

Chelsea A. Rodriguez

# Digital Newspapers, Material Knowledge: Grappling with the TimesMachine Digital Archive as a Repository of Knowledge

**Abstract:** This article explores the functionality and character of the TimesMachine, the digital archive of the *New York Times*, as it relates to research in the history of knowledge and history of education. The TimesMachine was designed specifically to improve the accessibility and contextualization of digital newspapers. However, the fog of digitization still complicates how researchers can (and should) use these digitized newspapers as sources of knowledge. Using the TimesMachine as a case study, this article aims to contribute to current conversations regarding the possibilities and limitations of digital archives and to demonstrate the importance of digital literacy for historians. There are several takeaways from this article that can help researchers make informed methodological choices in their pursuit of history, including the recentering of materiality in the digital space and the integration of interdisciplinary and digital methods in the history of knowledge.

**Keywords:** digital archive, The New York Times, digital newspapers, sources of knowledge, digital methods

## Introduction

*The New York Times* has played an instrumental role as a producer of knowledge and has served as the standard of authoritative journalism for its readership throughout the twentieth century.<sup>1</sup> In 2014, *The Times* developed its own digital archive, the aptly named TimesMachine, which was designed to be a utopia of digital newspaper archives and includes high-quality, full-page scans of its entire published body of work from 1851 to 2002.<sup>2</sup> Keyword searchable, creatively contextualized, and empirically rich, the TimesMachine and its unique immersive interface is

---

<sup>1</sup> For the sake of brevity, *The New York Times* is also referred to as *The Times* in this text.

<sup>2</sup> Jane Cotler and Evan Sandhaus, "How to Build a TimesMachine," *The New York Times Open Blog*, February 1, 2016 (archived), accessed April 13, 2023, <https://open.blogs.nytimes.com/2016/02/01/how-to-build-a-timesmachine/>. The TimesMachine is not a public archive but is accessible to subscribers of the *New York Times*, including academic institutions.

indicative of how digitization has shifted the outer bounds of what is possible for historians.<sup>3</sup>

However, while offering wide-ranging opportunities for research, untangling the forms and circulation of ideas through the prism of this digital database requires care and attention to the historical and material context of the newspaper as a cultural artifact. The value of newspapers as historical source materials is clearly established, owing to their rich historical context and unique periodicity, which makes them ideal for researching social continuity and change.<sup>4</sup> However, as digitization continues to develop more user-friendly interfaces, the question still remains: how should historians best grapple with digital newspaper archives such as the TimesMachine, which aim to replicate the material experience of the newspaper within an abstracted digital space? There is a widely felt need to develop more historically informed understandings of digitized newspapers and to set standards to help researchers overcome both technical and philosophical challenges in digital archives.<sup>5</sup>

Scholarly debates concerning the digitization of newspapers and the ensuing implications of digitalization on historical research have kept pace with the proliferation of digital sources and archives over the last decade. The most consequential discussions for historians have concerned the impacts of source abundance and mass digitalization on the craft of history,<sup>6</sup> the need for historians to develop

---

3 David Larsson Heidenblad, "The Emergence of Environmental Journalism in 1960s Sweden: Methodological Reflections on Working with Digitized Newspapers," in *Histories of Knowledge in Post-War Scandinavia: Actors, Arenas, and Aspirations*, ed. Johan Östling, Niklas Olsen, and David Larsson Heidenblad (London: Routledge, Open Access, 2020), 59.

4 Bob Nicholson, "The Digital Turn: Exploring the Methodological Possibilities of Digital Newspaper Archives," *Media History* 19, no. 1 (2013): 59–73, accessed April 13, 2023, doi:10.1080/13688804.2012.752963; John Tosh, *The Pursuit of History: Aims, Methods and New Directions in the Study of Modern History*, 6th edition (New York: Routledge, 2015).

5 Melody Beals and Emily Bell, *The Atlas of Digitised Newspapers and Metadata: Reports from Oceanic Exchanges*. Online resource (Loughborough: Oceanic Exchanges, 2020), 1, <https://doi.org/10.6084/m9.figshare.11560059.v2>.

6 Just a few of the most relevant examples include Richard Abel, "The Pleasures and Perils of Big Data in Digitized Newspapers," *Film History* 25, no. 1–2 (2013): 1–10, accessed January 31, 2023, <https://doi.org/10.2979/filmhistory.25.1.2.1>; Andreas Fickers, "Towards a New Digital Historicism? Doing History in the Age of Abundance," *VIEW Journal of European Television History and Culture* 1, no. 1 (2012): 19–26; Sarah Van Ruyskensvelde, "Towards a History of E-ducation? Exploring the Possibilities of Digital Humanities for the History of Education," *Paedagogica Historica* 50, no. 6, (2014), accessed January 31, 2023, doi:10.1080/00309230.2014.955511; and, most recently, Estelle Bunout, Maud Ehrmann, and Frédéric Clavert, *Digitised Newspapers – A New Eldorado for Historians?: Reflections on Tools, Methods and Epistemology* (Berlin & Boston: De Gruyter Oldenbourg, 2023), accessed January 31, 2023, <https://doi.org/10.1515/9783110729214>.

digital literacy and, accordingly, a new kind of source criticism skills,<sup>7</sup> and the effects of different design features and coding methods on the quality of digital archives and sources for historical researchers.<sup>8</sup> Essential work has also been carried out to establish interdisciplinary baselines concerning terminology, useability, and access related to digital newspaper archives, thereby allowing researchers from different disciplines and national contexts to better communicate and collaborate on digitized newspaper research.<sup>9</sup> However, most of these discussions about digital newspapers and archives have occurred outside the history of knowledge research field, meaning that there are few examples of scholarship explicitly focusing on the added value (or challenges) of using digitized newspapers to explore historical structures and circulators of knowledge.

