harm; malinformation - information that is genuine and used to inflict harm.[5] Information is the raw material of journalism and documentary AV production. Reputable media organisations have a mandate to only report verifiable and true information, but in a world where anyone with a smartphone or a connected device can become a so-called citizen journalist, a large number of people now get their news via social media platforms where information sourcing and reliability can be very different. A dedicated fact-check team at Agence France-Presse (AFP) has a mission to verify news content that courses through these social media platforms.

AFP, in fact, has the world's leading fact-checking network, with 150 journalists working on five continents in 26 languages.[6] Over the course of their work, they regularly find videos that have been edited or have been shared out of context, in order to represent an alternate reality. Times of conflict, crisis and a paroxysm of confusion are particularly rife with manipulated videos, such as the war in Ukraine,[7] the conflict between Israel and Hamas in Gaza,[8] or Iran's air strikes against US military bases in Iraq.[9] Manipulated video often appears in narratives of emotive societal issues, such as migration,[10] health crises[11] or climate change.[12] In the process of fact-checking and digital verification, journalists track images up to the point of first appearance. Accurate metadata, the hidden extra data (or the data about the data) that describe provenance and other basic information such as where and when the images were shot and the channel of first broadcast/first sharing, are crucial to this workflow called 'debunking' and the ability to irrevocably prove the nature of the images and the events they depict.

In the fight against mis- and disinformation, many news generators and gatherers and equipment manufacturers are working together within the Coalition for Content Provenance and Authenticity (C2PA)[13] to develop verifiable, tamper-evident signatures that indicate that neither the image nor its associated metadata has undergone undisclosed alterations. By formulating technical standards and protocols for embedding metadata into digital content, such as images and videos, C2PA lays the groundwork for verifying the genuineness and origin of such content. C2PA unifies the efforts of the Adobe-led Content Authenticity Initiative[14] which focuses on systems to provide context and history for digital media, and Project Origin,[15] a Microsoft- and BBC-led initiative that tackles disinformation in the digital news ecosystem.

Well-embedded metadata fosters transparency and a clearer understanding of content creation and alteration processes. It empowers journalists, fact-checkers, non-governmental organisations, and news consumers to discern credibility more effectively. The news generators see it as fundamental to fighting the erosion of trust in the news media.[16] When working on the MediaNumeric white paper, we became interested in how audiovisual archives might integrate this technology into their content management systems. Below the partners in conversation will introduce how they see the role of the archive when it comes to misinformation, but also how misinformation has a grey area overlap with misrepresentation, which needs us to understand all media-making as work that is specific to a given time and place, as well as to personal, professional and organisational cultures and goals.

3 Looking to the Past to Contextualise the Present: How Audiovisual Archives Can Preempt Misinformation

Archives are important as they can, through drawing on original materials and historical references in their collections, provide the evidence that fact-checkers need to demonstrate the original use of the audiovisual content as part of their debunking process. This depends, of course, on the original material being discoverable and well-checked. Jacqueline Pietsch spoke at length with Antoine Bayet, Editorial Director at the National Audiovisual Institute (INA) in France. INA are striving to get ahead of the misinformation curve. In order to do so, INA is opening up to the wider public. For archives such as INA, combating misinformation and disinformation through contextualisation aligns well with older

goals of educating professionals and the general public about the past. After all, archives provide an important means of contextualising today's strong convictions and misconceptions.

INA's opening up to the wider public evolved in 2015 and 2016. As video exploded on the web and social media platforms, INA began participating in *France TV Info*, a domestic rolling news channel that began broadcasting on the web in France on August 31, 2016 and on television a day later. This story starts a decade earlier, in 2006, when INA, the legal repository for France's radio and television audiovisual archives, created a new website, www.ina.fr, to give public access to the archived collections of 100,000 historical programmes.[17] The website crashed shortly after launch under the sheer number of connections as many tried to access the content.[18] INA realised that there was incredible interest in audiovisual archives, not just from media professionals, but also among the general public. In the beginning, the decision about what to promote on the website was relatively traditional: commemorations of past events or what happened on a specific date in the past. "We started with the past and tried to give past events new life today," Bayet said.[19] In 2016, with the launch of *France TV Info,* INA turned this on its head. From this point onwards, INA, along with its partners *France Télévisions*, *Radio France* and *France Média Monde*, started with a topical news event and scoured the archives to provide historical context to explain current events. As Bayet continues:

> We no longer start from yesterday and rebroadcast yesterday today. We start with today and look back into the past to better understand what is happening now. The evolution has been radical and has led to a significant change among viewership. They have more connection with the archives as they relate to today and not just about the past.[20]

