Web Archives as a Source of Data for Critical and Inclusive Uses of AI in History

Sophie Gebeil

21/10/2025

Texte de la conference donnée le 21/10/2025 et diaporama utilisé pour la presentation.

Web Archives as a Source of Data for Critical and Inclusive Uses of AI in History

Thank you to Nadescka and all the organisers for giving me the opportunity to share my research with you today

I am Sophie Gebeil, I am A researcher in Web history at the Aix Marseille University in the south of France in which I am also the Vice-President Delegate for Human and Mediterranean Sciences. The main focus of my research is how the past is portrayed online through the French Web Archives. And I am also interested in the epistemological and methodological implications of using web archives as historical material.

The European program Polyvocal Interpretation Of Contested Colonial Heritage (PICCH) which was concluded in 2024, aimed at exploring, how archival documents created from a colonial perspective could be reappropriated and reinterpreted to become an effective source for constructing an inclusive future society. In France, the term 'decolonization' has been heavily instrumentalized, losing the meaning attributed to it by historical thinkers like Achille Mbembe. In this project, the idea of "decolonizing the archives" is to render the materials from former colonial powers more inclusive and respectful towards populations still facing discrimination today, challenges that have been driving archivists worldwide for years (Ghaddar & Caswell, 2019). One of the project's objectives was to refine the metadata of television archives as well as web data concerning narratives of events related to the colonial past or post-colonial issues. We scrutinized the media coverage of the 1983 March for Equality and Against Racism in France, from a transmedia perspective, based on web and television video corpora from the INA. One of the goals was to examine the visibility accorded to the marchers by the media. Indeed, in 1983, they were young suburbanites, born to immigrant parents, perceived as Maghrebi or Black and this led to an essentialization of the mediatic discourse. In spite of these young people having been the principal initiators of the March, they were quickly relegated to the periphery of the journalistic narrative from the 1980s until around 2013, with the media focusing on the white figure of the Père (Father) Delorme and the creation of the mainstream SOS racism association in 1984.

Given the volume of data (such as archived web pages, voice-over text from videos, and video metadata), we employed AI programs to automate the identification of the marchers, both through text (so names, nicknames) and facial recognition in the videos.

Based on this case study, I will deliberately avoid using the interpretation of the online media coverage of the march and will concentrate more on the methodological and hermeneutical questions raised by cultural biases when employing deep learning AI

programs to analyse web data. **It seeks to investigate under what conditions the application of AI programs to analyse archived web data can render these historical actors more visible.**

Firstly, I will present the research topic, its key issues, and the corpus that has been assembled.

Secondly, I will focus on the methodology used to study the media coverage of the marchers in the television and web archives of the National Audiovisual Institute, with particular attention to how AI is employed to analyse both textual and visual content from the videos.

Finally, I will discuss the insights gained from the study and reflect on the lessons learned from using AI to process web data, especially in relation to hegemonic biases.

I. Remembering the 1983 March for Equality and Against Racism in France through Television and Web Archives

Before delving into the analysis itself, I would like to begin with a brief overview of the 1983 March for Equality and Against Racism in France, often referred to as the "Marche des Beurs." And, since not everyone may be familiar with the web archiving domain, I will also briefly introduce what web archives are and why they are increasingly important for historical and media research. This research is therefore situated within the field of contemporary cultural history, while also contributing to the growing domain of web archiving studies.

I.1. 1983 March for Equality and Against Racism

The triumphant arrival of the March for Equality and Against Racism on December 3, 1983, in Paris is, in many respects, a significant event in contemporary France. Studying it enables a better understanding of the identity tensions that agitate present society. From a media perspective, it signifies the emergence of the second generation of post-colonial immigration, previously considered a temporary phenomenon. The press and cameras focused particularly on these children of North African immigrant workers born in large housing estates, who had become young adults denouncing racist crimes and, more broadly, the mechanisms of exclusion they faced. The Maghreb-focused lens led journalists to label this unprecedented anti-racist initiative as the "beurs' march", a designation imbued with colonial heritage and reductionism. Indeed, the March is also a

post-colonial event in the sense that the violence induced by the Algerian War (1954–1962) still lingers.

Furthermore, the March represents a significant moment in the immigrant and anti-racist social movement. It originated in the working-class neighborhood of Minguettes in Lyon against a backdrop of tensions between the local youth and law enforcement, whip up (whèp up) by a surge in racist crimes in France (Hajjat 2013). Toumi Djaïdja, then 19 years old and president of the association Avenir Minguette, was injured by arbitrary police gunfire and hospitalized. This incident sparked the idea of crossing France to denounce racist violence and demand better treatment for immigrants and their children. Inspired by figures such as Gandhi (1930), Martin Luther King (1963), and the Larzac farmers (1978), the March pick up support from Father Christian Delorme and the CIMADE network (Inter-Movement Committee for Evacuees) from its inception. The first seventeen marchers gathered in Marseille on October fifteenth,2023, welcomed by local support committees. They journeyed across France until reaching Paris, where they were received at the Élysée Palace by President François Mitterrand, who pledged to grant a 10-year residency permit for immigrant workers (Hajjat 2013). The following year saw the establishment of the SOS Racisme association, following the lead of the Socialist Party, one of the historical anti-racism associations behind the memorable 1985 concert. The "Don't Touch My Buddy" badge marked an entire generation, but for the marchers and anti-racist activists, SOS Racisme was criticized as a political exploitation of the March, embodying the unfulfilled promises made by the Socialist Party to the inhabitants of working-class neighborhoods.

