Semantic analysis of web archive historical data: 1983 “Marche pour l’égalité et contre le racisme”

Davide Rendina, Sophie Gebeil, Mathieu Génois, Patrice Bellot

Abstract: Based on a corpus composed by data obtained from the web archive of the French National Audiovisual Institute, including web pages referencing the history of the 1983 March for Equality and Against Racism, we explored how the memory of a historical event is built through the recounting of web media and the possibilities afforded by computational text analysis methods for the study of large corpuses of historical data from the archived web. This chapter presents the methodology and results of Davide Rendina's master's thesis in computer sciences under the supervision of Sophie Gebeil, Mathieu Génois, and Patrice Bellot. The objective is to demonstrate how historians can utilize archived HTML pages to study the media coverage of historical subjects on the web.

Keywords: anti-racism, media web archive, memory studies, topic modeling, 1983.

Introduction

The chapter delves into the historiographical challenges and digital humanities methodologies encountered in examining the mediatization of the March for Equality and Against Racism that took place in Paris on December 3, 1983, through the French media web archives. This event, a pivotal moment in contemporary French history, marked the emergence of second-generation post-colonial immigrants who confronted racial prejudices and demanded societal recognition. The chapter situates the research at the interdisciplinary intersection of historical studies, digital humanities, and computational sciences[1].

Initially labeled as the “Marche des beurs” by the media, the March holds significance not only as a post-colonial event but also as a manifestation of anti-racist and immigrant social movements. Its historical context is deeply rooted in the aftermath of the Algerian War and the rise of xenophobic sentiments, notably exemplified by the electoral success of the National Front. This event's complex narrative encompasses themes of identity, social justice, and political activism, making it fertile ground for

[1] This chapter is based on the Master Thesis Report: Rendina, D., Gebeil, S., Génois, M., and Bellot, P. (2024). Semantic analysis of web archive historical data the 1983 "Marche pour l'égalité et contre le racisme" [Zenodo]. https://doi.org/10.5281/zenodo.11199667

Davide Rendina, Aix-Marseille University, France, davide.rendina@gmail.com, 0009-0001-3001-8864
Sophie Gebeil, Aix-Marseille University, France, sophie.gebeil@univ-amu.fr, 0000-0002-9883-733X
Mathieu Génois, Aix-Marseille University, France, mathieu.genois@univ-amu.fr, 0000-0001-5492-8750
Patrice Bellot, Aix-Marseille University, France, patrice.bellot@univ-amu.fr, 0000-0001-8698-5055

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Davide Rendina, Sophie Gebeil, Mathieu Génois, Patrice Bellot, *Semantic analysis of web archive historical data: 1983 “Marche pour l’égalité et contre le racisme”*, © Author(s), CC BY 4.0, DOI 10.36253/979-12-215-0413-2.22, in Sophie Gebeil, Jean-Christophe Peyssard (edited by), *Exploring the Archived Web during a Highly Transformative Age. Proceedings of the 5th international RESAW conference, Marseille, June 2024*, pp. 259-273, 2024, published by Firenze University Press, ISBN 979-12-215-0413-2, DOI 10.36253/979-12-215-0413-2

scholarly inquiry. The study aims to explore representations of the March in audiovisual media and on the web, with a particular focus on its semantic treatment online. This interdisciplinary research intersects fields such as memory studies, the history of representations of the past, digital humanities, and computational sciences.

The chapter delves into digital media representations of the March, particularly on the web, probing how these evolve amidst immigration debates and colonial legacies, notably the Algerian War. Employing distant reading, the first section aims to dissect online semantic treatment, uncover recurring themes, and decode underlying narratives. Methodologically, the second section integrates natural language processing methods and network analysis, facilitating systematic exploration of a vast web archive corpus from the INA (French National Audiovisual Institute). Subsequent sections detail data retrieval and methodology, addressing challenges through innovative strategies such as automatic indexing and community detection. The chapter culminates by presenting results and discussions, illuminating temporal trends, identifying entities and topics, and unveiling structural patterns. These insights offer nuanced understanding of media narratives surrounding the March, reflecting societal dynamics and political discourses over time.

1. Historiographical issues and digital humanities challenges

This interdisciplinary research intersects multiple fields: the history of representations of the past, digital humanities, and computational sciences.