Now, the archive team is integrated into the heart of the decision-making workflow: they participate in the morning news conference alongside INA's journalists when the day's topics are discussed. They also work at speed to find the content and publish, within a few hours, stories with context and background drawn from the archives to match with a rolling news environment. Bayet is convinced that this approach contributes to the fight against misinformation, disinformation and malinformation:

> When there is an operation to manipulate information, there may not necessarily be anything false in what is broadcast. It can be based on a truth but it is transformed, deformed, amplified and it ends up being decontextualised. People could then link it to a specific event in the past when in reality, there's really no connection, and it can change peoples' perception of an event. INA's role in the reality that is malinformation or disinformation is to contextualise and to prevent the decontextualisation of news.[21]

In daily life misinformation and disinformation are often conflated. From the point of view of fact-checking and of archival expertise, they need to be understood in their own right. Moreover, although misinformation is not intentionally created to deceive, in the context of journalism it may overlap with how inadequate preparation by a journalist, lack of oversight by an editor, or simply the desire to conform to accepted social norms, produce a partial representation of reality.

Jacqueline Pietsch also took up the issue with Willemien Sanders, a media scholar affiliated with Utrecht University (UU) and Sound & Vision. Sanders is specialised in data-driven storytelling. Her research shows that audiovisual archives contain powerful stories, originated from both contemporary content and the use of archives to change the narrative of modern-day viewpoints.[22] Sanders gave an example of how slanted versions of reality come to be widely accepted. Not, perhaps, misinformation but surely misrepresentation. When certain groups of people are always represented in a specific light, the constant reuse of a specific kind of image reinforces and perpetuates a trope of that group of people, be they rich, poor, or from specific ethnic, religious or social backgrounds.[23] This may fix people into one non-evolving identity through their representation. Alert archivists can use the archives to change today's stereotypes. People with Moroccan or Turkish backgrounds, for example, continue to be identified by some in the Netherlands as passive and not integrated into society. Using the Sound & Vision archives to investigate, Sanders found numerous pieces of footage of these migrant groups as active and engaged members of Dutch society when they first arrived in the country, countering a pervasive and problematic stereotype:

> This is an image or a representation that we don't often see nowadays. So by showing and really searching what was there before and diving into the archive […] there's this alternative representation that sheds another light on how we look back. In the very contemporary discourse on foreign workers or foreign people in the Netherlands, it's really useful to have an alternative representation that also shows: 'Hey, these people are here because we asked them to come over and we needed them to do the work that we didn't want to do and they're not just passive or unwilling or whatever.[24]

In a similar example, INA provided crucial context to the #MeToo movement against sexual violence against women in France. Here accepted notions of the privileges of certain groups in society are dismantled. In 2019, INA archivists located an excerpt from a 'high-brow' literary talk show called *Apostrophes* on *Antenne 2* that dated from 1990.[25] In the clip, Gabriel Matzneff, a well-known and well-respected author explains his preference for sex with underage girls and boys, some as young as 14. Only one of the other guests, Denise Bombardier, challenges Matzneff on his behaviour and calls it out as the sexual abuse of minors.[26] The way in which she was pilloried for her criticism of Matzneff goes some way to explaining how sexual abuse among France's literary and film circles was tolerated and covered up for a long time.[27] Access to archival material gives us insight into events of the past, as well as the opportunity to reflect on our reactions as individuals and as society to them. The archive opens our eyes, Bayet suggests:

> With each new case [allegation of sexual misconduct], when we rewatch sequences that were broadcast on the television in the past and which didn't shock anyone at the time, we are surprised to see that this happened in full view of everyone.[28]

4 The Challenge of Contextualising Well in using the archive and according metadata

How to reissue archival content so that it becomes accessible to contribute constructively to debate today is a challenge for archivists and journalists. When speaking about this with Bayet, he said that in the case of historic sexual abuse in France, the archivists and journalists at INA had considered the 'what, whether and how' extensively:

> Each time we deal with this subject [allegations of sexual misconduct], the amount of contextualisation needed is quite dizzying. Our purpose is not to simply broadcast powerful clips for shock value. Our role is to transmit as a fact what was broadcast at a specific point in time, […] We always ask ourselves how it will contribute or not [to the debate]. There is constant arbitration. One of the conclusions we could make would be to not rebroadcast because with the eyes of 2024, this clip is unbearable to watch and as we haven't found the right way to reuse it, it will just be shocking. But on the other hand, self-censorship is neither virtuous nor recommended. These questions are very, very, very important and we systematically ask ourselves these questions. The right method often lies with ultra-contextualisation.[29]