A complex and divisive event, the March remained largely overlooked in the 1990s and 2000s. It was not until its thirtieth anniversary that celebrations began to emerge, including several exhibitions and, notably, the 2013 film La Marche by Nabil Ben Yedir. Following an initial qualitative study highlighting the ambivalences of the memory of this event, which oscillated between nostalgia and bitterness (Gebeil 2013), we aimed to study its representations in audiovisual media and on the web within the PICCH project.

II.2. Web Archiving Research Domain

To examine how audiovisual representations of this event have evolved since 1983, we drew on both French television archives and web archives.

Web archiving refers to the process of collecting, preserving over the long term, and providing access to web pages. It is almost as old as the web itself, as illustrated by the creation of the Internet Archive in 1996 and its well-known Wayback Machine. Since 2001, the European Commission has encouraged member states to establish national web archiving initiatives, and in 2003, UNESCO officially recognized web pages as part of our digital heritage.

Although web archiving is not a new practice, it remains underused by historians, despite a growing body of transnational research, exemplified by the RESAW network. Yet, web archives are essential for historians working with the web as a source to study recent phenomena. I refer to this approach as born-digital history.

Web archives are crucial for several reasons. First, they ensure the citability of born-digital sources. For example, if you cite a URL from the live web and that link becomes inactive, your argument may be weakened. Being able to access that URL—even ten years later—is vital. Citing the archived version strengthens both the traceability of your sources and the evidentiary basis of your claims. In historical research, sources must be accessible and verifiable; we are not simply to be taken at our word.

Second, web archives offer a rich repository of sources. They allow historians to retrieve previous versions of now-defunct websites or to study how a single site has evolved over time. Unlike screenshots or PDFs, archived websites are navigable, preserving the structure and interactivity of the original.

Finally, web archives constitute our future digital heritage. As such, they raise important questions about the preservation of tomorrow's sources—questions that directly concern us as historians.

However, using web archives is not without challenges. They are not exact mirrors of the live web, but rather reconstructions of web pages based on what was captured. This is why we speak of reborn digital heritage. Technical limitations in the capture process can result in errors or missing elements, making these archives difficult to work with.

The research community has invested significant effort in adapting tools originally designed for the live web—such as URL graph analysis—for use in the closed and fragmented environment of web archives. This adaptation process reveals what I call the cost of entry to working with web archives, a barrier that may discourage many historians from engaging with these materials.

This is precisely why I have recently felt the need to reflect more deeply on how historians use web archives, in all their complexity. I believe this engagement challenges some of the core concepts of our discipline—particularly our relationship to sources—and that it is through our encounters with web archives that we begin to grasp the broader epistemological shift brought about by the digital turn.

To return to today's topic, for this project I worked with the French web archives. In France, two institutions have been responsible for archiving the national web since the introduction of legal deposit for online content in 2006. The INA (National Audiovisual Institute) archives media-related websites, including those connected to radio, television, cinema, and digital audiovisual platforms, while the BnF (National Library of France) is in charge of archiving other types of websites.

For this study, we collaborated with the team in charge of the legal deposit of media websites at INA, as well as with the INA Lab. The INA Lab supports research on media by providing access to INA's collections and facilitating quantitative analysis and automated processing of audiovisual data.

Let me now turn to the corpus used in this study.

It brings together a selection of audiovisual and web materials that allow us to trace how the 1983 March has been represented and remembered over time.

I.3. Corpus

The study corpus is composed of three types of sources:

- Television archives,
- Born-digital web content, and
- Archived web sources.

It spans a 40-year period, from 1983 to 2023.

A television corpus was created in 2022 using materials from the French National Audiovisual Institute (INA). This qualitative corpus includes 465 programs broadcast on French television and web media, all preserved under the legal deposit framework.

Here, you can see a graph showing peaks in media visibility for each decade since 1983.

This qualitative corpus was then complemented by a large-scale web corpus drawn from the INA's web media legal deposit, using a textual search strategy. After deduplication and cleaning, we retained around 25,000 web pages. This part of the corpus covers the period from 2003 to 2022.

To ensure coverage of the 40th anniversary of the March in 2023, we also created a live web corpus using Hyphe, a web crawling tool developed by the Sciences Po MediaLab.

Finally, this collection also includes a Twitter dataset, captured just before the end of free access to the Twitter API in 2022. All of these materials—television, web, and social media—are now archived by the INA.

Let us now move on to the second part of this presentation, which focuses on the methodological challenges we encountered.

In particular, I will discuss how we approached the analysis of both textual and visual data across time, using tools from computational humanities.

I will also present the neural topic modeling approach we adopted to explore the evolution of media discourse, and reflect on its relevance and limitations for historical inquiry.

II. Methodological Challenges: Textual and Visual Data Across Time. Neural Topic Modeling Approach for Historical Inquiry

Based on this corpus, our objective was to study how media coverage of the 1983 March evolved over time, with particular attention paid to the visibility of the marchers themselves. And the sheer volume of data made it necessary to automate parts of the analysis to meet this goal.