One notable exception is the work of historian of knowledge David Larsson Heidenblad, who has recently used his research on the emergence of environmental journalism in Sweden to explore which considerations must be made when working with digitized newspapers.<sup>10</sup> In line with other scholars, Heidenblad points to issues of decontextualization, while also highlighting the positive aspects of expanded accessibility and the possibilities offered by digitization in terms of transnational research.<sup>11</sup> Digital sources are increasingly being used to craft complex and diverse histories in various disciplines, and there are many examples of

---

7 See, for example, Helle Strandgaard Jensen, "Digital Archival Literacy for (All) Historians," *Media History* 27, no. 2 (2021): 251–265, accessed April 12, 2023, doi:10.1080/13688804.2020.1779047; Jon C. Giullian, "'Seans Chernoi Magii Na Taganke': The Hunt for Master and Margarita in the Pravda Digital Archive," *Slavic & East European Information Resources* 14, no. 2–3 (April 1, 2013): 102–126, accessed January 31, 2023, <https://doi.org/10.1080/15228886.2013.813374>; and Tim Hitchcock, "Confronting the Digital," *Cultural and Social History* 10, no. 1 (2013), doi:10.2752/147800413X13515292098070.

8 Two of the most recent and relevant works are those of Sarah Oberbichler et al., "Integrated Interdisciplinary Workflows for Research on Historical Newspapers: Perspectives from Humanities Scholars, Computer Scientists, and Librarians," *Journal of the Association for Information Science and Technology* 73, no. 2 (February 2022): 225–239, accessed January 31, 2023, <https://doi.org/10.1002/asi.24565>; and Maud Ehrmann, "Historical Newspaper User Interfaces: A Review," Athens, Greece: IFLA (2019), accessed January 31, 2023, <http://library.ifla.org/2578/>.

9 See, for example, Beals and Bell, *Atlas of Digitised Newspapers*; Lara Putnam, "The Transnational and the Text-Searchable: Digitized Sources and the Shadows They Cast," *The American Historical Review* 121 (2016): 377–402, accessed January 31, 2023, doi:10.1093/ahr/121.2.377; and Benjamin Charles Germain Lee et al., "The Newspaper Navigator Dataset: Extracting and Analyzing Visual Content from 16 Million Historic Newspaper Pages in Chronicling America," Cornell University [Cs] (May 2020), accessed January 31, 2023, <http://arxiv.org/abs/2005.01583>.

10 Heidenblad, "Environmental Journalism."

11 Ibid., 68–69.

quality research predicated on the analysis of historical, digitized newspapers.<sup>12</sup> However, unlike in digital history circles, where questions of digitization are frequently debated and circulated, most of these studies typically lack methodological reflections on the differences between material and digital newspapers in historical analyses.<sup>13</sup> Researchers are familiar with the challenges of digitization in general. However, it is valuable to explore these challenges within the specific space of the TimesMachine digital archive, which has not previously received a great deal of attention in academic circles.

In this article, I use empirical findings from my research into the history of post-war education news coverage to make a new contribution to the broader methodological discussions on digital source materials in the history of knowledge. Integrating concepts and methods from the history of knowledge into this project has allowed me to develop new perspectives on education news as a broad political, social, and cultural phenomenon. Particularly the notions of knowledge actors and circulation have allowed me to center *The New York Times* as a historical actor and arena of contestation that has set the limits on the circulation of knowledge about education for millions of readers throughout the twentieth century.<sup>14</sup> This type of in-depth institutional and historical research has been enabled by mass digitization, but there are drawbacks that need to be addressed.

The unique context and character of *The Times*, as it exists in the TimesMachine digital archive, is used as a case study to explore the potential and limita-

---

12 A short selection of examples in different disciplines and national contexts include Jennifer L. Cohen, "Teachers in the News: A Critical Analysis of One US Newspaper's Discourse on Education, 2006–2007," *Discourse: Studies in the Cultural Politics of Education* 31, no. 1 (February 2010): 105–119, accessed January 31, 2023, <https://doi.org/10.1080/01596300903465450>; Joakim Landahl, "De-Scandalisation and International Assessments: The Reception of IEA Surveys in Sweden during the 1970s," *Globalisation, Societies and Education* 16, no. 5 (October 20, 2018): 566–576, accessed January 31, 2023, <https://doi.org/10.1080/14767724.2018.1531235>; Jani Marjanen et al., "A National Public Sphere? Analysing the Language, Location and Form of Newspapers in Finland, 1771–1917," *Journal of European Periodical Studies* 4, no. 1 (30 June 2019): 54–77; and Archie Thomas, Andrew Jakubowicz, and Heidi Norman, *Does the Media Fail Aboriginal Political Aspirations? 45 Years of News Media Reporting of Key Political Moments* (Chicago: Aboriginal Studies Press, 2021).

13 See Alexandra Chassanoff, "Historians and the Use of Primary Source Materials in the Digital Age," *The American Archivist* 76, no. 2 (2013): 458–480, <https://americanarchivist.org/doi/10.17723/aarc.76.2.lh76217m2m376n28>; Donghee Sinn and Nicholas Soares, "Historians' Use of Digital Archival Collections: The Web, Historical Scholarship, and Archival Research," *Journal of the Association for Information Science and Technology* 65, no. 9: 1794–1809, <https://asistdl.onlinelibrary.wiley.com/doi/full/10.1002/asi.23091>.

14 Based on the conceptualizations outlined in Johan Östling and David Larsson Heidenblad, "Fulfilling the Promise of the History of Knowledge: Key Approaches for the 2020s," *Journal for the History of Knowledge* 1, no. 1 (2020): 2, accessed January 31, 2023, <https://doi.org/10.5334/jhk.24>.

tions of analyzing digitized newspapers as sources of knowledge. This exploration is organized in four parts: first, I introduce *The New York Times* newspaper and its history in the post-war era. This section focuses, in particular, on how the newspaper viewed and structured itself and how its practices shaped a distinct reality and knowledge (about education). Second, I delve into the institutional, technical, and user-oriented aspects of the TimesMachine to explain how the medium of digital archives affects what researchers may find and perceive about the past.<sup>15</sup> I also focus on the archive's efforts to overcome common methodological limitations, particularly concerning historical and material contextualization, that came with previous digital newspaper archives.<sup>16</sup> While the design of the TimesMachine certainly addresses some of these issues, in the third section, I look more closely at which research prospects are still limited and how certain gaps between the digital and analog forms of the newspaper persist. Finally, in the fourth section, I present examples of how recentering the materiality of the newspaper in the archive can help researchers better navigate the fog of digitization. This article concludes with a discussion on how the case of the TimesMachine further illustrates the possibilities and limitations historians must grapple with when using digital newspapers and archives as sources of knowledge.