Historiographically, 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. The process of memory work only began in the 1990s, with the French state officially considering the war as a mere “law enforcement operation” until 1999 (Branche 2005, 24–44). France’s defeat in 1962 left its mark on public opinion and fueled racism towards those perceived as Algerian (Gastaut

2000). Nostalgic groups for French Algeria carried out racist attacks, such as the one in Marseille in 1973. The National Front, a xenophobic far-right party founded by Jean-Marie Le Pen in 1972, himself a former soldier who served in Algeria, began to achieve electoral success in the cantonal elections of 1982 and municipal elections of 1983.

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, fueled 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 garnered 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 15, 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[2]. This involves understanding: How do representations of

[2] Polyvocal Interpretations of Contested Colonial Heritage (PICCH 2022–2024) is a European project involving five national partners and coordinated by Prof. Daniela Petrelli (Sheffield Hallam University). It explores how archival material created in a colonial mindset can be re-appropriated and re-interpreted to become an effective source for decolonization and the basis for a future inclusive society. The French team at Aix-Marseille University, coordinated by Sophie Gebeil (PI), comprising Véronique Ginouvès (archivist), Christine Mussard (historian), Pauline Savéant (Ph.D.

the March transform with the debates related to immigration in France since the early 2000s? What portrayal is given of the marchers themselves? What role do references to the colonial past play in media coverage of the March? Especially the Algerian War? These questions draw upon corpora composed of audiovisual archives, web videos, and web pages collected by the INA. In this chapter, we focus on the media coverage of the March on the web, particularly the semantic treatment of the event online. Through distant reading, we aim to understand how the event is described within online media and identify recurring names and themes associated with its depiction.

In addition to these inquiries within the fields of memory studies and the history of representations of the past (Gebeil 2021), this research also falls within the realm of digital humanities. Indeed, it raises questions about leveraging the analysis of archived HTML pages from the web to study the media coverage of a subject in web history. This implies an interdisciplinary reflection as it also addresses significant questions in the study of semantic networks and computer science.

2. Corpus and data from the INA web archive

To retrieve data related to the event, the web archive was queried for three different expressions: “Marche des Beurs” OR “Marche pour l’égalité et contre le racisme” OR “Marche de 1983”, with OR being the standard union operator. The web pages were scraped using Boilerpipe, a library designed to process HTML files and extract the main content, thereby filtering out text associated with navigation links and other extraneous elements. It is important to note that since websites are archived every time a single byte changes, the same web page may appear multiple times in the archive. Therefore, a deduplication was performed, retaining only the first (chronologically) URL of each web page. The resulting data formed a JSON file. In this project, only some of the available information from each webpage was kept and extracted in a CSV file:

- id: the unique identifier of the web page.
- url: the original URL.
- title: the HTML title.
- date: the extraction date of the web page, from which the year was used for the deduplication, which can be further used as a proxy for the publication date for a diachronic analysis of the corpora.

student) and Mara Bertelsen (research assistant), in partnership with the INA (Institut National de l’Audiovisuel), mobilize diverse archives to study colonial and postcolonial narratives.

- webpage_text: the text scraped by the boilerpipe algorithm from the web page.

In total, the final corpus contains 12,688 entries, whose distribution across the years is highly skewed, with 44% being from 2013[3]. This is most likely due to the 30th anniversary of the March, coinciding with the release of a movie and a documentary. The temporal heterogeneity in the corpus is important to consider, as it may affect comparisons in diachronic analysis. Additionally, the data includes 558 different domains, ranging from radio (franceinfo.fr) and television (non-stop-people.fr) websites to blogs and forums. Therefore, the data is expected to be heterogeneous in both content and form as well as in time.

3. Methodology: pipeline

The main challenge in handling such a corpus is its size (12,688 documents), which makes manual investigation impossible. Furthermore, the documents are ‘raw’ texts, potentially containing multiple independent sections (e.g., a blog page with several articles, a media home page with titles and excerpts, a forum page with different messages, etc.). The first step in exploring the corpus is thus to make it browsable, i.e. enabling targeted retrieval of documents related to specific questions. Given the prohibitive number of documents in the corpus for manual tagging, we tested whether Natural Language Processing (NLP) methods could facilitate automatic indexing of the documents. Specifically, we focused on two approaches: identifying entities and identifying topics.