So, we had to overcome several challenges. First, the corpus was transmedia, combining television and web archives. Second, it was composed largely of video content. And third, we wanted to examine diachronic change—that is, to approach the material from a historical perspective.

We therefore designed an experimental protocol in collaboration with the INA Lab based on two types of data from video corpus: visual and textual data.

The first data treatment pipeline focused on visual similarity using Snoop, a tool already used by the INA Lab. The processing chain used in this software involved segmenting videos into still frames, extracting key visual features, and allowing users to retrieve segments like a given source excerpt. This tool allows for targeted searches within the corpus, such as retrieving specific images or excerpts. For example, I can search for images showing Toumi Djaïdja, who initiated the 1983 March after being hospitalized due to a police shooting. Similarly, I can look for footage of the Arc de Triomphe in my video archive corpus. The latter made it possible to sift through the visual media excerpts / that the automated search engine had gathered[1].

Snoop has two key features:

It uses an indexing structure that enables real-time search across massive image databases.

[1] And it was possible to search for Visual media excerpts were made searchable through the Snoop visual search engine (see Joly & Buisson, 2008, and subsequent work: https://www.ina.fr/institut-national-audiovisuel/equipe-recherche/projet-snoop. and the latter made it possible to sift through the visual media excerpts / that the automated search engine had gathered
Afficher moins

It incorporates active learning, allowing users to refine results by annotating relevant and irrelevant matches.

It's important to note that the feature extraction process—used to identify visual similarities—is powered by an open-source neural network model trained on large-scale image datasets external to our corpus. The model used here is Inception V3

For the textual data from videos videos, our goal was to facilitate the analysis of the discourse surrounding the media coverage of the 1983 March. To do this, we applied a two-step pipeline, combining deep learning-based tools for automatic speech recognition (WhisperX) and neural topic modeling (BERTopic), and we examine how media narratives surrounding the 1983 March for Equality and Against Racism have evolved over four decades. Today, I will go into more detail about this pipeline, as it represents a key methodological achievement of this research. Since the approach is replicable and adaptable to other audiovisual corpora in historical research, it will be presented in an upcoming special issue on AI and History proposed by the journal Digital History.

Among the many available tools and methods for topic modelling, we chose the one implemented in the Python library BERTopic (Grootendorst, 2022), primarily for its ease of use and its built-in visualization tools. This work was initially explored by Davide Rendina in collaboration with our team, and later developed further with Arthur Lezer from the INA Lab. For those who are unfamiliar, let me precise that Bert Topic is an artificial intelligence of the NLP (Natural Language Processing) type, i.e. a set of machine learning techniques applied to natural language processing. It combines multiple AIs and models for language understanding, clustering, dimensionality reduction for visualization, and Language Statistics for keyword extraction. Topic modelling technique automatically identifies recurring lexical patterns—called topics—within the transcripts of our corpus.

We transcribed and analyzed a corpus of 150 hours of French television and web media archives (145 TV broadcasts and 320 web videos) preserved by the French National Audiovisual Institute (INA).

For each television and web video, speech was transcribed using the WhisperX automatic speech recognition system (Bain et al., 2023). Then, we compared several variants of the BERTopic pipeline for the task of dynamic topic modelling on ASR-generated text. Most notably, we assessed the impact of several variants of pre-trained BERT neural models on topic quality, both qualitatively and quantitatively, using coherence and diversity metrics. This comparative experimentation constitutes a methodological contribution to AI-assisted historical research. Moreover, by systematically evaluating different configurations of the BERTopic pipeline—such as embedding models, clustering

parameters, and dimensionality reduction techniques—this study demonstrates how algorithmic choices shape the interpretability and granularity of historical insights. We also investigate the potential of BERTopic's visualization capabilities to serve as a 'corpus explorer' for an audiovisual collection over time. At the hermeneutics layer, this takes the form of Python code to perform the following steps : transcribe the A/V media, fit the Bertopic model on the transcribed text, compare the performance of BERT models, generate Plotly visualisations that allow users to navigate the topics in the form of clusters of documents or time-periods, along with archive extracts and metadata.

By doing this, we investigated how AI can help surface marginalized narratives and support historians in navigating complex, transmedia corpora.

Let us now turn to the results obtained using machine learning models and AI programs to process this data.

III. Results: Mapping Media Narratives of the 1983 March — A Distant Reading Approach to Inclusive Historical Inquiry

These results highlight, first, the potential of distant reading techniques to navigate large-scale trends in the media discourse surrounding the 1983 March.

It also invites us to reflect critically on the contributions and limitations of AI-based methods when it comes to writing histories that take into account marginalized or minoritized populations.

III.1. Biases and Mitigation Strategy

In our use of Snoop, we observed a clear bias: facial recognition tended to work more effectively for white individuals. We thought that this was a symptom of bias in the training data used for the V3 Inception model, which was trained on large-scale, royalty-free image datasets—primarily from platforms like Flickr, as discussed in Fuica and Lezer (2023).

As for the textual corpus, we encountered two layers of bias. Firstly, within the corpus itself where the names of Arabic or North African origins often lacked standardized spelling, which complicated automated identification—for example, Toumi Djaïdja. Second, language models are typically less trained on such names or terms, as they are underrepresented in French-language corpora.