The art and craft of contextualising well is as important when discussing specific examples when reissuing material as it is when recording metadata. Archivists and researchers have ongoing internal discussions about how to classify sequences and what kind of metadata should be used.[30] A key question is whether, if, and, how hurtful and discriminatory conventions of the past should be repeated or not in the evolution of archival records? Sanders explained that while reviewing children's programmes of yesteryear, researchers came across a children's show that contained an offensive racial slur against Black people. Should the N-word be included in the metadata? Some argued that it was a quote, others felt that the slur should be described but not directly quoted as it is highly offensive. How to look forward when adding metadata, what type of terminology might future generations use in their search? Sanders adds:

> It's a real challenge to make metadata sustainable for the future. In a few years, those people who are still young will probably use different terms to search for material. How can you make sure they find it?[31]

How to do this is a question that needs European archives to agree on a common standard for metadata; collaboration is necessary but also discussion about how the seemingly neutral activity of archiving and retrieving archival materials is deeply political. To be invested in the #MeToo cause is a political choice. To research and redress how the Dutch think about migrant groups is too. At issue is whether materials relevant to these questions and causes can be described in such a way that local differences in attitudes today towards societal issues are respected rather than lead to fully different ways of describing content within the archives. For example, LGBTQ+ rights are considered differently across the European Union and so the metadata describing the content will be based on different principles. Sanders counsels against underestimating the difficulty of such questions:

> It's naive to think that we can create some sort of objective way to describe certain phenomena or groups or people because the terms that you use to describe certain phenomena already have meaning and create meaning and assign meaning. […] So really the question is, can we agree on an approach?[32]

With misrepresentation coming close to the categories of misinformation and disinformation, this is a crucial issue. More often than not, objectivity and impartiality are assumed to be a well shared amongst professionals. This is hardly the case as they too work in and for organisations that have specific goals and aims. In the MediaNumeric interviews we came across similar questions and assessments.

Yvonne Ng, a digital archivist at the human rights advocacy non-profit organisation WITNESS, said finding objectivity and common ground in archival practice is extremely challenging.[33] The perception that the science of archiving is grounded in objectivity is widely shared. Value judgments however are being made every day, starting with the collection process. "I think all collections are biased, a term that is often read with a negative," Ng said, "But you know all collecting is done with a certain perspective. There is no neutral position so it's more about being transparent."[34]

Sanders concurs:

> Despite all the technological developments, basically we still use a lot of human beings to develop those technologies and also to do a lot of the work. Humans with all their flaws and shortcomings and preferences and ideas and everything, so it will never be perfect. […] It should be a tool, not an end in itself.[35]

Not only is technology human work, it touches on ideologies and convictions. Better therefore, when a venture such as coming to high quality future-proof sustainable metadata, that it is organised together with as many stakeholders as possible.

5 Using AI Tools in the Classification of Audiovisual Archives

Rapid expansion of their digital collections is a real challenge for archives. To use and reuse materials in archives, they need to be findable, which depends on having comprehensive and accurate metadata. The amount of content stored at INA, Sound & Vision or similar archives is so large that it is impossible for people to categorise everything manually. Archivists are therefore experimenting with artificial intelligence (AI). AI's Large Language Models can be a game changer but they also pose a fresh raft of difficulties. Tools such as automatic and real-time speech-to-text transcription, speech segmenters, face and object recognition (to identify publicly recognised people or objects), language translation or the extraction of named entities (subjects from the real world such as names of people, locations, businesses, topics and themes) can process mundane tasks in seconds that would take archivists hours, if

not days, to perform accurately. However, what preconceptions and prejudices will the AI incorporate into the metadata?

Virginia Bazán-Gil serves as the Head of Archives at RTVE Radiotelevisión Espanola. Her archive actively works with AI developers to develop, test and improve retrieval systems. They participate in the biennial Albayzín Evaluation Challenge for Speech Technologies in Spanish,[36] a series of technological evaluations open to the scientific community to propose challenges and data sets in the broad area of speech technologies.[37] As in Trusted European Media Data Space (TEMS), another such collaborative venture, an interesting exchange is taking place.[38] RTVE provides data sets for testing and in return learns about different technologies in development and can decide whether it is worthwhile to put those technologies into production. Bazán-Gil shares:

> From our point of view, the most challenging part is tagging the content. We spend a lot of time, effort and money in speech transcription, diarisation and tagging of the content so researchers can use this data to train and develop their own systems.[39]

With so many different ways to tag content regarding its provenance and meaning, she advocates for careful consideration ahead of time of the different types of metadata to add and how this can be done in a process she described as 'data economy': "Why should we create new data if we can reuse data generated by others?"[40]