## ***The New York Times* as a Producer of Knowledge (about Education)**

No newspaper in the United States has sought to establish itself with as much authority and influence in the lives of its readers as *The New York Times*. The very phrase used to describe an authoritative newspaper with a wide reach, “a newspaper of record,” was actually created by *The Times* to describe itself, thus perpetually intertwining its practices and image with notions of objective, authoritative reporting.<sup>17</sup> While *The Times* is still an active and influential newspaper, many aspects of its organization, printing structure, leadership, and format have changed since the 1990s. This section explicitly concerns *The New York Times* during the

---

15 Inspired by Jensen's conceptualization of the archive, as elaborated in Jensen, “Digital Archival Literacy,” 253, 257. Jensen equates the (digital) archive to other mediums of communication (e.g., radio or websites) as they channel cultural production (collections) in similar ways. A key thing to remember in the context of digital archives is the cybertechnical infrastructure, absent from analog archives, which mediates all interactions with users and the items they contain.

16 Nicholson, “Digital Turn,” 60.

17 “The Newspaper of Record,” Advertisement, *The New York Times*, October 26, 1924, p. S5, col. 5.

post-war era (1945–1990) and describes the character and nature of the paper as it is preserved in the TimesMachine, based on primary and secondary sources.

For most of its history, *The New York Times* provided daily local, national, and international news coverage to a predominantly middle-to-upper class readership, not only in the densely populated tri-state area (New York-New Jersey-Connecticut) or the metropolitan United States, but also across the globe.<sup>18</sup> Founded in 1851 and led by the Ochs-Sulzberger family since 1896, *The Times* has been described as “the fullest, most expensive daily record of history compiled by man,” while the journalists responsible for the stories strewn across its grey pages “left their fingerprints on the first drafts of history.”<sup>19</sup> Arguably, no other paper possessed comparable resources, esteem, or influence during the twentieth century, and certainly not in the post-war era: when *The Times* spoke, even the chancellors of Europe and the Kremlin took notice.<sup>20</sup> Every evening, hundreds of thousands of copies were printed and flown to all reaches of the world, and other major news outlets such as the Associated Press, Reuters, Agence France-Presse, and TASS would be transmitting chief dispatches by *The Times* by wireless and satellite before copies even hit the shelves.<sup>21</sup>

Its expansive staff and budget meant that *The New York Times* was always ready for the “next Titanic,” and it would frequently send reporters to follow up on major stories weeks, months, or even years after an initial story broke.<sup>22</sup> *Timesmen*, a popular nickname for its journalists, strove for an almost “sociological” approach to news coverage, seeking to avoid sensational tabloid stories by framing their news coverage as investigative and authoritative.<sup>23</sup> While some Timesmen were credited with bylines during the twentieth century, a majority of stories in *The Times* were anonymous, reinforcing the illusion that the Grey Lady, not the journalists, was the one responsible for bringing knowledge of the world to its readership.<sup>24</sup> This notion was also reinforced by the paper’s visuals, as its reserved

---

18 Harrison Salisbury, *Without Fear or Favor: The New York Times and Its Times* (New York: Times Books, 1980), 4–5.

19 Meyer Berger, “The Gray Lady Reaches 100,” *Life Magazine*, September 17, 1951, 153; Seth Mnookin, *Hard News: The Scandals at The New York Times and Their Meaning for American Media* (New York: Random House, 2004), 4.

20 Mnookin, *Hard News*, 4.

21 Ibid.

22 Ibid., 10–14; Edwin Diamond, *Behind the Times: Inside The New York Times* (Chicago: University of Chicago Press, 1994), 12.

23 Diamond, *Behind the Times*, 11–12.

24 Robert D. McFadden, “150<sup>th</sup> Anniversary: 1851–2001; 150 and Counting: The Story So Far,” *The New York Times*, November 14, 2001, accessed January 30, 2023, <https://www.nytimes.com/2001/11/14/news/150th-anniversary-1851-2001-150-and-counting-the-story-so-far.html>.

use of large headlines and photo layouts presented an orderly, structured world for readers and became the standard by which the relative importance of events was judged.<sup>25</sup>

During the twentieth century, *The Times* was published in two formats: the *daily edition*, a two-section newspaper published Monday through Saturday, which ranged from 40 to 80 pages in length, and the *Sunday edition*, a multi-sectional news behemoth filled with upwards of 300 pages of content. The Sunday paper was perhaps the most prestigious, with its own dedicated executive editor and section-specific team of editors specializing in subjects such as Business and Finance, Travel, Local News, and Education. Besides extended news coverage, the extra volume on Sundays can also be attributed to the addition of *The New York Times Magazine*, *The New York Times Review of Books*, and extra advertising. The Sunday editor of *The Times* operated independently from the weekday managing editor up until the 1970s, meaning that news decisions made Monday through Saturday were different and often disjointed from the news decisions made for the Sunday edition.<sup>26</sup>

Counterintuitively, the presence of special sections on Sundays may actually have limited the circulation of knowledge and ideas regarding these subjects. Having a dedicated section on education in *The Times* provided the education editor with space for in-depth, nuanced coverage of education issues, but that education section only occupied half a page in the fourth section (Review of the Week) of the three-hundred-page newspaper. This means that most casual readers would likely not have encountered this education news section unless they intentionally sought it out. This demonstrates how the reach of certain policy and social ideas was often dependent on the placement of stories and topics within the paper.<sup>27</sup>

With the exception of a few months of strikes during the 1960s and 1970s, *The Times* has been shaping the contours of a *mediated* world for readers on a daily basis throughout the twentieth century. Mediated in this context refers to the power of the newspaper in communicating knowledge and information to readers, not only through the literal information it shares but also through the specific medium and prism of newsprint. Think here of media theorist Marshall McLuhan's assertion in media studies that "the medium is the message," as the material and cultural elements of a media artifact (newspaper) also influence the messages it aims to communicate.<sup>28</sup> Certainly at least since being purchased by Adolph Ochs

---

<sup>25</sup> Ibid.