To address this, we created a curated entity directory, manually reviewed and corrected by interns working on the project. This task, while time-consuming, proved to be highly instructive, helping them to better understand how machine learning models work—as well as of course the concept of digital labor (Caselli).

This dictionary has become a valuable tool in its own right: it helps document the March, highlights the presence of the marchers, and improves model performance.

III.2. Interpreting the Results: Visibility, Bias, and Contested Memory

In terms of media coverage, our collaboration with the INA Lab allowed us to explore the corpus in new ways. The HTML files enabled us to visualize how automatically generated themes evolved over time, and to link them directly to audiovisual archive references.

The data visualisation revealed a growing media presence of the marchers. For instance, Toumi Djaïdja remained a marginal figure in the corpus until the 30th anniversary, and gained more visibility around the 40th.

However, this increased visibility did not necessarily reflect sympathetic coverage—nor consensus around the March's demands that was to denounce racist violence, economic and social exclusion in the suburbs, and systemic discrimination.

On the contrary, our analysis shows that—despite the celebratory tone of 2013, which framed the March as a reclaimed event by SOS Racisme and the Socialist Party—it remains a deeply polarizing historical moment. It continues to serve as a projection surface for controversies around Islam, the marginalization of working-class neighborhoods, and police violence. As a result, the original demands of the March are often either historicized as past concerns or drowned out by present-day political tensions.

Taking stock of the March's legacy—acknowledging it as a movement born in the working-class suburbs, led by young people of immigrant descent, victims of specific forms of violence and discrimination, including from the police, and claiming full belonging to the French nation—remains, even after 2013, a contested historical narrative.

Moreover, when the discussion shifts to the current situation of marginalized communities—religious or racial discrimination, or police violence disproportionately affecting them—it often triggers radical or even hateful discourse, which hinders any meaningful reflection on the March's significance.

While numerous documentaries have been produced and the event has gained visibility, the core demands of the 1983 marchers still face strong resistance in the political sphere.

III.3. Reflections on Historical Methodology and Epistemology : Designing a Replicable Pipeline for Diachronic Analysis of Large-Scale Audiovisual Archives

From the standpoint of historical methodology, the project yields a significant result through the development of a replicable pipeline for the diachronic analysis of large-scale audiovisual corpora, enabling both systematic exploration and critical engagement with long-term media narratives.

These interactive visualizations generated by BERTopic serve as a powerful exploratory interface for historians by mapping topic distributions over time and across clusters of documents. They facilitate a form of distant reading of media discourse that remains anchored in archival specificity. Thus, a key contribution of this study lies in the methodological transparency and reproducibility of its AI-assisted analytical pipeline. Each step—from speech transcription to topic modeling—is thoroughly documented and guided by quantitative evaluation metrics, enabling both interpretability and replicability in the analysis of large-scale audiovisual corpora.

This pipeline enables historians to navigate their archives differently, combining distant reading with qualitative video analysis.

Furthermore, the project contributes to ongoing reflections on how artificial intelligence is transforming historians' practices in source analysis. It thus engages with broader epistemological questions about the changing nature of knowledge production in the digital age. This dimension must be critically addressed to ensure a reflective and responsible use of AI in historical research. This leads me to the third point.

IV. Lessons to be learnt and perspectives

The PICCH project offers valuable insights—not only into the history of French society and its relationship with racism, but also into how historians might critically engage with the idea of "decolonizing archives." While this concept holds meaning in the context of heritage and memory, it cannot be applied mechanically. For historians, the goal is not to erase the past, but to better understand it through a critical lens.

That said, we would like to emphasize here the potential and limitations of using AI when working with data related to minority groups. For any media historian considering the use of AI in this context, several key lessons emerge.

The value of this work is significant. The pipeline we developed is replicable and offers a new way of navigating the corpus—enabling a distant reading of audiovisual archives. Topic modeling, in particular, allows us not only to trace the presence of the marchers over time, but also to identify the themes associated with their representation.

This makes it a powerful tool for historians working with archived web data. It helps identify major trends and provides visualizations that can be cross-referenced with qualitative analysis of the videos themselves.

However, several precautions must be taken. First, we must remain aware that both the corpus data and the training data tend to amplify socio-cultural biases: individuals from lower-income backgrounds and non-white populations are less accurately identified.

To mitigate this bias, significant human involvement and time are required to improve the data and, ideally, to retrain models on more representative datasets. Another option is to draw on the state of the art to identify models that have already been trained on such data—though, by definition, these are relatively rare.

In addition, data processing must be documented with full transparency to ensure the reproducibility of analyses, especially in interdisciplinary contexts.

This also opens up broader reflections—particularly when working with multilingual corpora or languages that use non-Latin scripts, such as Arabic. These challenges raise important questions about how experimental approaches are recognized within the historical discipline. For instance we should using LLM and Generative AI in order to Fine-tune BERTopic.

Traditionally, the field has valued contributions to knowledge and scholarly erudition. But to meet these new challenges, the discipline must also recognize **methodological work**—work that redefines disciplinary approaches and benefits the entire scholarly community.