3.1 Named Entity Recognition

Named Entity Recognition (NER) enables the identification of entities—words referring to objects that can be denoted with a proper name—in a text, and classifying them into categories (Ehrmann 2021). Among the different techniques that can be used for NER, we relied on a pre-trained deep learning model developed by Babelscape (Tedeschi 2021). Deep learning models hold significant advantages, as they are based on transformers that are able to capture semantic relationships and contextual information in texts without the need for manual feature engineering. Babelscape[4] as a further advantage of being usable on a multilingual dataset.

Though performant, automatic methods are not perfect. The model

[3] Web page distribution across the years, available online: Figure 5.1: Web page distribution across the years, Zenodo. https://doi.org/10.5281/zenodo.11203104

[4] Babelscape, https://web.archive.org/web/20240324020754/https://babelscape.com/

initially identified 127,195 unique entities, a significant portion of which proved unusable due to detection errors. Manual filtering was performed to select only readable entities (34,467 entities). This filtering process, conducted by individuals familiar with the 1983 March, also allowed to tag the entities based on their relevance to the corpus topic. Interestingly, only 2,134 entities were clearly linked to the March, illustrating the 'broad capture' of the initial query. Table 1 provides further insights, showing the number per category, as classified automatically by the Babelscape model.

Table 1. Number of usable named entities extracted per category.

	Relevant	Other
Person	781	23,155
Location	1171	3,184
Organization	73	2,857
Misc.	109	3,317

Named entities provide an initial entry point into the corpus. They enable targeted searches for documents mentioning specific individuals, locations, etc., which can then be manually analyzed. Moreover, they enable cross-searches for documents mentioning multiple entities simultaneously.

More interestingly, the analysis of the list of entities itself opens a window for studying the corpus in its globality. It provides a broader and more exhaustive list of entities related to the main subject, minimizing the risk of overlooking relevant items. Investigating why entities that were *a priori* unrelated to the March appear in the corpus can yield new insights on the subject. For example, focusing on the 20 most frequent persons appearing in the corpus. Remarkably, while political figures and actors from the movie about the March are present, the most frequent person is a journalist, presumably due to her prolific contributions to articles on the subject[5].

3.2 Topic Modeling

The second approach we used is Topic Modeling (TM). In essence, a TM algorithm identifies words that co-occur and groups them into clusters,

[5] See figure online: Figure 5.2, Frequency of top 20 PER NE extracted, p. 32, Zenodo. https://doi.org/10.5281/zenodo.11203104

which are then labeled as ‘topics’. Among the various methods available, we tested two pre-trained deep learning models: BERTopic and Top2Vec. Deep learning approaches based on large language models offer several advantages: they can capture semantic relationships between words and phrases, they do not require extensive text preprocessing tasks, and they can process raw text. More importantly, they automatically detect the number of topics within a given corpus. After testing the coherence of topics generated by both methods (Röder 2015), we focused on BERTopic, as it yielded better results, identifying 106 topics. Figure 3 shows the 20 most frequent topics.

However, these two methods have limitations regarding the size of the text they can analyze. Since text extracted from documents can sometimes be very long, these were split into non-overlapping ‘chunks’ of a maximum size of 512 tokens. Another limitation is that only one topic can be assigned to each chunk. To address the fact that a document may often mention several topics, we assigned to each webpage all identified topics from its chunks.

Table 2. Top 10 topics with their final labels.

Topic	Count	Label	Top 10 Keywords
0	2517	Islam and Muslim Culture	islam, musulmane, ramadan, islamique, mosquée, signes, communauté, école, autres, islamophobie
1	2134	Anti-Racism Activism in Marseille in 1983	égalité, racisme, 1983, octobre, jeunes, immigrés, marseille, collectif, mouvement, violences
2	1766	Music, Television and Entertainment	détails, musique, stop, 25, légales, partenaires, twitter, hanouna, réagissez, adolescence
3	1719	Documentary and Film Festivals	documentaire, films, blog, commentez, festival, caméra, production, réel, email, changer
4	1699	Youth Education	jeunesse, indifférence, éducation, générale, publicité, changé, hôtellerie, sports, trente, guides