<sup>26</sup> Mnookin, *Hard News*, 13.

<sup>27</sup> Allen Bell, *The Language of News Media* (New Jersey: Wiley-Blackwell, 1991), 14.

<sup>28</sup> Marshall McLuhan, "The Medium is the Message," in *Understanding Media: The Extensions of Man* (New York: McGraw Hill, 1964).



in 1896, *The New York Times* has been conscious of its image and influence on the worldview of its readers. Ochs and his daughter Iphigene would often share an anecdote with their staff about stonecutters and cathedral builders, which exemplifies the mentality of the newspaper: *The Times* does not employ stonecutters, who produce generic news stories with little care regarding their potential; *The Times* hires cathedral builders, who see their journalistic work as contributing to something greater, to building the kingdom of news and public knowledge that is *The New York Times*.<sup>29</sup> So, naturally, when the time came to digitize their newspaper archive, *The Times* “set out to reimagine the archive browsing experience” by developing an immersive, “more sophisticated” digital archive: the TimesMachine.<sup>30</sup>

## What Can You Do with a TimesMachine? Exploring the Digitized Past

Until the creation of the online version of the newspaper in 1995, *The New York Times* was only available in print. As most text-based historical sources, archived editions could be found in select libraries and archives, mostly on microfilm.<sup>31</sup> The first version of the digital TimesMachine was launched in 2008 and consisted of full-page PDF scans spanning from 1851 to 1922.<sup>32</sup> Its launch coincided with the creation of many early digital newspaper archives, including the British Library of 19<sup>th</sup> Century Newspapers, Chronicling America, Hemeroteca Nacional Digital de México, and Delpher.<sup>33</sup> Sensing the limitations of full-page PDFs, notably the large file size and lack of internal metadata for keyword searches, the archive was redesigned in 2014 to expand its collection and improve the existing online interface.

---

29 Gay Talese, *The Kingdom and the Power: Behind the Scenes at The New York Times: The Institution That Influences the World* (New York: Random House, 2007), 15–16.

30 Evan Sandhaus, “Introducing the New TimesMachine,” The New York Times Open Team Blog, July 11, 2013, accessed March 11, 2023, <https://open.nytimes.com/introducing-the-new-timesmachine-e4686183261f>

31 Microfilm is obviously a different carrier medium involving different methodological and analytical challenges, even though it is not all that different from the initial full-page digital scans of *The New York Times* characterizing its first foray into digital archiving. An excellent overview of the history of microfilm and its challenges can be found in Ian Milligan, *The Transformation of Historical Research in the Digital Age* (Cambridge: Cambridge University Press, 2022). doi:10.1017/9781009026055.

32 Cotler and Sandhaus, “How to Build a TimesMachine.”

33 Ten of these large-scale digital newspaper archives created between 2000 and 2012 are explored and analyzed in depth in Beals and Bell’s *The Atlas of Digitised Newspapers and Metadata*.



Demonstrating its persistent ambition to establish itself at the forefront of journalism, *The Times* adopted a new method of structuring digital newspaper archives. By employing image tiling strategies from digital mapping software and combining this mapping with optical character recognition (OCR), the TimesMachine is able to display full-page, browsable scans of their newspapers in a keyword-searchable, but also page-turnable, interface.<sup>34</sup> This page-turning ability and full newspaper visualization features are what set the TimesMachine apart from other digital newspaper archives that typically retrieve individual articles and (unavoidably) conceal the context of the surrounding issue to which they belong.<sup>35</sup> The archive's interactive design sought to remove the barriers between users and the digitized newspaper by replicating the experience of holding a newspaper as close as the web permits, leveraging a vast collection of scanned images and metadata with accessibility and a user-friendly organization.<sup>36</sup> This immersive archive of 12.8 million articles and 4.1 million pages visually exemplifies the size and reach of *The Times* as a producer of knowledge.<sup>37</sup>

Perhaps the most relevant aspect for historians, however, is that the TimesMachine is intentionally structured to address a problem frequently plaguing (digital) archives: a lack of context.<sup>38</sup> Newspaper archives are often made up of scanned clippings or individual pages of newspapers, which are then retrieved for users via a search tool. As a result, the sources are presented as suspended in this digital space, stretched beyond the historical and material context in which they first appeared. For example, a researcher using keywords to search for information on Cold War student movements might find a front page story in the TimesMachine about the brutal suppression of student protestors by Warsaw police on October 5, 1957.<sup>39</sup> However, a contemporary reader absorbed in the newspaper that morning would have been considerably more interested in the headline story that the Soviet Union had successfully launched Sputnik, the first man-made satellite, into outer

---

34 Both Google Books and the Barney Newspaper Collection from Gale use a combination of geographic mapping, image location, and OCR to enhance the user experience, although they stop short of replicating the experience of reading a physical book or newspaper in their software design.

35 Ehrmann, Bunout, and Clavert, "Digitised Historical Newspapers," 4.

36 Cotler and Sandhaus, "How to Build a TimesMachine."

37 Ibid. and Salisbury, *Without Fear*, 5–8. The archive has not had a major update in terms of content since 2014, because articles written post-2002 are retrievable as archived webpages on NYTimes.com where the online version of the paper has appeared since 1996.

38 Colter and Sandhaus, "How to Build a TimesMachine" and Sandhaus, "Introducing the New TimesMachine."

39 Sydney Gruson, "Warsaw Crushes New Protest; Clubs, Tear Gas Rout Students," *The New York Times*, October 5, 1957.

space.<sup>40</sup> Knowing where the article on student protestors appeared (the bottom right of the front page), as well as what else occurred that day makes the source far more valuable to a historian, as it allows for the analysis of an article's content as well as the reach and social importance of that knowledge to contemporary readers in its time.