Data economy will need to take the systemic bias of AI into account. While AI may well promise to unlock the vast unlabeled contents of audiovisual and other archives, it will also destroy the value of archival materials if it mislabels content from a biased white male perspective, as has been shown to be the case by feminist data research as well as by engaged journalists. There is much discussion among people developing and/or using AI tools about systemic bias that may have been built into the LLMs that were used to train the system. As the technology passes into the hands of private companies, it becomes difficult to know how the LLMs were put together, with what content and what kind of algorithms. Algorithm bias can totally change the type of results the AI provides. Image classification and identification on Google and Facebook, for example, have both in the past identified Black people as gorillas or as primates.[41] [42] The AI sector is still dominated by white men who may not easily see what is missing from the data they use to train the AI algorithms.[43]

Alexandre Rouxel, Senior Project Manager specialised in Data and AI at EBU, says the organisation needs to be at the forefront of technology. To avoid the possibility of systemic bias in the AI tools it uses, EBU therefore decided to build their own LLM. They used public data sets available in the field of AI research and supplemented this with other content which was cross-referenced with information from news sources that have a track record and a mission to report accurately. In all, EBU used about 4 million news articles to design the architecture, develop, train and test the models. Among other things, EBU is developing a facial recognition tool to scour the archives of its member organisations which it will make available for its members. Having their own tool has the added advantage of being able to train it to recognise local and national celebrities who may not be well known internationally and therefore indiscernible to off-the-shelf software.[44]

Developing proprietary tools means greater control, precision and accuracy over your own algorithms. It is possible to modify the algorithms as needs and demands on the archive evolve and, crucially, any problems with the algorithms themselves are much easier to address. Algorithms can therefore also be trained to respond to biases that may already exist in the archives when content was produced during periods of authoritarian rule or when societal prejudices were more prevalent. As EBU has control over the LLM, it can also make sure its tools comply with European General Data Protection Regulation (GDPR), an evident but not easy to realise requirement for a European organisation. It now only registers the names of people within the public domain. Overall, Rouxel is optimistic about what AI will achieve, and also when checking content for disinformation:

> A number can be true depending on the context so you have to understand the context in which the number was presented. That can be difficult for AI. A journalist can understand humour which can refer to things that

> happened long ago. This is why it is difficult to make a firm decision. There are not many AI tools that can do this because it depends too much on the context.[45]

The EBU factored in this challenge when building its tools. It attributes a 'reliability score' to articles, which is based on analysis of the language used in the context of disinformation. The tool warns journalists that certain statements should be checked carefully. Such software is based on the same speech-to-text tools which in theory will enable large audiovisual archives to annotate huge amounts of content. A possible 'game-changer' for archive repositories, Rouxel says:

> It's impossible to multiply by 10 or 100 the number of archivists. The fact that artificial intelligence is arriving with mature technology, may change the way that they work. The more mundane and less interesting tasks could be taken over by artificial intelligence, which means that archivists could take on roles that are more creative.[46]

Effective and reliable data enables media partners to share and exchange content and data in ways that were previously impossible and thereby strengthen the resilience of their alliances. Through these new search tools, journalists and storytellers are able to access more relevant content than ever before and identify stories that had hitherto been hidden. More is possible, though. Whether alerted by AI for possible misinformation or disinformation, or served with far more audiovisual material than before in archival research, the one hard condition for this to work well is that archives have access to how the LLMs were built and so that they can verify and teach the algorithm. There is more than enough work checking for contemporary as well as historical bias without AI magnifying and embedding racism and sexism (amongst others) in how historical records can be accessed.

6 Journalists Finding Their Way (in)to the Archive

The Covid pandemic beginning in 2020 was a turning point for the relationship between audiovisual archives and journalists. The broad range of uses for archival collections and the development of tools to unlock archival material that are now taken very seriously, are also a result of how journalists 'discovered' audiovisual archives. With almost no ability to film new content during mandatory lockdowns, television producers and journalists turned to audiovisual archives to fill their news and television programmes. Hélène Rauby Matta, a Business and Training Development Manager at the EBU Academy and a MediaNumeric interviewee, remembers how the relationship between the news desk and archivists changed during that period:

> Quickly we had a request from our members who found themselves working from home, including journalists and reporters and producers. These people had little material to work with and suddenly started, I wouldn't say rediscovering their archives, but certainly paying more attention. Shooting was really difficult for several months. So they started going back to their colleagues to establish that connection and started, maybe, talking again to the archivists.[47]

As development manager, Rauby Matta speaks optimistically about training an increasing number of journalists in AI, data and big data tools at EBU Academy:

> We train dozens of journalists on building constructive stories, and that is also offering multifaceted approaches to issues and storytelling that is not just based on the one event, but may be going back in the past. I think this has helped redefine journalism a bit for some of our members. I think this has had a positive impact on reconsidering how you can use archives to actually feed and enrich the stories. It's the same approach in a way when, as a public service media, you try to debunk fake news or you try to educate your team. It's also about substantiating news stories with past elements with historical facts for which again you may have to dig into your archives and do a bit more research.[48]

Willemien Sanders shares this sentiment: “When used well, archival research can debunk strongly held public misconceptions. Such research however is also about what is archived. Archives do not hold neutral or fully verified data. Idealised notions of the archive also need debunking. That is to say that audiovisual archives, whether national ones, or those of individual (public) broadcasting organisations hold material that is imprinted with the norms and values of the historical period in which it was made.”