5	1443	Democracy and Politics	démocratie, croissance, modèle, social, question, élus, travail, délégation, amendement, politiques
6	1393	Police Violence	police, article, violence, intérieur, amendement, justice, ministère, rumeur, jeunes, sécurité
7	1187	Immigration and Integration	immigration, immigrés, intégration, travailleurs, immigré, migratoires, immigre, familles, population, africains
8	1144	Anti-Semitism	juifs, antisémitisme, palestiniens, israéliens, humoriste, paix, antisémites, football, sionisme, allemands
9	1005	Commemoration of Algerian War	guerre, 1962, algérien, indépendance, histoire, mémoire, peuple, colonial, française, nationalisme

As for entities, labeling documents with their topics enables targeted exploration of those mentioning specific subjects. Cross-searches are also possible, both between topics and between topics and entities. Similar to entities, investigating the list of topics may offer new insights. For example, while the 1983 March initially addressed equal rights and racism, the most prominent topic in the corpus is Islam, which suggests a potential bias or specific presentation of the subject within the corpus[6].

3.3 Network approaches

While automatic labeling of documents with NER and TM yielded compelling results, we went one step further in exploring the corpus. We used a network-based approach to discern whether entities and topics could unveil an underlying structure. We defined four networks:

1. the **document-document network based on topics**: Each document is a node, and a link exists between two documents sharing at least one topic. The weight of a link is then the number of shared topics.
2. the **document-document network based on entities**: Each document is a node,

[6] Top 20 topics extracted, with for each the top 5 most relevant words. Figure available online: Rendina_5.7.png, Zenodo. https://doi.org/10.5281/zenodo.11203104

and a link exists between two documents with at least one common entity. The weight of a link is then the number of common entities.

3. the **topic-topic network**: Each node is a topic, and a link exists between two topics if they appear together in at least one document. The weight of a link is either the number of documents in which they co-appear, or the Pointwise Mutual Information (PMI) score, which adjusts for differing topic frequencies.
4. the **entity-entity network**: Each node is an entity, and a link exists between two entities if they appear together in at least one document. The weight of a link is the number of documents in which they co-appear.

We then used community detection algorithms to ascertain whether these networks were structured, i.e. if nodes form groups. For simplicity, we used the Louvain algorithm (Fortunato 2009).

In all four networks, we observed that their structure is far from random or uniform. Both topics and entities group themselves into communities, suggesting correlations in their occurrences within documents: some topics (respectively entities) are related to each other. In particular, topics appear to form two distinct communities[7].

Document-document networks also exhibit non-trivial community structures (see Figure 1). This indicates that shared topics or shared entities are indeed related to the existence of sub-corpora within the corpus, which may share a focus on a specific aspect of the main subject, a particular discourse, etc.

[7] Topic-topic network with PMI. Zenodo. https://doi.org/10.5281/zenodo.11203104