Conclusion

To conclude, this interdisciplinary study contributes to a deeper understanding of the audiovisual media coverage of the 1983 March for Equality and Against Racism. It also offers a methodological pipeline that enables new ways of navigating audiovisual corpora from a diachronic perspective. More broadly, it contributes to a critical engagement with the use of AI in historical research.

These promising results should not obscure the substantial effort required to overcome the high entry threshold faced by interns and early-career researchers involved in the project. A significant amount of time was devoted to explaining to students both the value and the limitations of web archiving, as well as to developing a methodology that would actually work.

At the end of their internships, several students remarked that “history has really changed!”—a telling sign of how such projects can reshape historical training and practice.

It underscores the need for transparent, critically informed methodologies capable of supporting both scholarly communities and student training in an era where generative AI is becoming increasingly embedded in everyday academic routines.

Since completing this project, we have been reflecting more deeply on the epistemological challenges that arise when historians engage with web archives. As part of my work with the Institut Universitaire de France, I am now leading a research project that explores the historiographical implications of web archiving. By examining both academic and non-academic practices, this project contributes to the renewal of theoretical and methodological frameworks—particularly through the integration of computational approaches.

It is in this context that you received a questionnaire. From the Maison Méditerranéenne des Sciences de l’Homme at Aix-Marseille University, my colleague and I have established the WebLab—a space dedicated to research and training on the history of digital media and web archiving. We offer tailored workshops for researchers in history, sociology, and other social sciences working on the Mediterranean region. The WebLab also serves as a hub for current developments in the field.

— End –

Annex

Slideshow

Web Archives as a Source of Data for Critical and Inclusive Uses of AI in History

Sophie Gebeil
Aix Marseille University (TELEMMe Laboratory) – WebLab MMSH
sophie.gebeil@univ-amu.fr
https://madi.hypotheses.org/

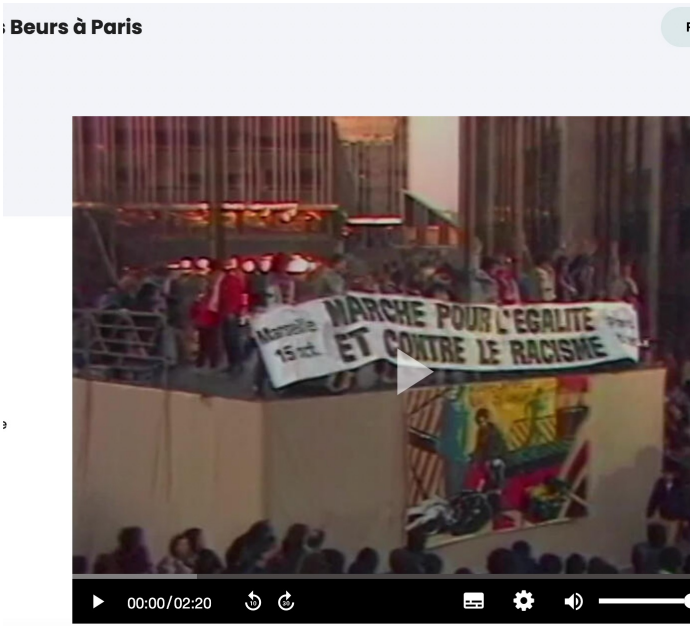
I. Remembering the 1983 March for Equality and Against Racism in France through Television and Web Archives

Context and Research Focus

Polyvocal Interpretation of Contested Colonial Heritage

European program Polyvocal Interpretation Of Contested Colonial Heritage (PICCH) – 2021/2024 dir. D. Petrelli (U. Sheffield). PI France> S. Gebeil Aim: Reinterpreting audiovisual colonial and postcolonial archives for inclusive historical narratives Context: A historical approach to the concept of "Decolonizing Archives"	Case study: 1983 March for Equality and Against Racism Media analysis: Television and web archives (INA) Objective: Examine visibility and representation of the event, and the marchers

https://web.archive.org/web/20240629211107/https://picch-project.org/



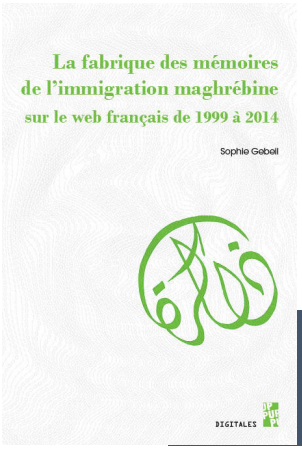
Marche des Beurs à Paris – March of « Beurs » in Paris, Television Archive, https://enseignants.lumni.fr/fiche-media/00000000444/arrivee-de-la-marche-des-beurs-a-paris.html

Case study: 1983 March for Equality and Against Racism

- Media analysis: Television and web archives (INA)
- Objective: Examine visibility and representation of the event, and the marchers

Web Archive as Historical Sources

Doctoral Thesis : *The Digital Construction of Maghrebian Immigration Memories on the French Web (1999–2014)*
Defended in 2015 – AMU



RESAW - a Research Infrastructure for the Study of Archived Web Materials

Home

RESAW, a Research Infrastructure for the Study of Archived Web Materials, is a community established in 2012, aiming at promoting a collaborative European research infrastructure for the study of archived web materials.

Want to host the RESAW Conference 2029?