As with most digital databases, TimesMachine employs a digital search and retrieval engine based on OCR that prompts users to enter keywords to explore the collection. Keywords in the TimesMachine are words serving as categorical identifiers, which then link to relevant articles, headlines, and letters to the editor published between 1851 and 2002. A keyword may represent different subjects, people, and locations, depending on what the researcher is trying to find. A conceptual historian, for example, could use a particular concept such as “mediocrity” or “intelligence” as a keyword to trace its use and development in the news.<sup>41</sup> Researchers interested in historical actors or networks can use names of individuals and institutions, such as “John F. Kennedy,” “Ford Foundation,” or “Harvard University,” while historians hoping to follow the news coverage of particular regions can use the names of cities, states, and countries to gather comprehensive coverage from these regions at different times. However, it is not possible to search for articles written by specific journalists, which represents a major limitation of the archive for those hoping to research journalists as historical actors in the production of knowledge.

Most researchers want to follow the news coverage of a particular historical event or phenomenon and see those events in their original context, which is not always easy to synthesize into a single keyword. The TimesMachine search interface makes it possible to conduct searches with slightly more nuance and focus. These include allowing users to (1) simultaneously search multiple keywords, (2) delineate the time period of the search from one day of news coverage to news coverage spanning multiple years, and (3) add another layer of depth by keyword and category searches *within* each individual newspaper. While the collection search tool is not compatible with Boolean operators, it does allow for a combined search of keywords, names, and locations that might appear together, separated by

---

<sup>40</sup> Walter Sullivan and William J. Jorden, “Soviet Fires Earth Satellite into Space; It is Circling the Globe at 18,000 M.P.H; Sphere Tracked in 4 Crossings Over U.S.,” *The New York Times*, October 5, 1957.

<sup>41</sup> Such an approach was employed in a recent article by Chelsea A. Rodriguez and Sarah Van Ruyskensvelde, “A Rising Tide of Discontent: Mediocrity, Meritocracy, and Neoliberalism in American Education, 1971–1983,” *Paedagogica Historica* (2021), April 13, 2023, doi:10.1080/00309230.2021.1999276.

a semicolon.<sup>42</sup> The timespan of these searches can be set by the users, allowing them to search for keywords that appear on a particular day, or within any month, year, or number of years over the one-hundred-and-fifty-year span of content. While keyword searches and time-delineation do not constitute features unique to the TimesMachine, the multi-level organization and individual indexing of newspapers allow researchers to explore the continuity and change of social ideas, concepts, and policies more easily.

In a recently published study, I employed these methods of keyword cross-referencing and time setting together to trace the development of “mediocrity” as an educational concept in *The Times*. To trace the evolution of “mediocrity” leading up to the release of *A Nation at Risk*,<sup>43</sup> which warned of a “rising tide of mediocrity” in American schools, I had to collect every article using this concept to discuss U.S.-American education in *The Times* between 1971 and April 27, 1983.<sup>44</sup> Using the keyword “education” resulted in 71,652 articles in this time period, while “mediocrity” resulted in 958 hits. However, when combining “education” and “mediocrity,” the collection whittled down to 187 results, a manageable amount for analysis. Close reading was still required to identify whether the categorization of education matched the research criteria for education news coverage, but these search features enabled me to easily collect source materials demonstrating how the concept developed as it moved between different school debates and policy discussions.<sup>45</sup> Keyword searches and delineating time periods are ubiquitous features in digital newspaper archives; however, this case illustrates how the mass digitization of newspapers has changed the ways in which we can apply mixed methods approaches in qualitative and quantitative analyses.

Perhaps the most crucial feature for researchers is that each daily paper in the TimesMachine is also individually indexed and searchable, a feature that is absent from most other digital newspaper archives. Users can use their own keywords to search the whole collection, but each digitized newspaper also contains an index with pre-coded categories created by the digital archivists. This allows readers to

---

<sup>42</sup> Boolean operators are commonly used conjunctions (such as AND, OR, or NOT) used as directives to combine or exclude keywords in a search, thereby resulting in more focused and productive search results in digital archives and databases.

<sup>43</sup> *A Nation at Risk: The Imperative for Educational Reform* (1983) was a report by the United States National Commission on Excellence in Education. Among other things, the report contributed to a discourse that American schools were failing and falling behind international competitors in terms of academic achievement and supremacy. While highly controversial and rhetorically charged, it set off a wave of local, state, and federal reform efforts that shaped education as we know it in the contemporary United States.

<sup>44</sup> Rodriguez and Van Ruyskensvelde, “A Rising Tide,” 4.

<sup>45</sup> Ibid., 17.

quickly find important topics, individuals, and stories without having to manually read each page.<sup>46</sup> Having pre-set as well as custom keyword searches caters to both casual users of the archive and experienced researchers seeking more control in the search process.<sup>47</sup> The presence of multi-level organization in the archive may not seem all that groundbreaking, but for as-of-yet indiscernible reasons, there are often inconsistencies between what you can find using a whole-collection search and a within-paper search.<sup>48</sup> Within-paper keyword searches allow researchers to collect articles that are more relevant and contextualized to their needs.<sup>49</sup>

Each individual newspaper index is typically organized into five pre-set categories: *People*, *Organizations*, *Creative works*, *Places*, and *Descriptors*. *People* includes all named individuals appearing in an article in the paper. *Organizations* includes a broad range of collective groupings such as universities, businesses, non-profit organizations, government agencies, and more social categories, such as the use of “Negroes” as an organization throughout the 1950s and 1960s.<sup>50</sup> *Creative works* organizes all content related to the arts and creative expression, while *Places* lists all news by the locations discussed in the articles, ranging from local towns to nations across the globe. Finally, *Descriptors* refer to a range of concepts and categories, from “Education and Schools,” to “Deaths” and relevant current events such as “The Presidential Election.” These pre-determined categories are hyperlinked, thus making it easy for the user to jump to the results in the paper and see the stories in their original contexts. The results in these five categories are sorted first by prevalence in the paper and then alphabetically. For example, on Wednesday, May 6, 1970, the index for that day’s paper reads as follows:

---

<sup>46</sup> As with many other archives, digital or otherwise, the methods used by the designers of the archive to create and organize these categories have not been disclosed.

<sup>47</sup> Adam Crymble, “Digital Library Search Preferences amongst Historians and Genealogists: British History Online User Survey” 10, no. 4 (2016), accessed January 31, 2023, <http://www.digitalhumanities.org/dhq/vol/10/4/000270/000270.html>.