In a project called CLARIAH,[49] Sanders analysed airtime for various politicians over a given timeframe in recent years. She was able to identify that certain political parties appeared comparatively infrequently in the media which meant that their voices and their ideas were not as represented. This says something about how the media portrayed the political landscape in the Netherlands at the time, but also about what the archive shows us and what it does not. Contrary to her earlier work in which she was able to uncover a different story about migrant groups, this project brought a stark reality of what an archivist is able to achieve:

> I used to refer to [Sound & Vision] as the history of the public debate in the Netherlands, but I increasingly realised that it’s a very limited public debate because a lot of groups were never included or never given a chair in that debate, they were never given an opportunity to participate in that debate. That is something that I think we should realise is problematic.[50]

What archives can offer depends on the material that comes their way. While one might assume that public broadcasting material represents society as a whole, this is of course not the case. In public broadcasting and the various forms of journalism and documentary making that are being archived, specific groups and issues will also be underrepresented. Ideally, archives adjust their collection policy accordingly and seek alternative sources to document the media and lives of underrepresented groups.

Such an approach is difficult to apply for archives managed by broadcasters. Broadcast archives typically only preserve content that was either produced or aired by the broadcaster itself. Of course, the root cause of underrepresentation lies not in the archives’ collection policies but in the editorial decisions of the broadcasters in this specific context. Broadcast archives as an internal department of the broadcaster have no control over these editorial choices. However, national audiovisual archives need to take on the extra burden of critically assessing what is assembled elsewhere and what extra sources, among them possibly social media, should also be collected. Journalists using the archive should explicitly be trained in recognising the limits of what they might well understand to be an all-encompassing treasure trove and critically assess the materials they draw from, just like they would with more traditional journalism sources.

7 Bringing community stakeholders into the archive

INA brought in professional outsiders into the archive. Willemien Sanders points out benefits to also bringing in ‘amateurs’. She is thinking of members of under-represented groups in society. They, better than others, can shine a light on what is available there and how content that has been stored portrays them. Community members can, presumably, help build more accurate metadata as well as identify what is missing from archives that mean to collect the ‘history’ of a nation. When archives are understood to be and to store community property (whether this is a national or a language community), we have a better chance of no longer thinking of ‘data collection in a vacuum’, as archivist Yvonne Ng puts it:

> Data is information about people and, especially in the context that I work in, we’re very conscious of that and we always like to work in partnership with local organisations or people who are directly impacted by the data. We’re not really interested in data collection for the sake of data collection but always for a purpose.[51]

Sanders and Ng do not only invite us to think about how useful it would be to have outsiders enter the archive, they help build a fully new notion of the archive as an open and connective space. Rethinking how open or closed an archive is includes reconsidering who can have access and make use of what is stored in the archive. Sound & Vision, for its part, facilitates research and data analysis of content for third-party journalists, researchers and political scientists. Johan Oomen, Head of Research and Heritage Services at Sound & Vision, cites Sanders' study into airtime given to different politicians and political parties on Dutch television. Interestingly, he said there is a general perception that public television in the Netherlands is progressive and left-leaning but the airtime analysis showed that centre-right politicians, who were in power during the time period covered by the study, dominated airtime. The results debunked the idea that public television is skewed towards left-leaning politicians.[52] Oomen shares:

> One of the very promising avenues we're exploring has to do with framing analysis, to look at how content is being reused. We can then determine how, why or which voices are amplified and which voices are less prominent. It's interesting to identify how our mainstream media works. The fact is that material is being reused and reused and reused, and then it becomes sort of canonical. Then, when the next journalist comes around and says, 'oh, yeah, I'm looking for that specific image', then that becomes a self-perpetuating mechanism, which is quite interesting from a media studies point of view.[53]

From the perspective of the archive, it seems imperative to notice such mechanisms and stop the building of imagery that starts to lead its own life and then feeds into social movements or influences representation in media, which, eventually, will reach policy making.