The RESAW Conferences are looking for a research institution to host the RESAW Conference in 2029.

Contact

In case of questions, comments, or suggestions please contact coordinator of the RESAW community, Professor Niels Brügger, Aarhus University, Denmark.

Web Archive as Historical Sources

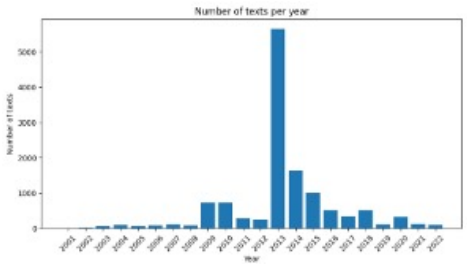
Web archiving is the process of collecting portions of the World Wide Web, preserving the collections in an archival format, and then serving the archives for access and use. (IIPC Web site, http://netpreserve.org/web-archiving/)

"Web archiving is the process of gathering up data that has been recorded on the World Wide Web, storing it, ensuring the data is preserved in an archive, and making the collected data available for future research." (Niu, Jinfang. 2012. "An Overview of Web Archiving." *D-Lib Magazine* 18 (3/4). https://doi.org/10.1045/march2012-niu1.)

In France, the Web Legal Deposit Low (2006) > the French national Library and the National Audiovisual Institute are allowed to archive the national Web

An international network and an interdisciplinary community > RESAW, IIPC

A Transmedia Corpus : Television, Web Archives and Social media



Corpus:

- 465 TV programs (1983–2023)
- 25,000+ archived web pages (2003–2022)
- live web corpus
- Twitter dataset

II. Methodological Challenges: Textual and Visual Data Across Time

Neural Topic Modeling Approach for Historical Inquiry

Visual Data Processing with Snoop

- Tool: Snoop (INA Lab)
- Process:
- - Video segmentation
- - Visual feature extraction
- - Search by image (e.g., Toumi Djaïdja, Arc de Triomphe)
- e.g. searching for images showing Toumi Djaïdja, who initiated the 1983 March after being hospitalized due to a police shooting.

Methodological Framework

Challenges in Data Analysis :

- Transmedia and diachronic complexity
- Cultural and technical biases in AI: Facial recognition less effective for non-white individuals / NLP struggles with non-standardized names

AI-Driven Topic Modeling on Audiovisual Archives

Corpus:
150 hours of French TV and web media (145 TV broadcasts, 320 web videos) from INA archives
Step 1 – Transcription:
Automatic Speech Recognition (ASR) using **WhisperX** (Bain et al., 2023)
Step 2 – Topic Modeling:
Comparative testing of **BERTopic** variants for dynamic topic modeling on ASR-generated text
Evaluation:
Assessed impact of different **pre-trained BERT models** using coherence and diversity metrics
Visualization:
Used **Plotly** to create interactive topic maps over time, enabling corpus exploration with metadata and archive excerpts
Historical insight:
Explores how AI can help **surface marginalized narratives** and support historians in navigating complex transmedia corpora
Dissemination:
Methodology to be published in *Digital History*'s special issue on **AI and History**

III. Results: Mapping Media Narratives of the 1983 March — A Distant Reading Approach to Inclusive Historical Inquiry

Textual Analysis Pipeline: From Speech to Topics

- **Objective**: Analyze the evolution of media discourse on the 1983 March over four decades
- **Two-step pipeline**:
  - **WhisperX**: Deep learning-based **automatic speech recognition** to transcribe video content
  - **BERTopic**: **Neural topic modeling** to detect and visualize recurring themes in the transcripts
- **Why BERTopic?**
  - Easy to use, with built-in visualization tools
  - Combines multiple AI models: NLP, clustering, dimensionality reduction, keyword extraction
- **Collaborations**:
  - Initial work by **Davide Rendina Rendina (2024) >** https://doi.org/10.36253/979-12-215-0413-2.22
  - Further developed with **Arthur Lezer** (INA Lab) > paper forthcoming in *Journal of Digital Hystory*
- **Relevance**:
  - Methodology is **replicable** for other historical audiovisual corpora
  - To be published in the *Digital History* journal's special issue on **AI and History**

Visual Data Processing with Snoop

- Tool: Snoop (INA Lab)
- Facial recognition tended to work more effectively for white individuals

Two layers of bias in the textual corpus:

- Corpus-level bias: Non-standardized spellings of Arabic or North African names (e.g., Toumi Djaïdja) complicate automated identification.
- NLP struggles with non-standardized names: Arabic and North African names often appear with inconsistent spellings (e.g., Toumi Djaïdja), making automated recognition difficult and amplifying model bias due to underrepresentation in training data.
- Model-level bias: Language models are less trained on underrepresented names and terms, limiting their ability to recognize and process them accurately.