<sup>48</sup> Perhaps the sheer volume of content and inconsistencies in metadata is to blame, but as a researcher, I have experienced these lapses and disconnections firsthand.

<sup>49</sup> It should be noted, however, that the within-paper indexes and keyword searches are also imperfect. For example, stories about certain universities or schools may not appear in a keyword search of “university” but will appear in the indexed categories as a specific organization (e.g., Kent State University).

<sup>50</sup> It is unclear how the index terms and categories were created in the archive. The use of now antiquated concepts as indexed categories suggests that, whether by AI or manual categorization, they were likely categorized based on terms and vocabulary originating from the historical newspapers themselves.

---

**Index**


---

**Wednesday, May 6, 1970**
[See All](#)


---



---

**PEOPLE**


---

Dickens, Charles John Huffam	2
Nixon, Richard Milhous	2
Abelson, Philip H	1
Alioto, Joseph L	1
Bauer, Yehuda	1

---



---

**ORGANIZATIONS**


---

Democratic Party	5
Kent State University	5
Congress	3
Black Panther Party	2
General Motors Corp.	2

---



---

**CREATIVE WORKS**


---

Chairs, The	1
Cherry Orchard, The	1
Endgame	1
Jeune Fille A Marier, La	1
Lacune, La	1

---



---

**PLACES**


---

United States	37
Vietnam	24
Cambodia	21
Cambodian Border Area	12
New York City	9

---

Continued

Index

Wednesday, May 6, 1970

See All

DESCRIPTORS

General Policies of Belligerents, Other Major Powers	18
Military Action	14
Deaths	13
Elections	10
Editorials	6

Even without any prior knowledge of U.S.-American history in this time period, the index presents the user with an inclination as to the dominant mood and most pressing concerns for readers. Someone familiar with this period in U.S.-American history would instantly recognize the links between the pre-selected categories: two days prior to the printing of this newspaper, four college students were killed by the U.S. National Guard at *Kent State University* while protesting the recent invasion of *Cambodia* by the *U.S. Military*, an action approved by newly elected *President Richard Nixon* as part of the ongoing *Vietnam War*. Thus, with some additional contextual knowledge, historians and researchers can with relative ease navigate the coverage of the events that shaped modern history.

## Limitations of the Digital Space

While TimesMachine offers many possibilities for researchers, there are certainly noteworthy limitations. Some pitfalls include the limited reach and reliability of keyword searches, as well as the lack of transparency regarding why only certain editions of the paper are accessible in the archive. While these limitations do impact research in the TimesMachine in particular ways, they represent common limitations found in a majority of (digital) archives.<sup>51</sup> A more unique limitation is the lack of metadata for advertisements in the TimesMachine, which makes it difficult or nearly impossible to conduct research into advertisements as sources of knowledge.

51 Ehrmann, “Historical Newspaper User Interfaces,” 15.

Keyword searches have limitations that most researchers are sorely aware of, as complex concepts cannot easily be distilled into searchable keywords and researchers must be creative and resourceful in how they choose their search terms. On top of this, while digital archives and search engines make it easier to mine data, the vast amount of data collected with keywords also makes it more difficult to frame and interpret this data and use it to write history.<sup>52</sup> For example, conducting a keyword search in the TimesMachine for all articles containing the keyword of “education” in the year 1957 will garner 5,723 results; yet, if you search for the keyword “school” in the same period, it garners over 15,000 results. Both concepts have multiple meanings and definitions, and AI or OCR systems cannot employ deductive reasoning or have a sense of what the researcher is searching for or hoping to find.<sup>53</sup> The search algorithm cannot differentiate between the subjective conceptual meaning and definitions of keywords, which is where the researcher must always come in.

Choosing the right keywords often requires researchers to possess perseverance and creativity, as well as an intimate knowledge of the culture and texts being studied.<sup>54</sup> It also requires researchers to distill the subject of their inquiry into a number of smaller, focused keywords to whittle down the thousands of hits and comb through the results by means of close reading: not a groundbreaking method, perhaps, but an effective one. Exactly how the metadata connects articles, categories, and keywords together in the TimesMachine is not public information. As with all archives (digital and otherwise), researchers are at the mercy of the unseen choices of these coders and archivists.<sup>55</sup> This means that we even more

---

52 Abel, “Pleasures and Perils,” 10.

53 The strengths and limitations of AI and OCR, particularly with regard to how their algorithms impact the presentation of archival materials, are increasingly discussed in the fields of digital history and archival research. One notable example of this scholarship is that of Gregory Rolan et al., “More Human Than Human? Artificial Intelligence in the Archive,” *Archives & Manuscripts* 47, no. 2 (2018): 179–203, <https://doi.org/10.1080/01576895.2018.1502088>.

54 Adrian Bingham, “The Digitization of Newspaper Archives: Opportunities and Challenges for Historians,” *Twentieth Century British History* 21 (2010): 231, as cited in Nicholson, “Digital Turn,” 67.

55 The mass digitization of source materials still does not resolve the issue that archives are curated, not just collected. They are comprised of artifacts and knowledge having been compiled through the choices of archivists motivated by professional, but also cultural and national inclinations. This means that minority voices and experiences are often silenced in archives, digital or otherwise, as explained in Rodney G. S. Carter, “Of Things Said and Unsaid: Power, Archival Silences, and Power in Silence,” *Archivaria* 61 (September, 2006): 215–233, accessed January 31, 2023, <https://archivaria.ca/index.php/archivaria/article/view/12541>.



so need to recenter the material aspects of newspapers that persevere in and inform the digitized versions we grapple with.

Another major limitation is the lack of transparency as to which editions of the newspaper are actually included in the TimesMachine. Besides the daily and Sunday versions of the paper, *The New York Times* was also printed in different editions depending on its target audience, including the New Jersey Edition, the Tri-State Edition, and the most-circulated City Edition. Researchers can determine which edition they are reading in the TimesMachine by going to the front page of the paper and looking to the top right corner of the masthead, once again reiterating the value and importance of full-page scans. When using the digital archive, it is important to be aware of these different editions as the TimesMachine primarily includes scans of the Late City Edition.<sup>56</sup> Crucially, the Late City Edition is the de facto version of the paper, but the “Late” indicates that additional news updates or edits have been made to the paper after the initial printing and distribution.