Oomen here touches on the politics of representation. This is an issue that is woven deeply into the tapestry of the archive. As others also point out, archival materials bear the imprint of when they were created. With underdeveloped or poorly developed metadata, they can become untraceable, or worse, only findable if they relate to specific individuals or groups in specific contexts. Social polarisation can be fed this way. Better to tackle the issue of what the (in this context audiovisual) archive represents head-on: is it the content produced by a specific organisation? Is it what is produced as public television and radio in a specific country? Or do the mandate and financing of the archive allow for a broader collections strategy? Can it bring in outsiders to provide insight and perspective into how materials are identified and described in metadata? Can it ask third parties what should also be collected? All of these are solid, practice-based questions that require answers. They make clear how the need to involve the archive in combating misrepresentation, disinformation and misinformation, in itself a sensible and valuable strategy, carries enormous responsibility.

That responsibility, it turns out, includes even more thinking of archives in terms of how its content, its descriptions, the valuations contained in its metadata often mark exclusions. Archives as they were discussed in the MediaNumeric project and in the conversations with Bayet, Sanders and Oomen are not co-owned institutions. They have specific mandates and forms of funding that in many cases limit access and re-using the collected materials. Much of the audiovisual content held by national audiovisual archives is under strict copyright and it may not be open to the general public. National audiovisual archives often provide a large-scale hosting service for television channels, but it is the channels who retain control over their content. Some archive institutions, like INA, are making more content available online. In most countries access is restricted to authorised viewers such as academics and researchers which vastly reduces the number of people who can access a country's heritage materials. Copyright can even extend beyond the content itself to all of the associated metadata which means that even those who are granted access need to create an additional layer of independent aggregated metadata over the top to be legally allowed to use the information.

Archives are costly to create, upgrade and to maintain and the data has a financial value. The reuse and sale of archive material might well become an important revenue source for television channels and radio stations. That would mean that only a select few are able to intervene on the content. Such a development would run counter to the vision our partners in conversation have for what audiovisual archives can achieve when they open to the various groups that make up the general public, when they build and maintain metadata in a self-reflective manner, when their material is used to debunk rather than fan exclusion and stereotyping.

Add to this, AI algorithm developers and exploiters are now actively courting national audiovisual archives in order to access their collections to build and train their systems. This poses a new set of complex strategic, legal, reputational and ethical concerns that have been comprehensively laid out by Brecht Declerq, President of the Fédération Internationale des Archives de Télévision / The International Federation of Television Archives (FIAT/IFTA) and Head of Archives at Radiotelevisione Svizzera Italiana, in an article titled "Neck-deep in digital oil? Public broadcaster's archives as AI training datasets."[54] Declerq says that AI developers are prepared to pay significant sums of money to access these archives because of their richness in information and because their data sets are so extensive. While the financial reward is tempting for a sector that struggles with financial investment, Declerq urges broadcasters to consider how any agreement would contribute positively or negatively to the organisation's long-term strategy. He stresses that: 1. It can be difficult to foresee how any harvested information might evolve beyond the initial agreement and taking AI developers to court for any infringement of the agreement is exorbitantly expensive; 2. The agreement could run counter to a public service mandate or rules governing financing; 3. There could be issues of copyright infringement and neighbouring rights regulations in Europe; 4. Generative AI and algorithms are perceived to present a risk in the spread of misinformation and to put journalism jobs under threat; 5. The internal technical development costs to prepare the data sets can be quite high; and 6. Biases, stereotypes and tropes from the past that exist in the archives can fan similar societal divisions today if the algorithms are not designed sensitively and ethically. Declerq says:

> Although the lure of financial benefits may seem attractive, especially for those who are increasingly the victim of heavy budget cuts, significant risks and limitations must not be overlooked. One may wonder whether it is ethically acceptable to work with companies that are not known for their adherence to public values. It is a decision that every broadcaster and every archive can - and I would even say must - make for itself.[55]

8 Conclusion

Archives have a significant contribution to make in the fight against misinformation, disinformation and malinformation but it is a fraught process. We need to be vigilant about the contexts in which stored audiovisual materials were made and what exclusions may have been applied, we need to understand metadata and the possibility of systemic bias. We need to realise we may have limited access and even more limited opportunities of use because of copyright restrictions. If we do all that, precious information can be found in archives that informs society of the past and enables us to contextualise events in the present. With news and information moving at breakneck speed, this contextualisation may well turn out to be vital to make sense of what is happening today. As with trust, when context is lost, misinformation spreads far and wide.

Thanks to the availability of AI tools, we can use archives much better than previously possible. Such tools are maturing to the extent that it is now conceivable to process extensive amounts of content at high speed, content which may, in the near future, become searchable with new, simpler techniques and queries. Using AI also has drawbacks related to ownership, privacy rules and, very importantly, the bias that can be part of how and on what data the AI algorithm is trained. This in turn makes it clear that the use of archival material and who gets access to the archive both need to be discussed widely and thoroughly.