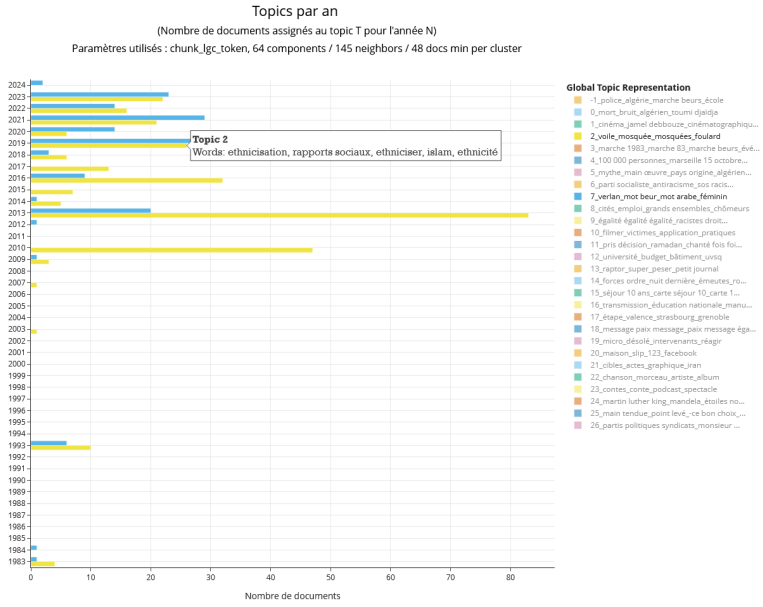
Interpreting the Results: Visibility, Bias, and Contested Memory

- **Increased media visibility** of the marchers over time
  - Example: *Toumi Djaïdja* remained marginal until the 30th anniversary; gained visibility around the 40th
- **Visibility ≠ positive recognition**
  - Greater presence does not imply sympathetic coverage or consensus on the March's original demands
- **Persistent polarization**
  - 2013 commemorations framed by SOS Racisme and the Socialist Party
  - The March remains a contested event, tied to debates on Islam, police violence, and suburban marginalization
- **Dilution of original demands**
  - Calls against racist violence and systemic discrimination often historicized or overshadowed by current political tensions
- **Legacy remains contested**
  - The March's identity as a grassroots movement led by youth of immigrant descent is still debated
- **Contemporary resonance triggers backlash**
  - Discussions on ongoing discrimination and police violence often provoke radical or hostile reactions, hindering historical reflection

Sophie Gebeil24

Contribution to Historical Methodology

Designing a Replicable Pipeline for Diachronic Analysis of Large-Scale Audiovisual Archives



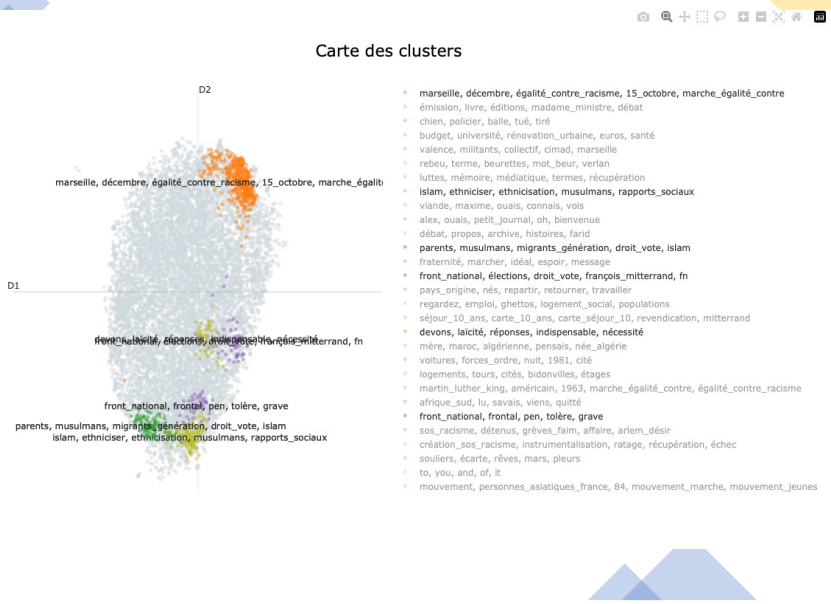
Mitigation Strategy

1) With D. Rendina (2022-2023)
Manual entity directory creation
BERTopic on web archived corpora
Creation of a curated entity directory, manually reviewed by research interns

Named Entities dictionary provided for filtering and normalization (Rendina, 23)

Do NOT CHANGE		to validate		
entity original (dictionary)	cnt	entity noramlized (dictionary)	type	flag
'. I. A. Pentago (LOC)'	1	. I. A. Pentago	LOC	
'1 (LOC)'	1	1	LOC	
'10 (LOC)'	4	10	LOC	
'10e arrondissement de Paris (LOC)'	4	10e arrondissement de Paris	LOC	
'10ème arrondissement de Paris (LOC)'	1	10ème arrondissement de Paris	LOC	
'11 (LOC)'	2	11	LOC	
'11e arrondissement de Paris (LOC)'	2	11e arrondissement de Paris	LOC	
'11eme arrondissement de Paris (LOC)'	2	11eme arrondissement de Paris	LOC	
'11ème arrondissement de Paris (LOC)'	1	11ème arrondissement de Paris	LOC	
'11ème circonscription des Français de l Etranger (LOC)'	1	11ème circonscription des Français de l Etranger	LOC	
'12e arrondissement de Paris (LOC)'	1	12e arrondissement de Paris	LOC	
'13 (LOC)'	1	13	LOC	
'13 arrondissement de Paris (LOC)'	1	13 arrondissement de Paris	LOC	
'13 Nord (LOC)'	1	13 Nord	LOC	
'13e arrondissement de Paris (LOC)'	1	13e arrondissement de Paris	LOC	
'13e circonscription de Paris (LOC)'	1	13e circonscription de Paris	LOC	
'13e circonscription des Hauts de Seine (LOC)'	1	13e circonscription des Hauts de Seine	LOC	