Until the printing process was expedited at the end of the twentieth century, each edition of *The Times* was finalized and approved for printing at 6 pm the night before and distributed to newsstands by 4 am the next morning. If something major happened in the meantime, or if a mistake was found by an editor or journalist post-printing, they were added or fixed in the Late City Edition, which was printed and distributed later in the day. It is not possible to compare the different editions of the newspaper using the TimesMachine, as it typically only displays one copy of the paper for each day. Sometimes, the only edition available in the TimesMachine on a given day is an even more limited edition of the paper, such as the short-lived Western Edition (1962–1964) or the International Edition (1946–1967). This subtle but important decision to prioritize the Late City Edition in the archive to an even greater extent means that users are not really seeing the newspaper as it was originally experienced by its initial audience, an important consideration to keep in mind when using these digitized newspapers as windows to the past. These possible differences between the versions of these newspapers that have been digitized (and those that have not) can have major implications on the type of information you can glean from them.<sup>57</sup>

It matters which newspaper edition the article came from because the content of some articles changed depending on the edition and its target audience. For example, on December 18, 1975, *The Times* ran a three-column story in the middle of

---

<sup>56</sup> The masthead is the heading on the front page of the newspaper. *The New York Times* masthead features the name of the paper at the center top and is flanked by the paper’s motto “All the News that’s Fit to Print” on the left, and edition and daily weather information on the right.

<sup>57</sup> As explained in the research saga detailed by Jon C. Giullian in “Seans Chernoi Magii Na Ta-ganke.”

the Late City Edition titled “More New Jersey Parents Are Sending Children to Private Schools in the City.” The scoop focused on some of the more prominent private schools in New York City, and the financial and time sacrifices New Jersey parents were making to help their children escape the “mediocrity” of their local public schools. However, the story changed when it appeared in the New Jersey City Edition, where it appeared on the front page at greater length, with a more negative tone, and more quotes from parents and evidence of failing local schools. While this certainly made it more attention-grabbing for the specific audience the story covered, the New Jersey version was likely seen by far fewer readers than the City or Late City editions.

Finally, advertising in the *New York Times* is not retrievable in the TimesMachine digital archive. When the archive was first launched, its developers assured that advertisements would be just as “legible, linkable, and shareable” as all other content in the archive.<sup>58</sup> However, the combination of different typefaces, complicated layouts, and the difficulty of algorithmically differentiating advertisements were later cited by *The Times* as reasons for the lack of metadata regarding advertisements.<sup>59</sup> This exclusion is significant, as a large number of ads in the paper were actually text-based, and special interest groups such as labor unions, political parties, and education associations often paid for advertised columns in order to target policy messaging to readers.<sup>60</sup> While they are visually similar to news stories and columns in the newspaper, these columns are classified in the same way as the more image-based (perhaps stereotypical) advertisements for travel, fashion, and food products. Crucially, these advertisements are not simply passive, isolated sources; readers often engaged in a dialogue with these targeted advertisements via letters to the editor, and those responses are findable via the keyword search tool.<sup>61</sup>

For example, Albert Shanker of the American Federation of Teachers (AFT) wrote an advertisement column “Where We Stand” in *The New York Times* for 27 years, offering “800 words of common sense, keen analysis and no-nonsense ideas about how to improve schools.”<sup>62</sup> This column, classified as an advertisement

---

<sup>58</sup> Sandhaus, “Introducing the New TimesMachine.”

<sup>59</sup> Ibid., also because the OCR used by the TimesMachine is not able to attribute metadata to images.

<sup>60</sup> The exclusion of metadata for these paid columns might be due to the fact that the TimesMachine archive focuses on cataloguing content of which *The Times* takes ownership. Paid advertising content, while it did appear in the original paper, might be subject to different copyright rules.

<sup>61</sup> Rodriguez and Van Ruyskensvelde, “A Rising Tide,” 15.

<sup>62</sup> Albert Shanker Institute, “Where We Stand,” Archives Online, accessed January 30, 2023, <https://www.shankerinstitute.org/resource/where-we-stand-archives-online>. However, it should be noted that while this archive seems to exist, efforts by the researcher to access it were not successful.

since the AFT paid for this space in the paper, was frequently commented on by readers and generated a wealth of debate in the letters to the editor section of the paper over school privatization and teacher pay.<sup>63</sup> These letters allow researchers interested in these columns to track down the original advertisements, as the readers always refer to the date on which the original column appeared in their response. But, without a response letter, these advertisements can only be found by manually flipping through the digital pages week-by-week. *The Times* did develop a crowd-sourced archive of their advertisements called Madison, named after the famous avenue in New York City where a majority of advertising agencies have their offices.<sup>64</sup> However, for unclear reasons, Madison is no longer accessible to users, which means that researchers interested in the history and knowledge-producing potential of advertisements in *The Times* are essentially limited to manually searching page by page.

## Making the Most of the Material: Recentering Context in the TimesMachine

If these digital limitations still persist, even in an archive designed to close the gap between digital and material, how then can we still make the most of databases like the TimesMachine? Here, I propose adopting a methodological approach to digital newspaper analysis that recenters the material aspects of the newspaper. I previously mentioned the Warsaw student protests and the Sputnik satellite launch, both of which made front page news on October 5, 1957. Both stories were newsworthy and touched upon similar Cold War anxieties, but their differing placement reveals a material aspect of newspapers that is often neglected or obscured in digital archives: the values and business motivations informing their placement.

When interpreting the knowledge produced by newspapers such as *The Times*, the placement of a story matters just as much as its content, if not more. Placement is a deliberate choice by the editor, and the location of a story within the paper can give us an indication of the perceived relevance and importance of a story to readers in comparison to other events. “News values,” the often unspoken theoretical values in journalism that determine the newsworthiness of an event, can help researchers understand the mechanisms through which such news stories were

---

<sup>63</sup> Rodriguez and Van Ruyskensvelde, “A Rising Tide,” 15.