Going over the interviews and the conversations with Antoine Bayet, Willemien Sanders and Johan Oomen, the MediaNumeric team is heartened by their optimism and sense of possibility and purpose. Audiovisual archives are so much more than information warehouses. They enable building a civic counterforce to how media manipulation hurts democratic processes and destroys trust. Yes, we need to understand better how to unlock the archive, we need ways to deal with copyright, we need to be aware of unintended exclusions whether materially (who gets access to the archive, literally) and digitally (how to provide for strong metadata). Despite all of this and because we are learning so much about the role of journalism, public broadcasting and the emancipatory role audiovisual media can play, this is an incredibly exciting adventure.

Acknowledgement

The MediaNumeric project was co-funded by the Erasmus+ Programme of the European Union. To learn more about the project visit **www.medianumeric.eu**.

Grant acknowledgment: The research leading to these results has received funding from the European Commission under grant agreement No. 621610-EPP-1-2020-1-NL-EPPKA2-KA.

Notes

1. The report "The World of Data-Driven Journalism in Storytelling and Fact-Checking" written by Jacqueline Pietsch and Daniel Sorabji as part of the MediaNumeric project and published in April 2024 can be directly downloaded from the MediaNumeric website www.medianumeric.eu here: **http://www.medianumeric.eu/wp-content/uploads/2024/04/Whitepaper_TheworldOfDataDrivenStorytellingInJournalismAndFactChecking.pdf**.
2. EBU, RTVE, Sound & Vision and INA, discussed in this work, have a vested interest in media archiving and in the content created by public broadcasters. There are of course differences between national archives and archives held by public broadcasting organisations themselves. Broadcasters hold the rights and can make audio and video more easily available than that in the national archives which will require a process of copyright identification and clearing before it can be used. While these are significant differences, for the purpose of this paper they are not of the utmost importance.
3. "Video Archiving", WITNESS Archiving, **https://archiving.witness.org/**.
4. Nick Couldry, *The Place of Media Power: Pilgrims and Witnesses of the Media Age* (London: Comedia, 1999).
5. Claire Wardle and Hossein Derakhshan, "Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making," September 27, 2017, *Council of Europe,* **https://rm.coe.int/information-disorder-report-version-august-2018/16808c9c77**.
6. Agence France-Presse, "About Us," date unknown, accessed on July 11, 2024, **https://www.afp.com/en/agency/about/about-us**.
7. Juliette Montesse, "A year of disinformation around the war in Ukraine," *AFP Fact Check*, February 22, 2023, **https://factcheck.afp.com/doc.afp.com.339R4ZG**.
8. Anuj Chopra and Bill McCarthy, "War of narratives: Syrian imagery falsely illustrates Gaza," *AFP Fact Check*, December 29, 2023, **https://factcheck.afp.com/doc.afp.com.349A4RK**.
9. Sami Acef and François D'Astier, "Iranian strikes in Iraq: decontextualized videos abound," *AFP Fact Check*, January 9, 2020, **https://factcheck.afp.com/iranian-strikes-iraq-decontextualized-videos-abound**.
10. Fabien Zamora, "Old brawl footage misrepresented as migrants in Lampedusa Italy," AFP *Fact Check*, September 21, 2023, **https://factcheck.afp.com/doc.afp.com.33VN7JD**.
11. AFP Hong Kong, "This video first appeared in 2014 reports about migrants who drowned off the Libyan coast," *AFP Fact Check*, April 10, 2020, **https://factcheck.afp.com/video-first-appeared-2014-reports-about-migrants-who-drowned-libyan-coast**.
12. Manon Jacob, "Climate misinformation overshadows record floods worldwide," June 14, 2024, *AFP Fact Check*, **https://factcheck.afp.com/doc.afp.com.34WF8VU**.
13. Coalition for Content Provenance and Authenticity, "Overview," 2024, **https://c2pa.org/**.
14. Content Authenticity Initiative, 2024, **https://contentauthenticity.org**.
15. Project Origin, "What Origin does," date unknown, **https://www.originproject.info/**.
16. Coalition for Content Provenance and Authenticity, "Overview," 2024, **https://c2pa.org/**.
17. Rédaction de l'INA, "6 janvier 1975, naissance de l'INA," *L'INA éclaire l'actu,* December 24, 2014, **https://www.ina.fr/ina-eclaire-actu/6-janvier-1975-naissance-de-l-ina**.
18. Antoine Bayet, Conversation with co-author Jacqueline Pietsch, June 6, 2024.
19. Antoine Bayet, idem.
20. Antoine Bayet, idem.
21. Antoine Bayet, idem.
22. Sanders, Willemien, Roeland Ordelman, Mari Wigham, Rana Klein, Jasmijn van Gorp, and J. J. Noordegraaf. "Developing Data Stories in Digital Humanities: Challenges and Protocol." *DH Benelux Journal* 5 (2023): 1-17.
23. Willemien Sanders, Conversation with co-author Jacqueline Pietsch, November 27, 2023.
24. Willemien Sanders, idem.
25. Schofield, B. H. "When sexual abuse was called seduction: France confronts its past," *BBC*, January 19, 2020, **https://www.bbc.com/news/world-europe-51133850**