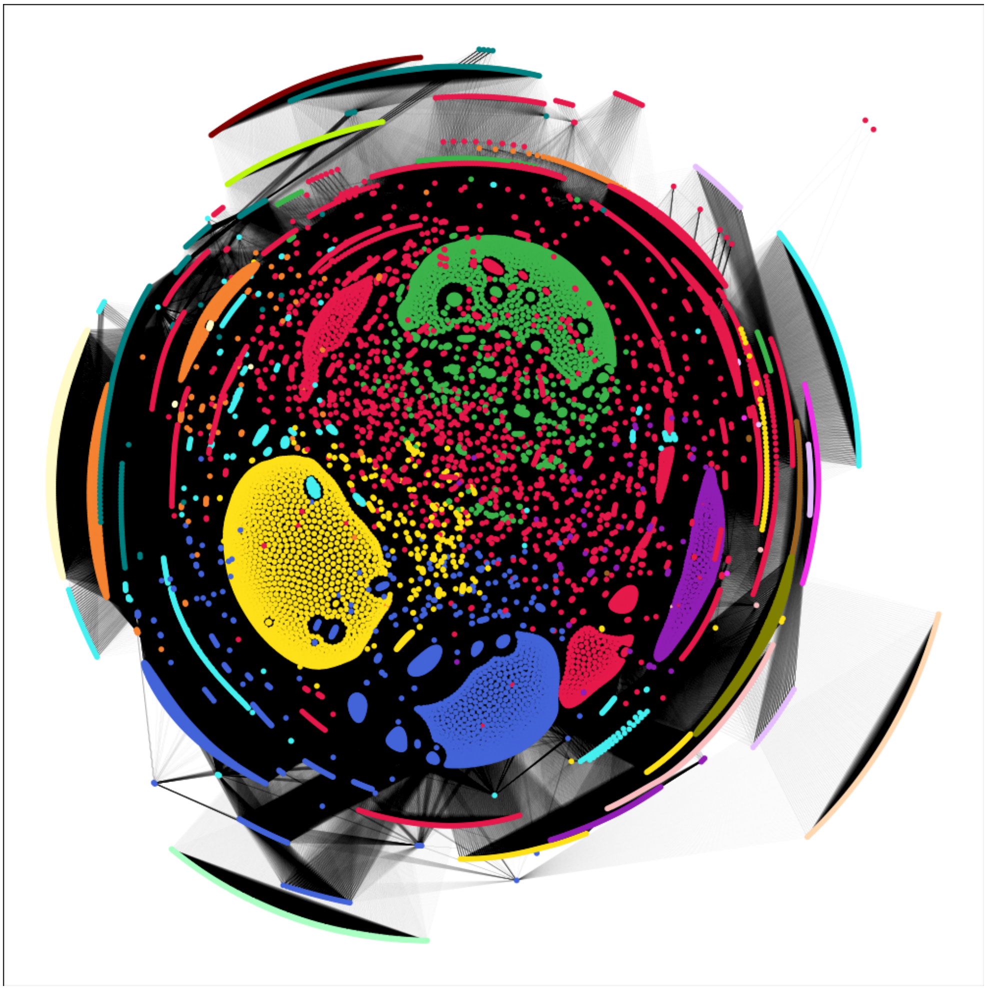

Figure 1. Document-document network based on shared topics. Colors indicate the main communities found, with red as ‘other’.

Exploring the underlying reasons for the existence of these communities can provide profound insights into the organization of discourse around the 1983 March. By using communities as additional labels for documents, we can specifically target them for analysis. The presence of topic and entity communities can be instrumental in selecting documents within the corpus associated with particular groups, and studying why and how these documents form a consistent object.

4. Results and discussion

Overall, the implemented pipeline provides several insights into the media coverage of the March through a distant reading.

The diachronic approach first illustrates the peak of media coverage in 2013 regarding the March online, corresponding to the event's 30th anniversary (Fig. 2). While this spike in 2013 also reflects corpus bias due to the abundance of data from that year, the ability to visualize the evolution of domains by year also shows that 2009, 2010, and 2018 are peaks as well. This corroborates findings from a qualitative survey conducted in 2014, which highlighted a resurgence of interest in the March due to the publication of testimonies or the political engagement of former marchers. Comparing themes per year also reveals an intensified association with anti-Semitism in 2014, which constitutes a particularly interesting avenue for further exploration. This demonstrates the significant role of media coverage in shaping perceptions of the March and its participants, for better or for worse. Indeed, in 2014, F. Belghoul, one of the oldest marchers, launched a crusade against the ABCD of equality program initiated by the Ministry of National Education, with the support of essayist Alain Soral consistently linked with anti-Semitic views. This radical shift from anti-racism to far-right ideology is extensively covered in the press and on television. The March, once seen as a symbol of anti-racism in the 1980s, is now invoked to underscore this radical political change.

Exploring web entities provides an overview of the media narratives of the March. First and foremost, it is evident that the term "Marche des beurs" persists, overshadowing the original name of the March. However, upon closer examination of the data, the term "Marche des beurs" is often enclosed in quotation marks, suggesting a potential avenue for further exploration.

Secondly, the corpus also highlights the main stages of the March, with Paris being the most represented, followed by Marseille, the starting point, and Lyon, along with references to the Minguettes and the name of Vénissieux (Rendina et al. 2023, 54). However, while Marseille is frequently mentioned, it is uncertain whether manual archive consultation would yield substantial documentary evidence about the departure from Marseille, as it received little media attention in 1983: the fact that Marseille is prominently mentioned does not necessarily mean that the events that unfolded in the city on October 15, 1983, are as well-documented as the history of the Minguettes).

The word cloud of organizations identified as entities serves as a reminder that the March remained, even in the 2000s, a focal point for various political parties (Rendina et al. 2023, 54–55). Occurring while the Socialist Party was in power, the names of leftist parties, along with SOS Racisme, are among the most frequently cited organizations. The Green Party is also mentioned. Interestingly, far-left parties such as the LCR, the NPA (Nouveau Parti Anticaptialiste), and Lutte Ouvrière are absent from the list of most-named entities. On the right, while the UMP garners mentions, it is the National Front that stands out prominently concerning the March, possibly reflecting the far-right's early adoption of web-based strategies (Gimenez andVoirol 2017; Mudde 2007).