<sup>64</sup> “Madison,” *NYT R&D, The New York Times*, 2015, accessed April 11, 2023. <https://nytlabs.com/projects/madison.html>.

often selected, placed, and emphasized depending on how the audience's perception and the publisher's motivations intertwined.<sup>65</sup> The *proximity* of a story to readers, the *consonance* of the topic with their values and beliefs, and the *recency* of the event are just a few examples of news values that influence where a story will likely appear within a paper.

Beyond news values and wanting to publish impactful news, *The Times* (just like other newspapers) is a business, and its front page was designed to entice customers and sell copies to fund its expansive news empire. This means that news and market values greatly influenced the placement of stories, particularly the ones appearing on the front page. Here, however, we encounter another crucial material aspect of the newspaper that gets lost with digitization: how it *folds*. In the TimesMachine, users are presented with the full span of the front page in all its glory, but, crucially, in its *unfolded* glory. Readers in the past did not receive their newspaper this way, and while it may seem like a small distinction, it is a clear example of how digitization can change what we see and how we see our source materials.

It is important to remember how our sources were actually used and functioned in their context, particularly when researching the circulation of knowledge. When the morning copies of *The Times* arrived at the newsstand, they were stacked and displayed folded in half. Copies of the paper delivered to suburban doorsteps came similarly folded in half and rolled up. With this in mind, the front page of *The Times* for much of the twentieth century was organized so that the most relevant stories occupied the upper half of the front page, with the most important story typically occupying the top-right position under the masthead. That is not to say that stories appearing on the bottom half of the front page were not important – only, perhaps, less so in the eyes of the editor. We know that the student protest in Warsaw was featured on the bottom-right half of the front page on October 5, 1957, but by employing news values and material knowledge of the newspaper, it is easy to predict that the Sputnik story occupied the main headline and top-right position of the front page. Analyzing newspapers with these material inclinations in mind may sensitize researchers to the priorities of the newspaper, the values embedded in its organization, and the types of knowledge it sought to platform.<sup>66</sup> It further illustrates the importance of actually understanding the material context of the sources we use instead of solely relying on

---

<sup>65</sup> Bell, *The Language*, 155–160; Michael Schudson, “The Sociology of News Production,” *Media, Culture, and Society* 11 (1989): 263–265.

<sup>66</sup> Simone Lässig, “The History of Knowledge and the Expansion of the Historical Research Agenda,” *Bulletin of the GHI Washington*, no. 59 (Fall 2016): 31, accessed January 31, 2023, <https://prae-perspectivia.net/publikationen/bulletin-washington/2016-59/0029-0058>.

keyword-retrieved articles. In other words, digital archives require us to focus not only on the contours of the knowledge map they present us, but also on the actual, original terrain those maps represent.<sup>67</sup>

## In Conclusion

Above all, this case study demonstrates the value of incorporating digital history methods and perspectives into the history of knowledge and history of education. The aim of this article was to explore how historians of knowledge should best grapple with digitized historical newspapers and the archives that house them. The case of *The New York Times* TimesMachine archive was used to illustrate the possibilities and limitations of using digitized newspaper archives in historical research. There are several important takeaways from this article that can help researchers make informed methodological choices in their pursuit of history, particularly when grappling with sources such as newspapers which produce and circulate various mediated forms of knowledge.

Before entering the digital archive, it is vital to remember that every news institution has its own historical context, internal logic, actors, and values that inform its practices. Newspapers, like other institutionally produced source materials, have specific features that require researchers to zoom out and assess before zooming in on the archive. Researchers need to know if they are dealing with stonecutters or cathedral builders to better understand the sources of knowledge produced by those institutions. The TimesMachine presents searchable, full-page scans of one hundred and fifty years of news coverage and aims to recreate the experience of flipping through a newspaper to overcome the common problem of decontextualization in digital newspaper archives. The immersive interface and multi-level newspaper indexation open up a number of research possibilities for historians, but there are still distinct limitations that make some research approaches difficult to execute.

While the archive presents the contours of the newspaper as an arena of knowledge circulation and production, there are certain limitations regarding keyword searches, source selection, and advertisements. As in other digital archives, keyword searches represent an imperfect process, and the multi-level indexing of the collection still means that the researcher needs to think creatively and use historical knowledge to effectively choose keywords and navigate search results. Moreover, there is a lack of transparency regarding how the collection was

---

<sup>67</sup> Beals and Bell, *Atlas of Digitised Newspapers*, i.

organized and which edition of the paper has been preserved. This has major implications for those trying to research the circulation of ideas in the paper, as some editions were distributed to a smaller number of readers. To alleviate these limitations, I advocate for a contextualized approach reemphasizing the material in this digital, abstracted space. Finally, advertisements, which serve as rich sources of social and cultural knowledge, are not indexed in the TimesMachine. This makes it very difficult to research advertised political columns and the circulation of political and social knowledge in advertising. However, I propose a few ways in which researchers may navigate around these limitations by cross-referencing letters to the editor that responded to certain advertisement columns. Just as is the case of traditional archives, users need to be aware of the logic of the digital archives they engage with and reflect on how digitization influences their sources and research.<sup>68</sup>

These findings and methods were developed in the context of a study on the history of education in the post-war era, but the integrative and generative capacity of the history of knowledge makes these approaches well-suited for studying the production of knowledge and circulation of ideas in several different news topics, themes, and contexts.<sup>69</sup> Similarly, while this study focuses on one unique digital archive, the main takeaways concerning how news content differed based on the days of the week, the edition of the paper, and the editor's inclinations and values are also relevant for research in all digital newspaper archives. Digital archives have undoubtedly expanded the possibilities for historians with regard to keyword searches and mass digitization, but the suspension of these newspapers in the digital space can still obscure key information on how the source material also informed the knowledge reproduced and circulated on its pages.

## About the contributor

**Chelsea A. Rodriguez** is a PhD candidate in the history of education at the University of Groningen (Netherlands). She is interested in educational debates and the mediated communication of educational ideas in the public sphere. Her dissertation investigates the history of education news in the United States, specifically how the institution of *The New York Times* circulated particular forms of knowledge about U.S.-American education to readers during the post-war era.

---

<sup>68</sup> Jensen, "Digital Archival Literacy," 252.

<sup>69</sup> Östling and Heidenblad, "Fulfilling the Promise," 1; Lässig, "History of Knowledge," 32.