26. Gabriel Matzneff à propos des adolescentes dans “Apostrophes” | Archive INA,” *INA Clash TV*, December 30, 2019, **https://www.youtube.com/watch?v=TjZmJkLdwN8**, *Matzneff.* (n.d.). Mediaclip. **https://mediaclip.ina.fr/en/i19364016-denise-bombardier-on-the-pedophile-practices-of-gabriel-matzneff.html**, *Réponse de Gabriel Matzneff face à Denise Bombardier.* (n.d.). Mediaclip. **https://mediaclip.ina.fr/en/i19364017-gabriel-matzneff-s-answer-to-denise-bombardier.html**
27. The same thing happened in 2023 when France’s film icon Gérard Depardieu was accused of sexual misconduct/aggression/violence. Le Parisien, “Affaire Depardieu : l’acteur visé par une nouvelle enquête pour agression sexuelle, après la plainte d’une décoratrice,” *Le Parisien,* March 5, 2024, **https://www.leparisien.fr/faits-divers/affaire-depardieu-lacteur-vise-par-une-nouvelle-enquete-pour-agression-sexuelle-apres-la-plainte-dune-decoratrice-05-03-2024-IXHZE4SETFFOXGIHNGDNDVHA7E.php**
28. Antoine Bayet, idem.
29. Antoine Bayet, idem.
30. Willemien Sanders, idem.
31. Willemien Sanders, idem.
32. Willemien Sanders, idem.
33. Yvonne Ng, MediaNumeric interview, April 6, 2021.
34. Yvonne Ng, idem.
35. Willemien Sanders, idem.
36. Iberspeech, “Albayazin Evaluation Challenge,” 2024, **https://iberspeech.tech/albayzin-evaluation-challenge**.
37. Albayzín Evaluations, “Spanish Thematic Network on Speech Technologies (RRTH),” date unknown, **https://catedrartve.unizar.es/albayzin.html**.
38. Trusted European Media Data Space, 2023, **https://tems-dataspace.eu/**.
39. Virginia Bazán-Gil, MediaNumeric interview, May 14, 2021.
40. Virginia Bazán-Gil, idem.
41. BBC, “Google apologises for Photos app’s racist blunder,” *BBC*, July 1, 2015, **https://www.bbc.com/news/technology-33347866**.
42. BBC, “Facebook apology as AI labels black men ‘primates’,” *BBC*, September 6, 2021, **https://www.bbc.com/news/technology-58462511**.
43. Cade Metz, “Who Is Making Sure the A.I. Machines Aren’t Racist?,” *New York Times*, June 23, 2023, **https://www.nytimes.com/2021/03/15/technology/artificial-intelligence-google-bias.html**.
44. Alexandre Rouxel, Conversation with co-author Jacqueline Pietsch, November 27, 2023.
45. Alexandre Rouxel, idem.
46. Alexandre Rouxel, idem.
47. Hélène Rauby Matta, MediaNumeric interview. April 23, 2021.
48. Hélène Rauby, idem.
49. Roeland Ordelman, Julia Noordegraaf, Jasmijn van Gorp, Mari Wigham and Mary-Joy van der Deure, “CLARIAH Media Suite: from 2014 to 2024 in a series of short stories,” *Zenodo*, June 24, 2024, **https://zenodo.org/records/12517693**.
50. Willemien Sanders, idem.
51. Yvonne Ng, idem.
52. Johan Oomen, Conversation with co-author Jacqueline Pietsch, May 29, 2024.
53. Johan Oomen, idem.
54. Brecht Declerq, “Neck-deep in digital oil? Public broadcaster’s archives as AI training datasets,” *FIAT/IFTA*, June 8, 2024, **https://fiatifta.org/broadcast-archives-as-datasets/**.
55. Declerq, “Neck-deep in digital oil?”

Biography

Jacqueline Pietsch works as a journalist at the French global news agency Agence France-Presse (AFP). Since joining AFP in 1996, she has worked in a myriad of roles, disciplines and countries, including head of AFP’s English-language TV production, correspondent in the Netherlands and video journalist responsible for Southeast Asia. Jacqueline currently oversees technical developments for AFP’s fact-check team.

VIEW Journal of European Television History and Culture Vol. 13, 25, 2024
URL: **https://doi.org/10.18146/view.334**
Publisher: Netherlands Institute for Sound and Vision in collaboration with Utrecht University, University of Luxembourg and Royal Holloway University of London.