2) With Arthur Lezer (INA Lab) 2023-2024 :
- Transcription of video corpus (WhisperX) and Neural Topic Modeling
- Then we compared several variants of the BERTopic pipeline for the task of dynamic topic modelling on ASR-generated text
> Improved model performance and documentation of the March


Carte des clusters
D2
D1
marseille, décembre, égalité_contre_racisme, 15_octobre, marche_égalité
front_national, frontal, pen, tolère, grave
parents, musulmans, migrants_génération, droit_vote, islam
islam, ethniciser, ethnicisation, musulmans, rapports_sociaux
marseille, décembre, égalité_contre_racisme, 15_octobre, marche_égalité_contre
émission, livre, éditions, madame_ministre, débat
chien, policier, balle, tué, tiré
budget, université, rénovation_urbaine, euros, santé
valence, militants, collectif, cimad, marseille
rebeu, terme, beurettes, mot_beur, verlan
luttes, mémoire, médiatique, termes, récupération
islam, ethniciser, ethnicisation, musulmans, rapports_sociaux
viande, maxime, ouais, connais, vois
alex, ouais, petit_journal, oh, bienvenue
débat, propos, archive, histoires, farid
parents, musulmans, migrants_génération, droit_vote, islam
fraternité, marcher, idéal, espoir, message
front_national, élections, droit_vote, françois_mitterrand, fn
pays_origine, nés, repartir, retourner, travailler
regardez, emploi, ghettos, logement_social, populations
séjour_10_ans, carte_10_ans, carte_séjour_10, revendication, mitterrand
devons, laïcité, réponses, indispensable, nécessité
mère, maroc, algérienne, pensais, née_algérie
voitures, forces_ordre, nuit, 1981, cité
logements, tours, cités, bidonvilles, étages
martin_luther_king, américain, 1963, marche_égalité_contre, égalité_contre_racisme
afrique_sud, lu, savais, viens, quitté
front_national, frontal, pen, tolère, grave
sos_racisme, détenus, grèves_faim, affaire, arlem_désir
création_sos_racisme, instrumentalisation, ratage, récupération, échec
souliers, écarte, rêves, mars, pleurs
to, you, and, of, it
mouvement, personnes_asiatiques_france, 84, mouvement_marche, mouvement_jeunes

IV. Lessons to be learnt and perspectives

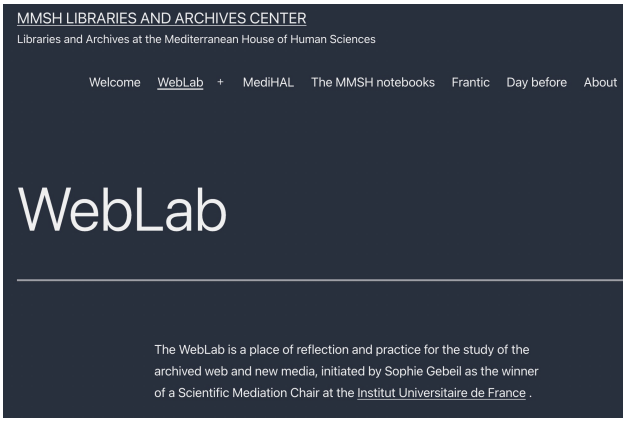
- A new way to explore an audiovisual and web corpora over time
- Distant reading of large corpora - Topic modeling
- Importance of human oversight and bias correction
- Ethical and Methodological Reflections
  - AI must be used critically and transparently
  - Need for representative training data
  - Recognition of experimental approach in historical researcher carreer
- AI can support inclusive historiography—but not without risks
- Future directions: Fine-tuning models, multilingual corpora, interdisciplinary collaboration

The WebLab is a space for both critical reflection and practical engagement with web, archived web and new media, based at the Maison Méditerranéenne des Sciences de l'Homme (Aix-en Provence, France)

WebLab (MMSH, Aix-Marseille University): research & training on digital media history & web archiving. https://pba.mmsh.fr/?page_id=1465

Upcoming guests: **Jefferson Bailey** (Internet Archive), **Ian Milligan**.

Subscribe to our mailing list! https://listes.services.cnrs.fr/wws/info/weblab



Conclusion

Research Contribution
•Deepens understanding of audiovisual media coverage of the 1983 *March for Equality and Against Racism*.
•Proposes a **methodological pipeline** for diachronic navigation of audiovisual corpora.
•Engages critically with the use of **AI in historical research**.
Key Insights
•High entry threshold for interns & early-career researchers → need for **pedagogical support**.
•Students' feedback: *"History has really changed!"* → signals transformation in historical training.
Future Directions
•Develop **transparent, critically informed methodologies** for AI-driven research.
•Address **epistemological challenges** of web archiving in historiography.
•Ongoing IUF project: renewing theoretical & methodological frameworks via computational approaches.