Another significant contribution of this semantic analysis lies in the identification of the key players in the media coverage of the March within the audiovisual sector. Among the prominently featured occurrences are the public group France Info (radio), France 2 (TV), and Canal+. While Canal+ is now owned by the Bolloré group, whose editorial stance aligns with conservative right-wing ideologies, marked by recurrent anti-immigrant and anti-diversity discourses, its involvement in the media coverage of the March in 2013 is not surprising. On one hand, the channel played a pivotal role in fostering comedic and anti-racist immigrant narratives, notably through personalities like Djamel Debbouze. On the other hand, journalist Maxime Musca, a member of the show "Le Petit Journal", spearheaded the "Refaire la Marche" initiative, with each stage broadcasted during daily shows until its culmination in Paris in December 2013.

Topic Modeling offers valuable insights into the thematic underpinnings surrounding the narratives of the 1983 March within the web milieu of audiovisual media in 2013. The commemorative events notably feature the promotion of Yabil Ben Yedir's film *La Marche*, starring Djamel Debbouze, as prominently evident in topics 3, 6, and 16.

Delving into textual data further confirms a pronounced focus on Islam and Muslims, reflecting the media's enduring preoccupation with the societal positioning of Muslims in France since the early 2000s. While religious demands were present among the marchers in 1983, they were not foregrounded or explicitly articulated. Rather, they symbolized a broader call for the full assimilation of post-colonial immigrant children into the national fabric, amid fervent expectations of equal rights in the face of racial violence and social marginalization. The figure of the young Maghrebi captivated attention in the 1980s, effectively reducing the March to the narrative of the 'beurs'. Three decades later, references to the offspring of the original marchers remain entwined with inquiries into the integration prospects of Muslims and their descendants, despite their longstanding French citizenship spanning multiple generations. This thematic discourse is intricately linked to the substantial presence within the reference corpus of Eric Zemmour, a right-wing provocateur convicted in 2011 for inciting racial discrimination. Zemmour's discourse, disseminated through media platforms such as "On n'est pas couché" (2006–2011) and later on RTL, pervades online discussions. Consequently, celebrations of the March are overshadowed by persistent controversies concerning French Muslims, jihadist terrorism, and the Israeli-Palestinian conflict.

Lastly, topic 9 warrants particular attention due to its alignment with the objectives of the PICCH project, which scrutinizes representations of the colonial past in contemporary media. It corroborates an associative linkage between the narratives of the March and the Algerian War, the French defeat that culminated in Algeria's independence in 1962. This data analysis not only facilitates the identification of contentious issues but also underscores the need to delve deeper into the nuanced significance attributed to the invocation of this event, which unfolded 21 years preceding the March.

Conclusion

This interdisciplinary study, drawing upon text extracted from HTML pages covering the media portrayal of the March, sourced from a vast corpus of the INA web archives, provides valuable insights into Memory Studies, as well as broader fields such as digital history and web archive studies. By harnessing NLP techniques and network analysis, it demonstrates the efficacy of integrating these methodologies for the exploration of large historical corpora. By leveraging the capabilities of NER, topic modeling, and network analysis, we have provided digital historians with valuable tools to navigate and extract meaningful insights from vast amounts of historical data.

The study first sheds light on the ambivalences surrounding the thirtieth anniversary of the March. This milestone prompted an unprecedented surge in content production, fueled in part by the release of Yabil Ben Yedir's film and its cast. However, the mention of the actors in this march, predominantly children of the first generation of immigrants perceived as Black and especially ‘beurs’ in the 1980s, sparked myriad controversies, prejudices, and hateful discourse, illustrating the enduring prevalence of the catch-all media figure variously labeled as ‘beur’, ‘Maghrebi’, ‘Algerian’, or ‘Muslim’. Consequently, the event appears saturated, drowned in a torrent of controversies that obscure the marchers’ initial message as they attempt to share their own vision of the March. This study thus opens numerous avenues for further exploration and inquiry, which will require examining discourse related to Islam and the Algerian War within the corpus.

Furthermore, this interdisciplinary approach, combining history and computer science, also contributes to the renewal of historical methods facilitated by the use of archived web pages as inherently digital sources, beyond the case of the media coverage of the March. It offers a distinct approach to historical sources using deep learning models, commonly termed AI-driven automated tools: not only for data processing but also for source compilation and analysis. Semantic data processing is enhanced by tools such as BERT, and the developed pipeline can be replicated across any web corpus. Additionally, the work on named entities provides researchers with a repertoire of words related to the media coverage of the March, which can be used to search for other web pages or to further analyze the corpus. This repository, available online, serves to document the event from a fresh perspective[8]. The exploration of semantic networks is also a novel approach, revealing clusters that remain to be interpreted. This reflects an important historiographical consideration for researchers working with data sourced from both the live and archived web, an area of study set to evolve within the WebLab created at the MMSH (Maison Méditerranéenne des Sciences de l'Homme) in April 2024 from a multidisciplinary perspective.

The interpretation of this distant approach stands to benefit from qualitative analysis of sources, focusing on content related to Islam and the Algerian War within the corpus. Lastly, the study's limitations prompt reflections for future research on corpora derived from archived web data, beginning with the case of the 1983 March. Efforts are underway to explore other semantic analysis tools in collaboration with the WebLab and the *CEntre de formation et de soutien aux Données de la REcherche* (CEDRE) at Aix-Marseille University[9]. Using the same corpus, comparisons between results obtained from different programs will be conducted. While diachrony remains a cherished aspect of historical research, it has been relatively overlooked here. Thus, this project marks the initial phase for expanding and refining the corpus to reproduce the pipeline on a television and web corpus spanning from 1983 to 2023, enabling an exploration of semantic network and topic evolution across successive commemorations.

[8] Named Entities Topics Analysis https://public.tableau.com/app/profile/davidearendina/viz/NE_Topics_analysis/NE_Topics_Analysis

[9] WebLab https://pba.mmsh.fr/?page_id=1465; CEDRE, https://www.univ-amu.fr/en/node/7879

References

Davide Rendina, Sophie Gebeil, Mathieu Génois, Patrice Bellot. 2023. "Semantic analysis of web archive historical data: the 1983 'Marche pour l'égalité et contre le racisme'." Master Thesis. Erasmus Mundus Joint Master's Degree in Big Data Management and Analytics (BDMA). Data Analysis, Statistics and Probability [physics.data-an]. ⟨dumas-04541382⟩

Davide Rendina, Sophie Gebeil, Mathieu Génois, Patrice Bellot. 2023. "Master Thesis Report – Semantic Analysis of Web Archive Historical Data the 1983 'Marche pour l'égalité et contre le racisme'." Zenodo, 10 August 2023. https://doi.org/10.5281/zenodo.10972646

De Lange, Sarah L. and Mudde Cas. 2005. "Political extremism in Europe." *European Political Science* 4(4): 476–88. http://www.cambridge.org/9780521850810

Ehrmann, Maud, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2024. "Named Entity Recognition and Classification on Historical Documents: A Survey." *ACM Computing Surveys* 56 (2): 1–47. https://doi.org/10.1145/3604931.

Fortunato, S. 2009. "Community detection in graphs." *Physics Reports*, 486 (3–5), 75–174. https://doi.org/10.1016/j.physrep.2009.11.002.

Gimenez, Elsa, and Voirol Olivier. 2017. "Les agitateurs de la toile. L'Internet des droites extrêmes. Présentation du numéro." *Réseaux* 202–203, no. 2–3: 9–37. https://doi.org/10.3917/res.202.0009

Pippa, Noris. 2003. "Preaching to the converted?: Pluralism, participation and party websites." *Party Politics* 9(1): 21–45.

Röder, M., Both, A., and Hinneburg, A. 2015. "Exploring the space of topic coherence measures." In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM 2015, Shanghai, China, February 2–6, 2015, X. Cheng, H. Li, E. Gabrilovich, and J. Tang, Eds., ACM, 399–408.

Tedeschi, S., Maiorca, V., Campolungo, N., Cecconi, F., and Navigli, R.. 2021. "Wikineural: Combined neural and knowledge-based silver data creation for multilingual NER." In *Findings of the Association for Computational Linguistics: EMNLP 2021*, Virtual Event / Punta Cana, Dominican Republic, 16–20 November, 2021, M. Moens, X. Huang, L. Specia, and S. W. Yih, Eds. Association for Computational Linguistics, 2521–2533.